Proceedings of Machine Learning Research vol 195:1–113, 2023 36th Annual Conference on Learning Theory

Tackling Combinatorial Distribution Shift: A Matrix Completion Perspective

Max Simchowitz MSIMCHOW@MIT.EDU
MIT, Cambridge, MA, 02139

Abhishek Gupta ABHGUPTA@CS.WASHINGTON.EDU
University of Washington, Seattle, WA, 98105

Kaiqing Zhang KAIQING@UMD.EDU
University of Maryland, College Park, MD, 20740

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

Obtaining rigorous statistical guarantees for generalization under distribution shift remains an open and active research area. We study a setting we call *combinatorial distribution shift*, where (a) under the test- and training-distributions, the labels z are determined by pairs of features (x, y), (b) the training distribution has coverage of certain *marginal* distributions over x and y separately, but (c) the test distribution involves examples from a product distribution over (x, y) that is *not* covered by the training distribution. Focusing on the special case where the labels are given by *bilinear embeddings* into a Hilbert space $\mathcal{H}$: $\mathbb{E}[z \mid x, y] = \langle f_\star(x), g_\star(y)\rangle_{\mathcal{H}}$, we aim to extrapolate to a test distribution domain that is not covered in training, or *bilinear combinatorial extrapolation*.

Our setting generalizes a special case of matrix completion from missing-not-at-random data, for which all existing results require the ground-truth matrices to be either *exactly low-rank*, or to exhibit very sharp spectral cutoffs. In this work, we develop a series of theoretical results that enable bilinear combinatorial extrapolation under *gradual* spectral decay as observed in typical high-dimensional data, including novel algorithms, generalization guarantees, and linear-algebraic results. A key tool is a novel perturbation bound for the rank-k singular value decomposition approximations between two matrices that depends on the *relative* spectral gap rather than the *absolute* spectral gap, a result we think may be of broader independent interest.

1. Introduction

While statistical learning theory has classically studied *out-of-sample generalization* from training data to test data drawn from the same distribution (e.g., Bartlett and Mendelson (2002); Vapnik (2006)), in almost all practical settings, one wishes to ensure strong performance on data which may be generated quite differently from the training data (Koh et al., 2021; Taori et al., 2020). This paper studies formal guarantees for a type of *out-of-distribution* generalization we call *combinatorial distribution shift*. Informally, we consider predictions from pairs of features (x, y) such that: (a) the *marginal* distributions of each of the features separately under the test data are covered by the training distribution, but (b) the *joint* distribution of the features may not be covered. We refer to *combinatorial extrapolation* as the process of generalization under combinatorial distribution shift. Our setting may encompass a broad swath of applications including: computer vision tasks which extrapolate to novel combinations of objects, backgrounds, and lighting conditions that have been

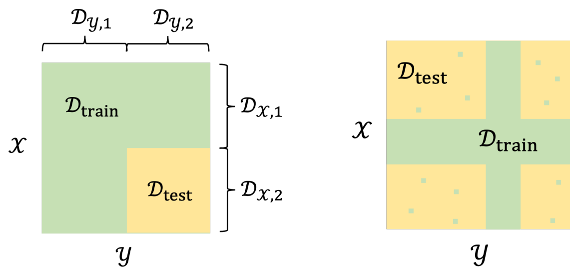

Figure 1: Illustration of the combinatorial extrapolation settings where the marginal distributions of $\mathcal{D}_{\mathrm{test}}$ is covered by $\mathcal{D}_{\mathrm{train}}$, while some combinations of them are not covered by $\mathcal{D}_{\mathrm{train}}$.

seen individually (Liu and Han, 2016); extrapolation to manipulating objects with novel combinations of masses, shapes, and sizes in robotic manipulation (Tremblay et al., 2018); extrapolation to predictions of the outcomes of medical intervention from one set of subgroups to others with novel combinations of salient traits (Gilhus and Verschuuren, 2015). See Figure 1 for an illustration.

Bilinearity, low-rank structure & matrix completion. A popular technique for compositional and combinatorial generalization is to *embed* features into a semantic vector space (Mikolov et al., 2013). For example, CLIP (Radford et al., 2021) learns embedding words and text into an *inner-product space* in order to achieve zero-shot generalization to new image classes. In this work, we adopt a matrix-completion perspective to study the potential of these bilinear approaches. Indeed, if the features (x, y) correspond to indices of a large data matrix, bilinear combinatorial extrapolation may be understood as *matrix completion*: complete an entire matrix $\mathbf{M} \in \mathbb{R}^{n \times m}$ from observing a subset $\Omega \subset [n] \times [m]$ of its entries. The estimation of accurate bilinear embeddings, then, corresponds to finding a low-rank approximate factorization of the data. We detail this connection in Appendix C. Whereas classical results study the *missing-at-random* (MAR) regime where Ω is drawn uniformly at random (Candes and Recht, 2012; Recht, 2011; Hastie et al., 2015), the absence of joint-distribution coverage makes our setting a special case of *missing-not-at-random* (MNAR) recovery (see, e.g., Ma and Chen (2019)). There is a rich literature on MNAR matrix recovery (see a detailed review in Appendix B). A common assumption in this literature of MNAR matrix recovery is that, the data matrix $\mathbf{M}$ is either *exactly low-rank*, or exhibits *sharp drop-offs* between adjacent singular values. This is in contrast to MAR matrix recovery, where it suffices that the singular values of $\mathbf{M}$ are only summable (Koltchinskii et al., 2011). While it is widely accepted that real data are approximately low-rank (Udell and Townsend, 2019), they tend to exhibit the more *gradual* singular value decay required by MAR matrix recovery, than the *rapid* decay necessitated by the existing MNAR-case results. Indeed, the spectra of random data matrices have continuous limiting distributions (Bai and Silverstein, 2010), and thus their singular values *do not* exhibit sharp cutoffs.

Our contributions. This paper demonstrates conditions under which *bilinear predictors* are statistically consistent under combinatorial distribution shift. We assume real labels z can be predicted from pairs of features $(x, y) \in \mathcal{X} \times \mathcal{Y}$ via bilinear embeddings into a Hilbert space $\mathcal{H}$: $\mathbb{E}[z \mid x, y] = \langle f^\star(x), g^\star(y) \rangle_{\mathcal{H}}$. We then state structural assumptions, inspired by a canonical case of matrix completion with MNAR data (see Figure 2 and Appendix C), which facilitate extrapolation from a training distribution $\mathcal{D}_{\mathrm{train}}$ over pairs (x, y) that has a full coverage of certain marginal distributions over x and y separately, to a test distribution $\mathcal{D}_{\mathrm{test}}$ containing samples from a product distribution over (x, y) that is *not* covered by $\mathcal{D}_{\mathrm{train}}$. In contrast to the MNAR matrix completion literature described above, we analyze a setting more akin to the kernel least-squares literature (Bissantz et al., 2007; Mendelson and Neeman, 2010), where a suitably defined feature covariance

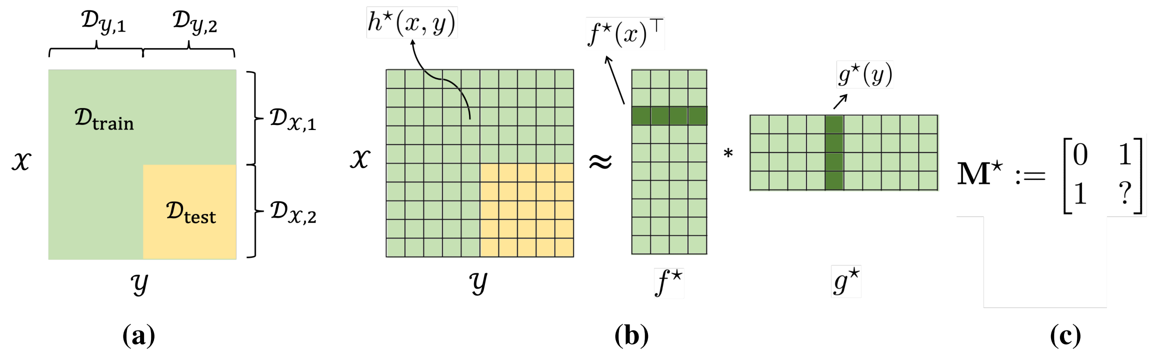

Figure 2: **(a)** Bilinear combinatorial extrapolation that satisfies the 2×2 block decomposition; **(b)** A basic case of **(a)** with discrete distributions, which can be viewed as matrix completion with MNAR data, and the bilinear representation of the distribution naturally appears; **(c)** An example of a matrix that does not satisfy Assumption 2.3 and thus fails to be completed uniquely.

matrix $\boldsymbol{\Sigma}^\star_{1\otimes 1}$ (Assumption 2.4) may exhibit spectral decay as gradual as $\lambda_i(\boldsymbol{\Sigma}^\star_{1\otimes 1}) \le Ci^{-(1+\gamma)}$ for some $\gamma > 0$ (Assumption 2.6). Our contributions are detailed as follows.

- Given finite-rank embeddings $\hat{f} : \mathcal{X} \to \mathbb{R}^r$ and $\hat{g} : \mathcal{Y} \to \mathbb{R}^r$, we establish a meta-theorem, Theorem 2, which establishes upper bounds for the excess risk $\mathcal{R}(\hat{f}, \hat{g}; \mathcal{D}_{\text{test}}) := \mathbb{E}_{\mathcal{D}_{\text{test}}}[(\langle \hat{f}, \hat{g}\rangle_{\mathcal{H}} - \langle f^\star, g^\star\rangle_{\mathcal{H}})^2]$ on $\mathcal{D}_{\text{test}}$ by the excess risk on $\mathcal{D}_{\text{train}}$, and the error on a sub-distribution $\mathcal{D}_{1\otimes 1}$ of $\mathcal{D}_{\text{train}}$, which corresponds to a dense diagonal block matrix in MNAR matrix completion.
- Using the meta-theorem, we show in Theorem 3 that if $(\hat{f}, \hat{g})$ above are trained via a *single stage* of supervised empirical risk minimization (ERM) (from a suitably expressive function class), then whenever it happens that $(\hat{f}, \hat{g})$ are well-conditioned (in a sense defined), $\mathcal{R}(\hat{f}, \hat{g}; \mathcal{D}_{\text{test}})$ scales with an inverse of some polynomials in the number of samples and in the rank r, provided that the exponent γ in the polynomial decay satisfies $\gamma > 3$.
- Finally, we introduce a *double-stage* ERM procedure (Algorithm 1), which produces final estimates $(\hat{f}, \hat{g})$ of the embeddings that (with high probability) are guaranteed to be well-conditioned, and have $\mathcal{R}(\hat{f}, \hat{g}; \mathcal{D}_{\text{test}}) \to 0$ for any decay exponent $\gamma > 0$ (see Theorem 4).

1.1. Relative singular-gap perturbation bound for the SVD approximation

Before describing our overall proof strategy, we highlight a key technical ingredient that we believe may be of more universal interest. Consider two real matrices $\mathbf{M}^\star, \hat{\mathbf{M}} \in \mathbb{R}^{n\times m}$, and let $\sigma_k(\cdot)$ denote the k-th largest singular value. The celebrated Davis-Kahan Sine Theorem and its generalization, Wedin's Theorem (see, e.g., Stewart and Sun (1990)), states that the principal angles between their (left or right) singular spaces scale with $\|\mathbf{M}^\star - \hat{\mathbf{M}}\|_{\mathrm{F}}/\delta_k^{\text{abs}}(\mathbf{M}^\star)$, where $\delta_k^{\text{abs}}(\mathbf{M}^\star) := \sigma_k(\mathbf{M}^\star) - \sigma_{k+1}(\mathbf{M}^\star)$ denotes the *absolute* singular gap. For the special case of multiplicative perturbations, $\hat{\mathbf{M}} = (\mathbf{I} + \boldsymbol{\Delta}_1)\mathbf{M}^\star(\mathbf{I} + \boldsymbol{\Delta}_2)$ with matrices $\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2$ close to zero, the perturbation scales with the (possibly much smaller) relative singular value gap (Li, 1998),

$$\delta_k(\mathbf{M}^\star) := \frac{\sigma_k(\mathbf{M}^\star) - \sigma_{k+1}(\mathbf{M}^\star)}{\sigma_k(\mathbf{M}^\star)}. \tag{1.1}$$

So far, we have reviewed bounds on the deviation in the singular value *subspaces* of the matrices $\mathbf{M}^\star$ and $\hat{\mathbf{M}}$. But in many cases, we do not know about these subspaces, but instead, know about the differences in the rank-k SVD approximations to these matrices. For this desideratum, we establish a perturbation bound which depends only on the *relative gap* and which, unlike the singular subspace bound of Li (1998), applies to *generic, additive perturbations*. Our result is as follows.

Theorem 1 (Perturbation of SVD Approximation with Relative Gap) *Let $\mathbf{M}^\star, \hat{\mathbf{M}} \in \mathbb{R}^{n\times m}$. Fix a $k \le \min\{n,m\}$ for which $\sigma_k(\mathbf{M}^\star) > 0$ and the relative spectral gap $\delta_k(\mathbf{M}^\star)$ (Eq. (1.1)) is positive. Then, if $\|\mathbf{M}^\star - \hat{\mathbf{M}}\|_{\mathrm{op}} \le \eta\sigma_k(\mathbf{M}^\star)\delta_k(\mathbf{M}^\star)$ for some $\eta \in (0,1)$, we have that the rank-k SVD approximations of $\mathbf{M}^\star$ and $\hat{\mathbf{M}}$, denoted as $\mathbf{M}^\star_{[k]}$ and $\hat{\mathbf{M}}_{[k]}$, are unique, and satisfy*

$$\left\|\hat{\mathbf{M}}_{[k]} - \mathbf{M}^\star_{[k]}\right\|_{\mathrm{F}} \le \frac{9\|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{F}}}{\delta_k(\mathbf{M}^\star)(1-\eta)}.$$

Theorem 1 is proven in Appendix D via a careful peeling argument. By contrast, a more naive application of Wedin's theorem incurs a dependence on absolute singular gap $\delta_k^{\mathrm{abs}}(\mathbf{M}^\star)$. Our bound is significantly sharper: for example, consider $\sigma_k(\mathbf{M}^\star) \sim \Theta(2^{-k})$, then $\delta_{k_i}^{\mathrm{abs}} = \sigma_{k_i}(\mathbf{M}^\star) - \sigma_{k_i+1}(\mathbf{M}^\star)$ is of order $O(2^{-(k_i+1)})$, while δ_{k_i} as defined in Eq. (1.1) is of order $\Omega(1)$. Having highlighted this particular technical result, we now turn to an overview of the entire analysis.

1.2. Overview of proof techniques and notation.

Throughout, the key technical challenge, from a matrix completion perspective, is generalizing the case with *sharp* spectral cutoffs to that with a *gradual* spectral decay. This challenge is considerably more difficult for *bilinear factorizations* than that for *linear* predictors studied in typical RKHS settings. Regarding the proof of our meta-theorem, Theorem 2: when distributions on (x, y) have finite support, the bilinear combinatorial extrapolation problem for discrete distributions can be reinterpreted as the completion of a block matrix $\mathbf{M}$ with blocks $\mathbf{M}_{ij}$, given data from blocks $\{(1,1),(1,2),(2,1)\}$. With a careful error decomposition, we argue that the extrapolation error is controlled by the recovery of a factorization of the top-left block $\mathbf{M}_{11}$ (see Proposition 4.1). More specifically, if we let $\mathbf{M}^\star = \mathbf{M}_{11}$ and let $\hat{\mathbf{M}}$ correspond to the estimates of a bilinear predictor $\langle \hat{f}, \hat{g}\rangle$ on the $(1,1)$-block, the key step is to show that if we can factor $\mathbf{M}^\star = \mathbf{A}^\star(\mathbf{B}^\star)^\top$ and $\hat{\mathbf{M}} = \hat{\mathbf{A}}\hat{\mathbf{B}}^\top$, then $\hat{\mathbf{M}} \approx \mathbf{M}^\star$ implies $\hat{\mathbf{A}} \approx \mathbf{A}^\star_{[k]}$ and $\hat{\mathbf{B}} \approx \mathbf{B}^\star_{[k]}$ in the sharpest possible sense, where k is some target rank and $(\cdot)_{[k]}$ denotes rank-k singular value decomposition (SVD) approximation of the matrix. While factor recovery guarantees do exist (notably Tu et al. (2016, Lemma 5.14)), all prior results require sharp spectral cutoffs. To this end, we provide a novel factor recovery guarantee (Theorem 5); this, in turn, relies on Theorem 1 above, as well as a careful *partition* of the singular values of a matrix we call the *well-tempered partition* (see Section 4.5). Limiting arguments pass from the matrix/discrete-distribution case to arbitrary distributions (Appendix J).

Given Theorem 2, the instantiation to a single stage of ERM (Theorem 3) is straightforward. Analyzing our double-stage ERM procedure (Algorithm 1) requires more care. Notably, the analysis depends on a careful characterization of what we term as the *balancing operator* – a linear algebraic operator which determines the change-of-basis in which the positive-definite covariance matrices are equal. Discussion of the algorithm and a proof sketch are given in Section 3.3, with a complete proof deferred to Appendix F; properties of the balancing operator are studied in Appendix L.

Notation. For two probability measures $\mathcal{D}, \mathcal{D}'$, we let $\mathcal{D} \otimes \mathcal{D}'$ denote the product measure, and $\frac{\mathrm{d}\mathcal{D}}{\mathrm{d}\mathcal{D}'}$ the Radon–Nikodym derivative of $\mathcal{D}$ with respect to $\mathcal{D}'$. Upper case bold letters $\mathbf{A}, \mathbf{B}, \mathbf{M}$ denote matrices, lower case bold letters $\mathbf{v}, \mathbf{w}$ denote vectors. Operators and elements of the Hilbert space $\mathcal{H}$ are denoted by bold serafs as $\boldsymbol{\Sigma}$ and $\boldsymbol{\mathsf{v}}$, respectively. Adjoints and transposes are *both* denoted with $(\cdot)^\top$; e.g., $\boldsymbol{\mathsf{v}}^\top$ and $\mathbf{v}^\top$ for $\boldsymbol{\mathsf{v}} \in \mathcal{H}$, $\mathbf{v} \in \mathbb{R}^d$. The i-th entry of a vector $\mathbf{v}$ is denoted by $\mathbf{v}[i]$, the i-th row of a matrix $\mathbf{A}$ by $\mathbf{A}[i,:]$, and the (i,j)-th entry by $\mathbf{A}[i,j]$. The space of symmetric (resp. positive semi-definite, resp. positive definite) d-by-d matrices are denoted as $\mathbb{S}^d$, (resp. $\mathbb{S}^d_+$, resp. $\mathbb{S}^d_{++}$). For $\mathbf{M} \in \mathbb{R}^{d\times d}$, $\sigma_i(\mathbf{M}) \geq 0$ denotes its i-th largest singular value; for symmetric $\mathbf{M}$, $\lambda_i(\mathbf{M})$ denotes its i-th largest eigenvalue, and if $\mathbf{M} \succeq 0$, $\mathbf{M}^{1/2}$ its matrix square-root; similar notation applies to operators $\boldsymbol{\Sigma}$ on $\mathcal{H}$. For $n \in \mathbb{N}$, $[n]$ denotes the set $\{1, \cdots, n\}$, and for finite sets $\mathcal{S}$, $|\mathcal{S}|$ denotes its cardinality. For any Hilbert space $\mathcal{V}$, we use $\langle x, y\rangle_{\mathcal{V}}$ to denote the inner product of $x, y \in \mathcal{V}$, and $\|x\|_{\mathcal{H}}$ to denote the Hilbert norm defined by the product. When $\mathcal{V}$ is omitted, they mean the inner-product and vector norm in the Euclidean space. $\log$ denotes the base-e logarithm.

2. Problem Formulation

In the *bilinear combinatorial extrapolation* problem, covariates $(x, y) \in \mathcal{X} \times \mathcal{Y}$ are regressed to real labels $z \in \mathbb{R}$. We are given access to a training distribution $\mathcal{D}_{\mathrm{train}}$ and a test distribution $\mathcal{D}_{\mathrm{test}}$ on $\mathcal{X} \times \mathcal{Y} \times \mathbb{R}$. We assume that the Bayes optimal predictor is identical between the two distributions, and is given by the inner product of bilinear embeddings defined below.

Assumption 2.1 (Bilinear Representation) *There is a Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ and two embeddings $f^\star : \mathcal{X} \to \mathcal{H}$ and $g^\star : \mathcal{Y} \to \mathcal{H}$ satisfying that $h^\star(x,y) := \langle f^\star(x), g^\star(y)\rangle_{\mathcal{H}}$ is the Bayes optimal predictor on $\mathcal{D}_{\mathrm{train}}$ and $\mathcal{D}_{\mathrm{test}}$, i.e., $\mathbb{E}_{\mathcal{D}_{\mathrm{train}}}[z \mid x, y] = \mathbb{E}_{\mathcal{D}_{\mathrm{test}}}[z \mid x, y] = h^\star(x, y)$. Also, $\mathbb{E}_{\mathcal{D}_{\mathrm{train}}}[\langle f^\star(x), g^\star(y)\rangle^2] < \infty$.*

Assumptions that facilitate extrapolation. The bilinear structure of $h^\star$ is insufficient for general combinatorial extrapolation; otherwise, in the finite-dimensional case, a matrix would have been completable from a single entry. We, therefore, assume that our training distribution can be decomposed into four blocks, such that the first three blocks, i.e., the blocks $(1, 1), (1, 2), (2, 1)$, are "covered" under $\mathcal{D}_{\mathrm{train}}$, but the fourth block, i.e., the block $(2, 2)$, may only be covered under $\mathcal{D}_{\mathrm{test}}$. It is formally introduced in the following assumption.

Assumption 2.2 (Coverage Decomposition) *There exist constants $\kappa_{\mathrm{trn}}, \kappa_{\mathrm{tst}} \geq 1$ and* marginal distributions *$\mathcal{D}_{\mathcal{X},1}, \mathcal{D}_{\mathcal{X},2}$ over $\mathcal{X}$, and $\mathcal{D}_{\mathcal{Y},1}, \mathcal{D}_{\mathcal{Y},2}$ over $\mathcal{Y}$, with their product measures $\mathcal{D}_{i\otimes j} := \mathcal{D}_{\mathcal{X},i} \otimes \mathcal{D}_{\mathcal{Y},j}$, such that the following is true for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$: (a)* ***Training Coverage:*** *for pairs $(i, j) \in \{(1, 1), (1, 2), (2, 1)\}$, $\frac{\mathrm{d}\mathcal{D}_{i\otimes j}(x,y)}{\mathrm{d}\mathcal{D}_{\mathrm{train}}(x,y)} \leq \kappa_{\mathrm{trn}}$, and (b)* ***Test Coverage:*** *$\frac{\mathrm{d}\mathcal{D}_{\mathrm{test}}(x,y)}{\sum_{i,j} \mathrm{d}\mathcal{D}_{i\otimes j}(x,y)} \leq \kappa_{\mathrm{tst}}$.*

The above condition means that the only part of $\mathcal{D}_{\mathrm{test}}$ not covered by $\mathcal{D}_{\mathrm{train}}$ is the samples (x, y) from $\mathcal{D}_{2\otimes 2}$. Thus, bilinear combinatorial extrapolation amounts to the *generalization* problem on these pairs. This condition represents the simplest case of the *Missing-Not-At-Random* (MNAR) matrix completion; see Figure 2 (a & b) and Appendix C for illustration and further discussion. As illustrated in Figure 2 (c), a unique completion requires that the top block has a rank equal to the other three blocks. Intuitively, we require an assumption that ensures that every feature which "appears" in $\mathcal{D}_{2\otimes 2}$ also "appears" in $\mathcal{D}_{1\otimes 1}$. We formalize this in the following assumption.

Assumption 2.3 (Change of Covariance) *There exists* $\kappa_{\mathrm{cov}} \geq 1$ *such that* $\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X},2}}[f^\star(x)f^\star(x)^\top] \preceq \kappa_{\mathrm{cov}} \cdot \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X},1}}[f^\star(x)f^\star(x)^\top]$ *and* $\mathbb{E}_{y\sim\mathcal{D}_{\mathcal{Y},2}}[g^\star(y)g^\star(y)^\top] \preceq \kappa_{\mathrm{cov}} \cdot \mathbb{E}_{y\sim\mathcal{D}_{\mathcal{Y},1}}[g^\star(y)g^\star(y)^\top]$.

Spectral assumptions. In addition to the above conditions, we require some control on the *richness* of the embeddings $f^\star, g^\star$. We shall assume that the covariances $\mathbf{\Sigma}_{f^\star} := \mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[f^\star(f^\star)^\top]$ and $\mathbf{\Sigma}_{g^\star} := \mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}[g^\star(g^\star)^\top]$ are trace-class operators on $\mathcal{H}$. We assume that we are in a basis of $\mathcal{H}$ for which $(f^\star, g^\star)$ are *balanced* in the following sense.

Assumption 2.4 (Balanced Basis) *The ground truth embeddings* $f^\star$ *and* $g^\star$ *are in an appropriate basis such that* $\mathbf{\Sigma}_{f^\star} = \mathbf{\Sigma}_{g^\star} =: \mathbf{\Sigma}^\star_{1\otimes 1}$ *are trace-class.*

The assumption $\mathbf{\Sigma}_{f^\star} = \mathbf{\Sigma}_{g^\star}$ may seem restrictive, but is achievable more-or-less without loss of generality by a change of basis (see Appendix L.3). Trace-class operators necessarily exhibit spectral decay. Hence, a key object throughout is the low-rank projections of our embeddings.

Definition 2.1 (Low-Rank Approximations) *Under Assumption 2.4, let* $\mathbf{P}^\star_k$ *denote the projection onto the top-k eigenspace of* $\mathbf{\Sigma}^\star_{1\otimes 1}$[1], $f^\star_k := \mathbf{P}^\star_k f^\star$, $g^\star_k := \mathbf{P}^\star_k g^\star$, *and* $h^\star_k(x,y) = \langle f^\star_k(x), g^\star_k(y)\rangle_{\mathcal{H}}$.

To take advantage of spectral decay, we shall reason extensively about the low-rank approximations $f^\star_k, g^\star_k$ to the ground-truth embeddings $f^\star, g^\star$. Our final condition ensures that low-rank approximations to $h^\star$ perform well on all the training data.

Assumption 2.5 *For all* $k \in \mathbb{N}$, $\mathbb{E}_{\mathcal{D}_{\mathrm{train}}}[(\langle f^\star_k, g^\star_k\rangle_{\mathcal{H}} - h^\star)^2] \leq \kappa_{\mathrm{apx}} \cdot \mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle f^\star_k, g^\star_k\rangle_{\mathcal{H}} - h^\star)^2]$.

We remark a sufficient (but strictly weaker) assumption which implies Assumption 2.5 is that $\mathcal{D}_{\mathrm{train}}$ is covered by the four-factor distributions in the sense that if $\mathrm{d}\mathcal{D}_{\mathrm{train}}/(\sum_{i,j=1}^2 \mathrm{d}\mathcal{D}_{i\otimes j}) \leq \tilde{\kappa}_{\mathrm{trn}}$; then one can check that Assumption 2.5 holds with $\kappa_{\mathrm{apx}} = 4\tilde{\kappa}_{\mathrm{trn}}\kappa_{\mathrm{cov}}^2$ if Assumption 2.3 holds. Note that such a case is easily satisfied by the standard matrix completion case, i.e., when the embeddings here are finite-dimensional. To make our results more concrete, we focus our attention on two classical regimes of spectral decay:

Assumption 2.6 (Spectral Decay) *There exist* $C, \gamma > 0$ *such that* ***either*** *(a)* $\lambda_i(\mathbf{\Sigma}^\star_{1\otimes 1}) \leq C i^{-(1+\gamma)}$ *(the "polynomial decay regime")* ***or*** *(b)* $\lambda_i(\mathbf{\Sigma}^\star_{1\otimes 1}) \leq C e^{-\gamma i}$ *(the "exponential decay regime").*

Notice that, for any $\gamma > 0$, the decay $\lambda_i(\mathbf{\Sigma}^\star_{1\otimes 1}) \leq C i^{-(1+\gamma)}$ does indeed ensure $\mathbf{\Sigma}^\star_{1\otimes 1}$ is trace-class.

1. When $\lambda_k(\mathbf{\Sigma}^\star_{1\otimes 1}) = \lambda_{k+1}(\mathbf{\Sigma}^\star_{1\otimes 1})$, $\mathbf{P}^\star_k$ is non-unique; in this case, assumptions stated in terms of $f^\star_k, g^\star_k$ can be chosen to hold for *any* valid choice of $\mathbf{P}^\star_k$.

Function approximation. As the spaces $\mathcal{X}, \mathcal{Y}$ are arbitrary, we require control of the statistical complexity of the embeddings $f_k^\star, g_k^\star$. We opt for the simplest possible assumption: for each $k \in \mathbb{N}$, the low-rank embeddings $f_k^\star, g_k^\star$ are captured by finite, uniformly bounded function classes.

Assumption 2.7 *Let B be the upper bound in Assumption 2.4. By inflating B if necessary, we assume that, for each $k \in \mathbb{N}$, there exist finite-cardinality function classes $\mathcal{F}_k \subseteq \{\mathcal{X} \to \mathbb{R}^k\}$ and $\mathcal{G}_k \subseteq \{\mathcal{Y} \to \mathbb{R}^k\}$ mapping into $\mathbb{R}^k$, such that (a) $\sup_{f\in\mathcal{F}_k} \sup_{x\in\mathcal{X}} \|f(x)\|_2 \le B$ and $\sup_{g\in\mathcal{G}_k} \sup_{y\in\mathcal{Y}} \|g(y)\|_2 \le B$, and (b) There exist some $(f, g) \in \mathcal{F}_k \times \mathcal{G}_k$ such that $\langle f(x), g(y)\rangle = \langle f_k^\star(x), g_k^\star(y)\rangle_{\mathcal{H}}$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. We define $\mathcal{M}_k := \log |\mathcal{F}_k||\mathcal{G}_k|$, and assume without loss of generality that $\mathcal{M}_k$ are non-decreasing as a function of $k \in \mathbb{N}$. Lastly, we also assume that for some $B > 0$, $\sup_{x\in\mathcal{X},y\in\mathcal{Y}} |\langle f^\star(x), g^\star(y)\rangle_{\mathcal{H}}| \le B^2$ and $\mathbb{P}_{(x,y,z)\sim\mathcal{D}_{\text{train}}}[|z| \le B^2] = 1$.*

Assumption 2.7 can easily be relaxed to accommodate infinite function classes with bounded covering numbers, classes with bounded Rademacher complexities (Bartlett and Mendelson, 2002), classes that satisfy more general tail conditions, and classes that only capture $f_k^\star, g_k^\star$ up to some error. As our bounds end up being polynomial in the log-cardinality of $\mathcal{M}_k$, we find Assumption 2.7 to be sufficient in capturing the essence of the function approximation setting.

3. Algorithms and Main Results

Additional notation. For any inner-product space $\mathcal{V}$ (e.g., $\mathcal{H}$ or $\mathbb{R}^r$ for $r \in \mathbb{N}$), we say (f, g) are $\mathcal{V}$-embeddings if $f : \mathcal{X} \to \mathcal{V}$, $g : \mathcal{Y} \to \mathcal{V}$; we say they are *isodimensional embeddings* if (f, g) are $\mathcal{V}$-embeddings for some $\mathcal{V}$. Given a probability distribution $\mathcal{D}$ on $(x, y) \in \mathcal{X} \times \mathcal{Y}$ pairs, we define the excess *risk* of the isodimensional $\mathcal{V}$-embeddings $(\hat{f}, \hat{g})$ as $\mathcal{R}(\hat{f}, \hat{g}; \mathcal{D}) := \mathbb{E}_{(x,y)\sim\mathcal{D}}[(\langle \hat{f}(x), \hat{g}(y)\rangle_{\mathcal{V}} - h^\star(x, y))^2]$. We often omit function dependence on (x, y) in expectations, i.e., writing it as $\mathcal{R}(f, g; \mathcal{D}) := \mathbb{E}_{\mathcal{D}}[(\langle g, f\rangle_{\mathcal{V}} - h^\star)^2]$ for short. We further define

$$\sigma_i(f, g) := \sigma_i \left(\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[ff^\top]^{\frac{1}{2}} \cdot \mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}[gg^\top]^{\frac{1}{2}} \right), \tag{3.1}$$

and if (f, g) are $\mathbb{R}^r$-embeddings, we say (f, g) are ***full-rank*** if $\sigma_r(f, g) > 0$. We adopt the shorthand $\boldsymbol{\sigma}_i^\star := \lambda_i(\boldsymbol{\Sigma}_{1\otimes1}^\star)$, and for $q \ge 1$, $\mathsf{tail}_q^\star(k) := \sum_{i>k} \lambda_i(\boldsymbol{\Sigma}_{1\otimes1}^\star)^q = \sum_{i>k} (\boldsymbol{\sigma}_i^\star)^q$. We use $a \lesssim b$ to denote $a \le c \cdot b$ for some absolute constant c; we use $a \lesssim_\star b$ to denote $a \le c \cdot b$ for some c that is at most polynomial in the problem constants $\kappa_{\text{cov}}, \kappa_{\text{trn}}, \kappa_{\text{tst}}, \kappa_{\text{apx}}$ in Assumptions 2.2, 2.3 and 2.5.

3.1. A meta-theorem for bilinear combinatorial extrapolation

We now provide a meta-theorem on the risk bound for bilinear combinatorial extrapolation. The bound depends on an upper bound ϵ_{trn} on the risk of the learned embedding $(\hat{f}, \hat{g})$ on the training distribution $\mathcal{D}_{\text{train}}$, on $\epsilon_{1\otimes1}$ that upper-bounds the risk on the top-block distribution $\mathcal{D}_{1\otimes1}$, as well as on $\sigma_r(\hat{f}, \hat{g})$ defined in Eq. (3.1).

Definition 3.1 (α-Conditioned & $(\epsilon_{\text{trn}}, \epsilon_{1\otimes1})$-Accurate Embeddings) *Given $\alpha \ge 1$ and $\epsilon_{\text{trn}}, \epsilon_{1\otimes1} > 0$, we say $\mathbb{R}^r$-embeddings $(\hat{f}, \hat{g})$ are α-conditioned if $\sigma_r(\hat{f}, \hat{g})^2 \ge (\boldsymbol{\sigma}_r^\star)^2/\alpha$ and $(\epsilon_{\text{trn}}, \epsilon_{1\otimes1})$-accurate if $\mathcal{R}(\hat{f}, \hat{g}; \mathcal{D}_{\text{train}}) \le \epsilon_{\text{trn}}^2$ and $\inf_{r'\ge r} \mathcal{R}_{[r']}(\hat{f}, \hat{g}; \mathcal{D}_{1\otimes1}) \le \epsilon_{1\otimes1}^2$[2], where $\mathcal{R}_{[s]}$ is the excess risk relative to $h_s^\star = \langle f_s^\star, g_s^\star\rangle$, evaluated on $\mathcal{D}_{1\otimes1}$:*

$$\mathcal{R}_{[s]}(\hat{f}, \hat{g}; \mathcal{D}_{1\otimes1}) := \mathbb{E}_{(x,y)\sim\mathcal{D}_{1\otimes1}}[(\langle \hat{f}(x), \hat{g}(y)\rangle - h_s^\star(x, y))^2]. \tag{3.2}$$

2. Because $h_{r'}^\star$ converges to $h^\star$ in $\mathcal{L}_2(\mathcal{D}_{1\otimes1})$ as $r' \to \infty$, $\inf_{r'\ge r} \mathcal{R}_{[r']}(\hat{f}, \hat{g}; \mathcal{D}_{1\otimes1}) \le \mathcal{R}(\hat{f}, \hat{g}; \mathcal{D}_{1\otimes1})$.

Theorem 2 (Main Risk Bound) *Given $\alpha \ge 1$, suppose $(\hat{f}, \hat{g})$ are α-conditioned and $(\epsilon_{\mathrm{trn}}, \epsilon_{1\otimes 1})$-accurate $\mathbb{R}^r$-embeddings, where $r \le \sigma_1^\star/(40\epsilon_{1\otimes 1})$. Then under Assumptions 2.1 to 2.4, we have*

$$\mathcal{R}(\hat{f}, \hat{g}; \mathcal{D}_{\mathrm{test}}) \lesssim_\star \left(r^4\epsilon_{1\otimes 1}^2 + \alpha r^2(\boldsymbol{\sigma}_{r+1}^\star)^2 + \mathsf{tail}_1^\star(r)^2\right) + \alpha\left(\frac{r^6\epsilon_{1\otimes 1}^4 + \epsilon_{\mathrm{trn}}^4 + \mathsf{tail}_2^\star(r)^2}{(\boldsymbol{\sigma}_r^\star)^2}\right) \quad (3.3)$$

Moreover, the condition $\epsilon_{1\otimes 1}^2 \le (1-\alpha^{-1})(\boldsymbol{\sigma}_r^\star)^2$ ensures that $\sigma_r(\hat{f}, \hat{g})^2 \ge (\boldsymbol{\sigma}_r^\star)^2/\alpha$.

Theorem 2 is proved in Section 4; its implications are best understood through its instantiation below. Here, we note an important point that the dependence on the "top-block" error $\epsilon_{1\otimes 1}^2$ is scaled up by polynomial factors of r. It is from this fact that the benefits of double-stage ERM derive.

3.2. Single-stage empirical risk minimization

A natural algorithm is to fix a target rank $r \in \mathbb{N}$ and compute a *single-stage empirical risk minimizer*, i.e., to find $(\hat{f}_{\mathrm{SS}}, \hat{g}_{\mathrm{SS}}) \in \arg\min_{f\in\mathcal{F}_r, g\in\mathcal{G}_r} \sum_{i=1}^n (\langle f(x_i), g(y_i)\rangle - z_i)^2$, where we draw $(x_i, y_i, z_i) \overset{\text{i.i.d}}{\sim} \mathcal{D}_{\mathrm{train}}$ and function classes $\mathcal{F}_r, \mathcal{G}_r$ are as given in Assumption 2.7. By combining Theorem 2, the fact that $\epsilon_{1\otimes 1} \le \kappa_{\mathrm{trn}}\epsilon_{\mathrm{trn}}$ by Assumption 2.2, and standard statistical learning arguments to bound ϵ_{trn}, we can obtain the following guarantee (whose proof is given in Appendix E.1).

Theorem 3 *Fix $\delta \in (0,1)$, $\alpha \ge 1$. Under Assumptions 2.1 to 2.5 and 2.7, with probability at least $1-\delta$, **if** $(\hat{f}_{\mathrm{SS}}, \hat{g}_{\mathrm{SS}})$ are α-conditioned, then $\mathcal{R}(\hat{f}_{\mathrm{SS}}, \hat{g}_{\mathrm{SS}}; \mathcal{D}_{\mathrm{test}}) \lesssim_\star \textsc{Err}_{\mathrm{SS}}(r,n,\delta)$ with*

$$\textsc{Err}_{\mathrm{SS}}(r,n,\delta) := \alpha\textsc{ApxErr}_{\mathrm{SS}}(r) + r^4\textsc{StatErr}_{\mathrm{SS}}(r,n,\delta) + \tfrac{\alpha r^6}{(\sigma_r^\star)^2}\textsc{StatErr}_{\mathrm{SS}}(r,n,\delta)^2,$$

where $\textsc{StatErr}_{\mathrm{SS}}(r,n,\delta) := \frac{B^4(\mathcal{M}_r + \log(1/\delta))}{n}$ captures the statistical error, and where $\textsc{ApxErr}_{\mathrm{SS}}(r) := r^4\mathsf{tail}_2^\star(r) + \mathsf{tail}_1^\star(r)^2 + r^2(\boldsymbol{\sigma}_{r+1}^\star)^2 + \frac{r^6\cdot\mathsf{tail}_2^\star(r)^2}{(\sigma_r^\star)^2}$. Moreover, under Assumption 2.6,

$$\textsc{ApxErr}_{\mathrm{SS}}(r) \lesssim_\star \begin{cases} C^2(1+\gamma^{-1})^2 r^{6-2\gamma} & \textit{(polynomial decay)} \\ C^2 r^6(\gamma^{-1}+r)^2 e^{-2\gamma r} & \textit{(exponential decay).} \end{cases} \quad (3.4)$$

To the best of our knowledge, Theorem 3 is the first result that establishes bilinear combinatorial extrapolation for (sufficiently fast) polynomial decay, $\gamma > 3$. However, the theorem has two weaknesses: first, our upper bound on $\textsc{ApxErr}_{\mathrm{SS}}(r)$ does not decay to zero under polynomial decay with $\gamma \le 3$. Second, α depends on the ratio of $\sigma_r(\hat{f}_{\mathrm{SS}}, \hat{g}_{\mathrm{SS}})$ to $\boldsymbol{\sigma}_r^\star$, and we do not (yet) know a way to control this quantity, except in the special case when $(\boldsymbol{\sigma}_r^\star)^2 > 2\kappa_{\mathrm{apx}}\kappa_{\mathrm{trn}}\mathsf{tail}_2^\star(r)$ (see Remark E.1). To see the culprit, consider the (somewhat trivializing) case where $\mathcal{D}_{\mathrm{train}} = \mathcal{D}_{1\otimes 1}$. Then $\epsilon_{1\otimes 1}^2 = \epsilon_{\mathrm{trn}}^2$, and by the Eckhart-Young theorem, $\epsilon_{1\otimes 1}^2 \ge \mathcal{R}(f_r^\star, g_r^\star; \mathcal{D}_{1\otimes 1}) = \mathsf{tail}_2^\star(r)$. In this case, we have (a) the upper bound in Theorem 2 is no better than $\frac{r^6\mathsf{tail}_2^\star(r)^2}{(\boldsymbol{\sigma}_r^\star)^2}$, which scales like $r^{6-2\gamma}$ for polynomial spectral decay, and (b) unless $\mathsf{tail}_2^\star(r) < (\boldsymbol{\sigma}_r^\star)^2$, we can not use Theorem 2 to ensure a lower bound on α. These issues exactly arise from our consideration of the *modest* spectral decay case, and would not cause trouble in a standard MNAR matrix completion case with a *sharp* spectral cutoff. In the next section, we present a more involved algorithm to circumvent these limitations.

3.3. Double-stage empirical risk minimization (ERMDS)

Given a desired rank cutoff r_{cut}, we also develop a *Double-Stage ERM* (ERMDS) algorithm, which learns $\mathbb{R}^{\hat{r}}$-embeddings $(\hat{f}_{\mathrm{DS}}, \hat{g}_{\mathrm{DS}})$ for a data-dependent $\hat{r}$ such that $\epsilon_{1\otimes 1} \ll r_{\mathrm{cut}}^3 \mathsf{tail}_2^\star(r_{\mathrm{cut}})$, for which $\mathsf{tail}_q^\star(\hat{r})$ is not much larger than $\mathsf{tail}_q^\star(r_{\mathrm{cut}})$. Hence, we can instantiate Theorem 2 with $r = r_{\mathrm{cut}}$, but *without suffering from the prefactor powers of* r_{cut} *premultiplying* $\epsilon_{1\otimes 1}$. Our procedure relies on a slightly stronger oracle:

Assumption 3.1 (Unlabeled $\mathcal{D}_{1\otimes 1}$-Oracle) *In addition to being able to sample i.i.d. data $(x, y, z) \sim \mathcal{D}_{\mathrm{train}}$, we can also sample* unlabeled *i.i.d. data $(x, y) \sim \mathcal{D}_{1\otimes 1}$.*

Moreover generally, Appendix E.3 shows that $\mathcal{D}_{1\otimes 1}$ can be replaced with any product distribution on $\mathcal{X} \times \mathcal{Y}$ with bounded density with respect to $\mathcal{D}_{1\otimes 1}$.

We summarize the details of ERMDS in Algorithm 1. The algorithm has three spectral parameters: an overparametrized rank p, a spectral cutoff σ_{cut}, and a rank cutoff r_{cut}. We first train high-dimensional $\mathbb{R}^p$-embeddings $(\tilde{f}, \tilde{g})$, where ideally $p \gg r_{\mathrm{cut}}$ is sufficiently large so that $\mathsf{tail}_2^\star(p) \ll r_{\mathrm{cut}}^6 \mathsf{tail}_2^\star(r_{\mathrm{cut}})^2/(\sigma_{r_{\mathrm{cut}}}^\star)^2$. We then perform an SVD-approximation of $(\tilde{f}, \tilde{g})$, first by estimating their covariance matrices, and then using these matrices to perform dimension reduction (the routine DIMREDUCE in Algorithm 2). The dimension reduction routine reduces to a rank-at-most-$\hat{r} \le r_{\mathrm{cut}}$ predictor $\hat{h}_{\mathrm{RED}}$, where $\hat{r}$ is determined by the estimated covariances matrices and spectral cutoff σ_{cut}. In a final distillation phase, we learn $\mathbb{R}^{\hat{r}}$-embeddings $(\hat{f}_{\mathrm{DS}}, \hat{g}_{\mathrm{DS}})$ by regularizing the supervised training error on labeled samples from $\mathcal{D}_{\mathrm{train}}$ with empirical risk on samples $(x', y', \hat{h}_{\mathrm{RED}}(x', y'))$, where (x', y') are drawn from $\mathcal{D}_{1\otimes 1}$ and labeled by $\hat{h}_{\mathrm{RED}}$. This is similar to the process of distillation in Hinton et al. (2015), where a larger deep network is used to supervise the learning of a smaller one. Algorithm 1 enjoys the following guarantee, the detailed version of which is given in Appendix E.2 and proved in Appendix F.

Theorem 4 *For any $r_{\mathrm{cut}} \gtrsim_\star \mathrm{poly}(C/\boldsymbol{\sigma}_1^\star, \gamma^{-1})$ and $\epsilon > 0$ and $\delta > 0$, there exists a choice of σ_{cut}, $p \lesssim_\star (r_{\mathrm{cut}})^c$ for some universal $c > 0$, and sample sizes $n_1, n_2, n_3, n_4 \lesssim_\star \mathrm{poly}(p, \mathcal{M}_p, \log(1/\delta), B, \epsilon^{-2})$, such that, Algorithm 1 with $\lambda = r_{\mathrm{cut}}^4$ and $\mu = B^2/n_1$ satisfies that with probability at least $1 - \delta$,*

$$\mathcal{R}(\hat{f}_{\mathrm{DS}}, \hat{g}_{\mathrm{DS}}; \mathcal{D}_{\mathrm{test}}) \lesssim_\star \epsilon^2 + C^2(1 + \gamma^{-2}) \begin{cases} r_{\mathrm{cut}}^{-2\gamma} & \textit{(polynomial decay)} \\ e^{-2\gamma r_{\mathrm{cut}}} & \textit{(exponential decay)} \end{cases}.$$

Proof Sketch of Theorem 4. We first show, by analogy to Theorem 3, that $\mathcal{R}(\tilde{f}, \tilde{g}; \mathcal{D}_{1\otimes 1}) \lesssim_\star \mathsf{tail}_2^\star(p) + o(n_1)$. We then learn a data-dependent $\hat{r}$, chosen by the DIMREDUCE procedure, so as to satisfy $\boldsymbol{\sigma}_{\hat{r}}^\star \gtrsim \sigma_{\mathrm{cut}}$, and to have lower bounded relative singular-value gap $(\boldsymbol{\sigma}_{\hat{r}}^\star - \boldsymbol{\sigma}_{\hat{r}+1}^\star)/\boldsymbol{\sigma}_{\hat{r}}^\star \gtrsim 1/r_{\mathrm{cut}}$. We then argue that $\hat{h}_{\mathrm{RED}}$ constructed in Line 4 is the correct analogue rank-$\hat{r}$ SVD approximation of $\langle \tilde{f}, \tilde{g} \rangle$, just as $h_{\hat{r}}^\star$ is the best rank-$\hat{r}$ approximation of $h^\star$ on $\mathcal{D}_{1\otimes 1}$. We then use our novel relative-gap SVD perturbation bound (Theorem 1) and limiting arguments to show that our bound $\mathcal{R}(\tilde{f}, \tilde{g}; \mathcal{D}_{1\otimes 1}) = \mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle \tilde{f}, \tilde{g} \rangle - h^\star)^2]$ implies $\mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\hat{h}_{\mathrm{RED}} - h_{\hat{r}}^\star)^2] \lesssim_\star r_{\mathrm{cut}}^2 (\mathsf{tail}_2^\star(p) + o(n_1))$. The factor of r_{cut}^2 arises from the relative singular-value gap at $\hat{r}$ mentioned above. In addition, we argue that DIMREDUCE chooses $\hat{r}$ large enough such that the tails $\mathsf{tail}_q^\star(\hat{r})$ and $\mathsf{tail}_q^\star(r_{\mathrm{cut}})$ are close. Finally, we show that the distillation step with a large λ forces $\langle \hat{f}_{\mathrm{DS}}, \hat{g}_{\mathrm{DS}} \rangle$ to be close to $\hat{h}_{\mathrm{RED}} \approx h_{\hat{r}}^\star$ on $\mathcal{D}_{1\otimes 1}$; this ensures that we can invoke Theorem 2 with $\epsilon_{1\otimes 1}^2 = \mathcal{R}_{[\hat{r}]}(\hat{f}_{\mathrm{DS}}, \hat{g}_{\mathrm{DS}}; \mathcal{D}_{1\otimes 1}) \approx \mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\hat{h}_{\mathrm{RED}} - h_{\hat{r}}^\star)^2] \lesssim_\star r_{\mathrm{cut}}^2 (\mathsf{tail}_2^\star(p) + o(n_1))$. In particular, by making $p \gg r_{\mathrm{cut}} \ge \hat{r}$, we can ensure $\hat{r}^3 \epsilon_{1\otimes 1}^2 \le r_{\mathrm{cut}}^3 \epsilon_{1\otimes 1}^2 \ll \mathsf{tail}_2^\star(r_{\mathrm{cut}})$, as desired. ■

Algorithm 1 Double-Stage ERM (ERMDS)

1: **Input:** Sample sizes $n_1, \dots, n_4$; over-parameterized rank p, under-parameterized cutoff r_{cut}, parameter σ_{cut}, regularization parameters $\mu, \lambda > 0$.

2: **Overparametrized Training.** Sample n_1 labeled triples $\{(x_{1,i}, y_{1,i}, z_{1,i}\}_{i\in[n_1]}$ i.i.d. from $\mathcal{D}_{\text{train}}$, and set

$$(\tilde{f}, \tilde{g}) \in \underset{(f,g)\in\mathcal{F}_p\times\mathcal{G}_p}{\arg\min} \frac{1}{n_1}\sum_{i=1}^{n_1}(\langle f(x_{1,i}), g(y_{1,i})\rangle - z_{1,i})^2.$$

3: **Covariance Estimation.** Sample n_2 *unlabeled* examples $\{(x_{2,i}, y_{2,i})\}_{i\in[n_2]} \sim \mathcal{D}_{1\otimes 1}$, and define covariance matrices $\hat{\mathbf{\Sigma}}_{\tilde{f}} := \frac{1}{n_2}\sum_{i=1}^{n_2}\tilde{f}(x_{2,i})\tilde{f}(x_{2,i})^\top$, $\hat{\mathbf{\Sigma}}_{\tilde{g}} := \frac{1}{n_2}\sum_{i=1}^{n_2}\tilde{g}(y_{2,i})\tilde{g}(y_{2,i})^\top$.

4: **Dimension Reduction.**

$$(\hat{r}, \hat{\mathbf{Q}}_{\hat{r}}) \leftarrow \textsc{DimReduce}(\hat{\mathbf{\Sigma}}_{\tilde{f}} + \mu\mathbf{I}_p, \hat{\mathbf{\Sigma}}_{\tilde{g}} + \mu\mathbf{I}_p, r_{\text{cut}}, \sigma_{\text{cut}}),$$

and $\hat{h}_{\text{RED}}(x, y) := \langle\tilde{f}(x), \hat{\mathbf{Q}}_r \cdot \tilde{g}(y)\rangle$

5: **Distillation.** Sample n_3 *labeled* examples $\{(x_{3,i}, y_{3,i}, z_{3,i})\}_{i\in[n_3]} \sim \mathcal{D}_{\text{train}}$ and n_4 *unlabeled* samples $\{(x_{4,i}, y_{4,i})\}_{i\in[n_4]} \sim \mathcal{D}_{1\otimes 1}$. Define the losses $\hat{L}_{(3)}(f, g) = \frac{1}{n_3}\sum_{i=1}^{n_3}(\langle f(x_{3,i}), g(y_{3,i})\rangle - z_{3,i})^2$ and $\hat{L}_{(4)}(f, g) = \frac{1}{n_4}\sum_{i=1}^{n_4}(\langle f(x_{4,i}), g(y_{4,i})\rangle - \hat{h}_{\text{RED}}(x_{4,i}, y_{4,i}))^2$, and select

$$(\hat{f}_{\text{DS}}, \hat{g}_{\text{DS}}) \in \underset{(f,g)\in\mathcal{F}_{\hat{r}}\times\mathcal{G}_{\hat{r}}}{\arg\min} \hat{L}_{(3)}(f, g) + \lambda\hat{L}_{(4)}(f, g).$$

Algorithm 2 $\textsc{DimReduce}(\mathbf{X}, \mathbf{Y}, r_0, \sigma_0)$

1: **Input:** $\mathbf{X}, \mathbf{Y} \succ 0$, $r_0 \in \mathbb{N}$, σ_0.

2: Compute $\mathbf{W} := \mathbf{X}^{\frac{1}{2}}(\mathbf{X}^{\frac{1}{2}}\mathbf{Y}\mathbf{X}^{\frac{1}{2}})^{-\frac{1}{2}}\mathbf{X}^{\frac{1}{2}}$.

3: Compute $\mathbf{\Sigma} := \mathbf{W}^{\frac{1}{2}}\mathbf{Y}\mathbf{W}^{\frac{1}{2}}$; set $r \leftarrow \max\left\{r \in [r_0] : \sigma_r(\mathbf{\Sigma}) \geq \sigma_0,\ \sigma_r(\mathbf{\Sigma}) - \sigma_{r+1}(\mathbf{\Sigma}) \geq \frac{\sigma_r(\mathbf{\Sigma})}{r_0}\right\}$.

4: Let $\mathbf{P}_r$ denote the projection onto the top r eigenvectors of $\mathbf{\Sigma}$.

5: Return $(r, \mathbf{Q}_r)$, where $\mathbf{Q}_r \leftarrow \mathbf{W}^{-\frac{1}{2}}\mathbf{P}_r\mathbf{W}^{\frac{1}{2}}$.

4. Proof Overview of the Meta-Theorem – Theorem 2

In this section, we provide an overview of the key techniques in our proof of the main result Theorem 2, which is completed in Appendix M.7. As noted above, the proofs of Theorems 3 and 4 are given in Appendix F.

4.1. Reformulation as matrix completion

To explain the intuition behind our proofs, it helps to consider the case when $|\mathcal{X}|$ and $|\mathcal{Y}|$ are finite, with elements $\{x_1, \dots, x_n\}$ and $\{y_1, \dots, y_m\}$. For $i, j \in \{1, 2\}$, we define the probabilities $\mathsf{p}_{i,\ell} = \mathbb{P}_{\mathcal{D}_{\mathcal{X},i}}[x = x_\ell]$ and $\mathsf{q}_{j,k} = \mathbb{P}_{\mathcal{D}_{\mathcal{Y},j}}[y = y_k]$. Because of the finite support of the distributions, we can regard any $\mathcal{H}$-embeddings (f, g) (including $(f^\star, g^\star)$) as embeddings into $\mathcal{H} = \mathbb{R}^d$, $d = \max\{n, m\}$, appending zeros if necessary. Define matrices $\mathbf{A}_i(f) \in \mathbb{R}^{n\times d}$ and $\mathbf{B}_j(g) \in \mathbb{R}^{m\times d}$ by assigning the rows to the scaled values of the embeddings $\mathbf{A}_i(f)[\ell, :] =$

$\sqrt{\mathsf{p}_{i,\ell}} f(x_\ell)^\top, \quad \mathbf{B}_j(g)[k,:] = \sqrt{\mathsf{q}_{j,k}} g(y_k)^\top$, and define $\mathbf{M}_{i\otimes j}(f,g) = \mathbf{A}_i(f)\mathbf{B}_j(g)^\top$. Each matrix $\mathbf{M}_{i\otimes j}(f,g)$ can be thought of as a look-up table, where $\mathbf{M}_{i\otimes j}(f,g)[\ell,k] = \sqrt{\mathsf{p}_{i,\ell}\mathsf{q}_{j,k}}\langle f(x_\ell), g(y_k)\rangle$ is the prediction of $\langle f,g\rangle$, scaled by the square root probability of x_ℓ and y_k. This reformulation yields the following equivalences, verified in Lemma J.2.

Lemma 4.1 *The following identities hold: (a)* $\mathcal{R}(f,g;\mathcal{D}_{i\otimes j}) = \|\mathbf{M}_{i\otimes j}(f,g) - \mathbf{M}_{i\otimes j}(f^\star,g^\star)\|_{\mathrm{F}}^2$ *and (b)* $\mathbb{E}_{\mathcal{D}_{\mathcal{X},i}}[ff^\top] = \mathbf{A}_i(f)^\top\mathbf{A}_i(f)$, *and similarly for* $\mathbb{E}_{\mathcal{D}_{\mathcal{Y},j}}[gg^\top] = \mathbf{B}_j(g)^\top\mathbf{B}_j(g)$.

Most of our technical results are easiest to establish for the matrix factorization formulation, and then are generalized to accommodate arbitrary distributions via some careful limiting arguments.

4.2. Balancing and singular value decomposition

Note that for any isodimensional embedding (f,g), any embedding $(f',g') = (\mathbf{T}^{-\top}f, \mathbf{T}g)$ for some invertible operator $\mathbf{T}$ satisfies $\langle f',g'\rangle \equiv \langle f,g\rangle$. We thus focus on **balanced** embeddings.

Definition 4.1 (Balanced Embeddings) *We say any isodimensional embeddings* (f,g) *are* balanced *if the covariance* $\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[ff^\top] = \mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}[gg^\top]$*; given* $\mathbf{M} \in \mathbb{R}^{n\times m}$, *we say* $(\mathbf{A},\mathbf{B}) \in \mathbb{R}^{n\times d} \times \mathbb{R}^{m\times d}$ *is a balanced factorization of* $\mathbf{M}$ *if* $\mathbf{M} = \mathbf{A}\mathbf{B}^\top$ *and* $\mathbf{A}^\top\mathbf{A} = \mathbf{B}^\top\mathbf{B}$.

Balancing is *orthogonally invariant*: for any orthogonal transformation $\mathbf{U}$ (of appropriate dimension), (f,g) are balanced if and only if $(\mathbf{U}f,\mathbf{U}g)$ are. Similarly, $(\mathbf{A},\mathbf{B})$ is a balanced factorization of $\mathbf{M}$ if and only if $(\mathbf{AU},\mathbf{BU})$ is. Moreover, when distributions are discrete, (f,g) are balanced if and only if $(\mathbf{A}_1(f),\mathbf{B}_1(g))$ is a balanced factorization of $\mathbf{M}_{1\otimes 1}(f,g)$. The matrix factorization interpretation reveals many useful properties of balanced embeddings/factorizations.

Lemma 4.2 *Suppose* (f,g) *are balanced* $\mathcal{H}$*-embeddings, and* $\mathcal{X},\mathcal{Y}$ *are finite spaces. Let* $\mathbf{P}_{[r]}$ *denote the orthogonal projection onto the top-*r *eigenvectors of* $\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[ff^\top] = \mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}[gg^\top]$. *Then, (a)* $\mathbf{A}_1(\mathbf{P}_{[r]}f)$ *is equal to the rank-*r *SVD approximation of* $\mathbf{A}_1(f)$, *and similarly for* $\mathbf{B}_1(\mathbf{P}_{[r]}g)$ *and* $\mathbf{B}_1(g)$*; (b)* $\mathbf{M}_{1\otimes 1}(\mathbf{P}_{[r]}f,\mathbf{P}_{[r]}g)$ *is equal to the rank-*r *SVD approximation of* $\mathbf{M}_{1\otimes 1}(f,g)$*; and (c) For any* $i \ge 1$, $\sigma_i(\mathbf{M}_{1\otimes 1}(f,g)) = \sigma_i(\mathbf{A}_1(f))^2 = \sigma_i(\mathbf{B}_1(g))^2$.

This lemma is a partial statement of a more complete result, Lemma J.2, given in the appendix. Importantly, the appropriate SVD approximation for balanced embeddings can be computed by projecting onto the top eigenvectors of the covariance matrix of f (or equivalently, of g). Via limiting arguments in Appendix J, this characterization can be extended to the case where spaces $\mathcal{X},\mathcal{Y}$ are continuous, and where the covariances can be computed from samples. One can also construct a balanced embedding from a non-balanced one. This is most succinctly stated as finite-dimensional *full-rank* embeddings; a more extensive statement and its proof are given in Appendix L.3.

Lemma 4.3 *For full-rank* $\mathbb{R}^r$*-embeddings* $(\hat{f},\hat{g})$, *there exists a unique* $\mathbf{T} \in \mathbb{S}^r_>$ *for which* $(\tilde{f},\tilde{g}) = (\mathbf{T}^{-1}\hat{f}, \mathbf{T}\hat{g})$ *is balanced; moreover,* $\sigma_r(\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[\tilde{f}\tilde{f}^\top]) = \sigma_r(\mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}[\tilde{g}\tilde{g}^\top]) = \sigma_r(\hat{f},\hat{g})$.

4.3. Error decomposition

We now specify our error decomposition result. First, we describe embeddings (f, g) into $\mathcal{H}$ which are consistent with the learned embedding $(\hat{f}, \hat{g})$, but are balanced, and are aligned with the top-k eigenspace of $\mathbf{\Sigma}^\star_{1\otimes 1}$. This allows us to reason about the differences between $f - f^\star$ and $g - g^\star$.

Definition 4.2 (Aligned Proxies) *We say $\iota_r : \mathbb{R}^r \to \mathcal{H}$ is an* isometric inclusion *if it preserves inner products, i.e., $\langle v, w\rangle = \langle \iota_r(v), \iota_r(w)\rangle_{\mathcal{H}}$. Fix a dimension $r \in \mathbb{N}$, and some $k \in \mathbb{N}$, and let $\hat{f} : \mathcal{X} \to \mathbb{R}^r$ and $\hat{g} : \mathcal{Y} \to \mathbb{R}^r$ be full-rank. We say (f, g) are* aligned k-proxies *for $(\hat{f}, \hat{g})$ if: (a) $f = (\iota_r \circ \mathbf{T}^{-1})\hat{f}$, $g = (\iota_r \circ \mathbf{T})\hat{g}$, where $\iota_r : \mathbb{R}^r \to \mathcal{H}$ is an isometric inclusion, and $\mathbf{T}$ is the balancing operator of Lemma 4.3, and (b) for[3] $\mathbf{P}^\star_k$ being the projection onto the top k-eigenvectors of $\mathbf{\Sigma}^\star_{1\otimes 1}$, we have*

$$\mathrm{range}(\mathbf{P}^\star_k) \subseteq \mathrm{range}(\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[f f^\top]). \tag{4.1}$$

Definition 4.3 (Key Error Terms) *Given aligned k-proxies (f, g) of $(\hat{f}, \hat{g})$, we define*

$$\mathbf{\Delta}_0(f, g, k) := \max\left\{\mathbb{E}_{\mathcal{D}_{1\otimes 1}}\langle f^\star_k, g^\star_k - g\rangle^2,\ \mathbb{E}_{\mathcal{D}_{1\otimes 1}}\langle f^\star_k - f, g^\star_k\rangle^2\right\} \qquad \text{(weighted error)}$$
$$\mathbf{\Delta}_1(f, g, k) := \max\left\{\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}\|f^\star_k - f\|^2,\ \mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}\|g^\star_k - g\|^2\right\} \qquad \text{(unweighted error)}$$
$$\mathbf{\Delta}_{\mathrm{train}} := \mathcal{R}(f, g; \mathcal{D}_{\mathrm{train}}). \qquad \text{(training error)}$$

Proposition 4.1 (Main Error Decomposition) *Suppose Assumptions 2.1 to 2.4 hold. Fix $r \geq k > 0$, let (f, g) be aligned k-proxies for full-rank $\mathbb{R}^r$-embeddings $(\hat{f}, \hat{g})$. Define the parameter $\sigma^2 := \min\{\sigma_r(\hat{f}, \hat{g})^2, \mathsf{tail}^\star_2(k) + \mathbf{\Delta}_0(f, g, k) + \mathbf{\Delta}_{\mathrm{train}}\}$. Then,*

$$\mathcal{R}(\hat{f}, \hat{g}; \mathcal{D}_{\mathrm{test}}) = \mathcal{R}(f, g; \mathcal{D}_{\mathrm{test}}) \lesssim_\star (\mathbf{\Delta}_1(f, g, k))^2 + \frac{1}{\sigma^2}(\mathsf{tail}^\star_2(k) + \mathbf{\Delta}_0(f, g, k) + \mathbf{\Delta}_{\mathrm{train}})^2.$$

The unweighted error, $\mathbf{\Delta}_1(f, g, k)$, measures how close the aligned proxies (f, g) track the best rank-k approximation $(f^\star_k, g^\star_k)$. The weighted error, $\mathbf{\Delta}_0(f, g, k)$, does the same, but only along the directions of $f^\star_k$ and $g^\star_k$ which have spectral decay. Thus, one can expect the weighted errors to be considerably smaller. This is important, because we pay for $\frac{1}{\sigma^2}(\mathsf{tail}^\star_2(k) + \mathbf{\Delta}_0(f, g, k) + \mathbf{\Delta}_{\mathrm{train}})^2$, so we need to ensure that $\mathbf{\Delta}_0(f, g, k)^2 \ll \sigma^2$ in order to achieve consistent recovery. Proposition 4.1 is proved, along with a more general statement, in Appendix M.

4.4. From error-terms to factor recovery, and concluding the proof of Theorem 2

We now aim for upper bounds on $\mathbf{\Delta}_i(f, g, k), i \in \{0, 1\}$ in terms of the parameter $\epsilon_{1\otimes 1}$ in Definition 3.1. In this section, we expose how to obtain the bound for distributions with finite support. This result is equivalent to a guarantee for *factor-recovery* in matrix completion. In the sequel, we adopt the finite-support setting, so that $\mathcal{H} = \mathbb{R}^d$. Define $\mathbf{A}^\star := \mathbf{A}_1(f^\star), \mathbf{B}^\star := \mathbf{B}_1(g^\star)$ so that $\mathbf{M}^\star := \mathbf{M}_{1\otimes 1}(f^\star, g^\star) = \mathbf{A}^\star(\mathbf{B}^\star)^\top$, and similarly set $\hat{\mathbf{A}} = \mathbf{A}_1(f), \hat{\mathbf{B}} = \mathbf{B}_1(g), \hat{\mathbf{M}} = \mathbf{M}_{1\otimes 1}(f, g) = \hat{\mathbf{A}}\hat{\mathbf{B}}^\top$. We further let $\mathbf{A}^\star_{[k]}, \mathbf{B}^\star_{[k]}$ denote the rank-k approximation of $\mathbf{A}^\star, \mathbf{B}^\star$, defined

3. In case of non-uniqueness, any choice of projection works.

formally in Eq. (I.3). Lastly, for an orthogonal matrix $\mathbf{R} \in \mathbb{O}(d)$, we define the following error terms

$$\Delta_0(\mathbf{R}, k) = \|(\mathbf{A}^\star_{[k]} - \hat{\mathbf{A}}\mathbf{R})(\mathbf{B}^\star_{[k]})^\top\|_{\mathrm{F}}^2 \vee \|\mathbf{A}^\star_{[k]}(\mathbf{B}^\star_{[k]} - \hat{\mathbf{B}}\mathbf{R})^\top\|_{\mathrm{F}}^2 \quad (4.2)$$

$$\Delta_1(\mathbf{R}, k) = \|\mathbf{A}^\star_{[k]} - \hat{\mathbf{A}}\mathbf{R}\|_{\mathrm{F}}^2 \vee \|\mathbf{B}^\star_{[k]} - \hat{\mathbf{B}}\mathbf{R}\|_{\mathrm{F}}^2. \quad (4.3)$$

One can then check (see Appendix J.3) that for the matrices defined above and $i \in \{0, 1\}$, we have $\Delta_i(\mathbf{R}, k) = \boldsymbol{\Delta}_i(\mathbf{R}^\top f, \mathbf{R}^\top g, k)$. Here, the matrix $\mathbf{R}$ allows us to rotate embeddings (f, g) to minimize the factor error. In sum, we have shown that the error terms in Proposition 4.1 are corresponding to the recovery of factors in matrix completion. We now establish an error bound on these factory-recovery terms, which is the main technical effort of this paper.

Theorem 5 *Let $\mathbf{A}^\star, \hat{\mathbf{A}} \in \mathbb{R}^{n\times d}$, $\mathbf{B}^\star, \hat{\mathbf{B}} \in \mathbb{R}^{m\times d}$, and suppose $(\mathbf{A}^\star, \mathbf{B}^\star)$ and $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ are balanced factorizations of $\mathbf{M}^\star = \mathbf{A}^\star(\mathbf{B}^\star)^\top$, and $\hat{\mathbf{M}} = \hat{\mathbf{A}}\hat{\mathbf{B}}^\top$. Let $r = \mathrm{rank}(\hat{\mathbf{M}})$. Fix $\epsilon > 0$ and $s \in \mathbb{N}$ such that $s > 1$, $\epsilon \geq \|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{F}}$, and $\epsilon \leq \frac{\|\mathbf{M}^\star\|_{\mathrm{op}}}{40s}$. Also, for $q \geq 1$, let $\mathsf{tail}_q(\mathbf{M}; k) := \sum_{i>k} \sigma_i(\mathbf{M})^q$. Then, there exists an index $k \in [\min\{r, s-1\}]$ and an orthogonal matrix $\mathbf{R} \in \mathbb{O}(p)$ such that*

$$\Delta_0(\mathbf{R}, k) + \mathsf{tail}_2(\mathbf{M}^\star; k) \lesssim s^3\epsilon^2 + s(\sigma_s(\mathbf{M}^\star))^2 + \mathsf{tail}_2(\mathbf{M}^\star; s)$$
$$\Delta_1(\mathbf{R}, k) \lesssim (\sqrt{r} + s^2)\epsilon + s\sigma_s(\mathbf{M}^\star) + \mathsf{tail}_1(\mathbf{M}^\star; s),$$

and moreover, $\mathrm{range}((\hat{\mathbf{A}}\mathbf{R})^\top \hat{\mathbf{A}}\mathbf{R}) \supset \mathrm{range}((\mathbf{A}^\star_{[k]})^\top (\mathbf{A}^\star_{[k]}))$.

The above theorem is a specialization of a more extensive guarantee, Theorem 7, stated and proved in Appendix I. There are a number of important points to make. First, the theorem requires specifying a target rank s, but the guarantee applies to a smaller rank k; this is explained in the proof sketch below. Still, care is ensured to guarantee that the upper bounds on $\Delta_i(\mathbf{R}, k)$ depend only on the tail-decay at s, but not k. Second, we observe that when instantiated with $\mathbf{M}^\star = \mathbf{M}_{1\otimes 1}(f^\star, g^\star)$ as above, $\mathsf{tail}_q(\mathbf{M}^\star; s) = \mathbf{tail}^\star_q(s)$, i.e., it is the tail of the spectrum of $\boldsymbol{\Sigma}^\star_{1\otimes 1}$. Third, the guarantee applies to an orthogonal transformation $\mathbf{R}$, and the guarantee of $\mathrm{range}((\hat{\mathbf{A}}\mathbf{R})^\top \hat{\mathbf{A}}\mathbf{R}) \supset \mathrm{range}((\mathbf{A}^\star_{[k]})^\top (\mathbf{A}^\star_{[k]}))$ ensures that, for $\hat{\mathbf{A}} = \mathbf{A}_1(f)$ as instantiated above, the transformed embeddings $(\mathbf{R}^\top f, \mathbf{R}^\top g)$ are aligned-k proxies. Lastly, observe that the weighted error is asymptotically *quadratically smaller* in ϵ than the unweighted one; this is also explained in the proof sketch below.

To conclude the proof of Theorem 2, we first extend, via limiting arguments, to the setting of bilinear embeddings with arbitrary distributions; this result, Theorem 8, and its proof, are given in Appendix J. This provides an upper bound on $\boldsymbol{\Delta}_0(f, g, k), \boldsymbol{\Delta}_1(f, g, k)$ in terms of the term $\epsilon^2_{1\otimes 1}$ in Definition 3.1. Finally, we conclude the proof of Theorem 2 in Appendix M.7 by plugging Theorem 8 into Proposition 4.1 and substituting $\boldsymbol{\Delta}_{\mathrm{train}} = \mathcal{R}(f, g; \mathcal{D}_{\mathrm{train}}) \leftarrow \epsilon_{\mathrm{trn}}$ as in Definition 3.1.

4.5. Proof sketch of Theorem 5

The proof of Theorem 5 is our most technically innovative result; we sketch some of these techniques here, deferring the formal proof to Appendix I. Though previous bounds for matrix recovery exist (notably Tu et al. (2016, Lemma 5.14) as restated in Lemma I.5), these results assume matrices to either have *exactly low-rankness*, or have *sufficiently large spectral gap*. Addressing more gradual spectral decay requires a far more subtle treatment.

Technical novelty #1: Relative singular-gap SVD perturbation. The first technical ingredient is the perturbation for the rank-k SVD approximation, Theorem 1, highlighted in the introduction, which replaces a dependence on *absolute eigengap* with one on *relative* eigengap.

Technique novelty #2: "Well-tempered" partition. Motivated by the advantages of considering a relative (as opposed to absolute) singular gap, we construct a certain partition of the spectrum of $\mathbf{M}^\star$, which we call a "well-tempered partition" (Definition I.3). This partition splits the indices of the top-s singular values of $\mathbf{M}^\star$ into intervals where: (a) the relative-singular gap separation between the *intervals* is sufficiently large, and (b) all singular values are of similar magnitude.

Specifically, we denote the subsets in this partition as $\mathcal{K}_i = \{k_i + 1, k_i + 2, \dots, k_{i+1}\}$; we call k_i the *pivot* and each $\mathcal{K}_i$ a *block*. We show that the partition can be constructed so as to ensure that the *relative* spectral gap $\delta_{k_i}(\mathbf{M}^\star)$, where for any k, is at least $\Omega(1/s)$. Here again, s is the target rank in Theorem 5. As noted above, the *absolute* singular gaps can be arbitrarily smaller.

Given this partition, we decompose the factor matrices $\mathbf{A}^\star_{[k]}, \mathbf{B}^\star_{[k]}, \hat{\mathbf{A}}, \hat{\mathbf{B}}$ into a sum over block-zero-masked matrices $\mathbf{A}^\star_{\mathcal{K}_i}, \mathbf{B}^\star_{\mathcal{K}_i}, \hat{\mathbf{A}}_{\mathcal{K}_i}, \hat{\mathbf{B}}_{\mathcal{K}_i}$, with each block corresponding to one element $\mathcal{K}_i$ of the well-tempered partition. We let $\mathbf{M}^\star_{\mathcal{K}_i} = \mathbf{A}^\star_{\mathcal{K}_i}(\mathbf{B}^\star_{\mathcal{K}_i})^\top$, with $\hat{\mathbf{M}}_{\mathcal{K}_i}$ being defined similarly. We use the triangle inequality to relate $\|\hat{\mathbf{M}}_{\mathcal{K}_i} - \mathbf{M}^\star_{\mathcal{K}_i}\|_{\mathrm{F}}$ to $\max_{j\in\{k_i,k_{i+1}\}}\{\|\hat{\mathbf{M}}_{[j]} - \mathbf{M}^\star_{[j]}\|_{\mathrm{F}}\}$, and bound the latter two using our SVD perturbation result (Theorem 1). This is to our advantage, since our choice of well-tempered partition guarantees that $\delta_j(\mathbf{M}^\star) = \Omega(1/s)$ for $j \in \{k_i, k_{i+1}\}$, and implies via Theorem 1 that $\|\hat{\mathbf{M}}_{\mathcal{K}_i} - \mathbf{M}^\star_{\mathcal{K}_i}\|_{\mathrm{F}}^2 \lesssim s^2\epsilon^2$. We then apply an existing matrix factorization lemma, Tu et al. (2016, Lemma 5.14) to these blocks. The rotation matrix $\mathbf{R}$ aligns the block-masked factor matrices to minimize factor error. Though Theorem 1 depends on *relative* gaps, the factor recovery error in block i in Tu et al. (2016, Lemma 5.14) depends on *absolute* ones, scaling with

$$\frac{\|\hat{\mathbf{M}}_{\mathcal{K}_i} - \mathbf{M}^\star_{\mathcal{K}_i}\|_{\mathrm{F}}^2}{\sigma_{k_i}(\mathbf{M}^\star)} \lesssim \frac{s^2\epsilon^2}{\sigma_{k_i}(\mathbf{M}^\star)}. \tag{4.4}$$

For the unweighted error, we select the rank cutoff k to ensure $\sigma_k(\mathbf{M}^\star)$ is sufficiently large; trading-off the tails $\mathsf{tail}_q(k; \mathbf{M}^\star)$ with $\sigma_k(\mathbf{M}^\star)$ leads to the unweighted error $\Delta_1(\mathbf{R}, k)$ to scale with ϵ, rather than ϵ^2. For the weighted error $\Delta_0(\mathbf{R}, k)$, we can weight the factor recovery errors in the i-th block by $\sigma_{k_{i-1}+1} = \max\{\sigma_j(\mathbf{M}^\star) : j \in \mathcal{K}_i\}$. We then use the second property of the well-tempered partition: all singular values indexed in $\mathcal{K}_i$ are of roughly constant magnitude; thus, weighting by $\sigma_{k_{i-1}+1}(\mathbf{M}^\star)$ cancels out the denominator of $\sigma_{k_i}(\mathbf{M}^\star)$ in Eq. (4.4), yielding a sharper estimate.

5. Conclusion

In sum, this paper explores the connection between combinatorial distribution shift and matrix completion, developing fundamental and novel technical tools along the way. Whether our results can be extended to more general coverage assumptions than those depicted in Figure 1 remains an exciting direction for future research.

Acknowledgements

The authors would like to thank the anonymous reviewers for the helpful comments. K.Z. acknowledges support from Simons-Berkeley Research Fellowship. M.S. acknowledges support from an Amazon.com Services LLC; PO# 2D-06310236 grant.

References

Anish Agarwal, Munther Dahleh, Devavrat Shah, and Dennis Shen. Causal matrix completion. *arXiv preprint arXiv:2109.15154*, 2021.

Zeyuan Allen-Zhu and Yuanzhi Li. Lazysvd: Even faster svd decomposition yet without agonizing pain. *Advances in neural information processing systems*, 29, 2016.

Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 2002.

Sohom Bhattacharya and Sourav Chatterjee. Matrix completion with data-dependent missingness probabilities. *IEEE Transactions on Information Theory*, 2022.

Nicolai Bissantz, Thorsten Hohage, Axel Munk, and Frits Ruymgaart. Convergence rates of general regularization methods for statistical inverse problems and applications. *SIAM Journal on Numerical Analysis*, 45(6):2610–2636, 2007.

Christian Borgs, Jennifer Chayes, Christina E Lee, and Devavrat Shah. Thy friend is my friend: Iterative collaborative filtering for sparse matrix estimation. *Advances in neural information processing systems*, 30, 2017.

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.

Emmanuel Candes and Benjamin Recht. Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119, 2012.

Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

Erhan Çinlar. *Probability and stochastics*, volume 261. Springer, 2011.

Mark A Davenport, Yaniv Plan, Ewout Van Den Berg, and Mary Wootters. 1-bit matrix completion. *Information and Inference: A Journal of the IMA*, 3(3):189–223, 2014.

Simon Foucart, Deanna Needell, Reese Pathak, Yaniv Plan, and Mary Wootters. Weighted matrix completion from non-random, non-uniform sampling patterns. *IEEE Transactions on Information Theory*, 67(2):1264–1290, 2020.

Nils Erik Gilhus and Jan J Verschuuren. Myasthenia gravis: subgroup classification and therapeutic strategies. *The Lancet Neurology*, 14(10):1023–1036, 2015.

Trevor Hastie, Rahul Mazumder, Jason D Lee, and Reza Zadeh. Matrix completion and low-rank svd via fast alternating least squares. *The Journal of Machine Learning Research*, 16(1):3367–3402, 2015.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *stat*, 1050:9, 2015.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavnass, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, 2021.

Vladimir Koltchinskii, Karim Lounici, and Alexandre B Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

Chi-Kwong Li and Gilbert Strang. An elementary proof of mirsky's low rank approximation theorem. *The Electronic Journal of Linear Algebra*, 36:694–697, 2020.

Ren-Cang Li. Relative perturbation theory: Ii. eigenspace and singular subspace variations. *SIAM Journal on Matrix Analysis and Applications*, 20(2):471–492, 1998.

Yihua Li, Devavrat Shah, Dogyoon Song, and Christina Lee Yu. Nearest neighbors for matrix estimation interpreted as blind regression for latent variable model. *IEEE Transactions on Information Theory*, 66(3):1760–1784, 2019.

Dawen Liang, Laurent Charlin, James McInerney, and David M Blei. Modeling user exposure in recommendation. In *Proceedings of the 25th international conference on World Wide Web*, pages 951–961, 2016.

Nian Liu and Junwei Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 678–686, 2016.

Cong Ma, Reese Pathak, and Martin J Wainwright. Optimally tackling covariate shift in RKHS-based nonparametric regression. *arXiv:2205.02986*, 2022.

Wei Ma and George H Chen. Missing not at random in matrix completion: The effectiveness of estimating missingness probabilities under a low nuclear norm assumption. *Advances in Neural Information Processing Systems*, 32, 2019.

Lester Mackey, Michael I Jordan, Richard Y Chen, Brendan Farrell, and Joel A Tropp. Matrix concentration inequalities via the method of exchangeable pairs. *The Annals of Probability*, 42 (3):906–945, 2014.

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research*, 11:2287–2322, 2010.

Shahar Mendelson and Joseph Neeman. Regularization in kernel learning. *Annals of Statistics*, 38 (1):526–565, 2010.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, 2021.

Reese Pathak, Cong Ma, and Martin Wainwright. A new similarity measure for covariate shift with applications to nonparametric regression. In *International Conference on Machine Learning*, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(12), 2011.

Shibani Santurkar, Dimitris Tsipras, and Aleksander Madry. Breeds: Benchmarks for subpopulation shift. *arXiv:2008.04859*, 2020.

Bernhard A Schmitt. Perturbation bounds for matrix square roots and pythagorean sums. *Linear algebra and its applications*, 174:215–227, 1992.

Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *international conference on machine learning*, pages 1670–1679. PMLR, 2016.

Devavrat Shah, Dogyoon Song, Zhi Xu, and Yuzhe Yang. Sample efficient reinforcement learning via low-rank matrix estimation. *Advances in Neural Information Processing Systems*, 33:12092–12103, 2020.

Dogyoon Song, Christina E Lee, Yihua Li, and Devavrat Shah. Blind regression: Nonparametric regression for latent variable models via collaborative filtering. *Advances in Neural Information Processing Systems*, 29, 2016.

Aude Sportisse, Claire Boyer, and Julie Josse. Imputation and low-rank estimation with missing not at random data. *Statistics and Computing*, 30(6):1629–1643, 2020.

Gilbert W Stewart and Ji-guang Sun. *Matrix perturbation theory*. Academic press, 1990.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 2020.

Jonathan Tremblay, Thang To, Balakumar Sundaralingam, Yu Xiang, Dieter Fox, and Stan Birchfield. Deep object pose estimation for semantic robotic grasping of household objects. In *Conference on Robot Learning*, pages 306–316. PMLR, 2018.

Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Ben Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *International Conference on Machine Learning*, pages 964–973. PMLR, 2016.

Madeleine Udell and Alex Townsend. Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science*, 1(1):144–160, 2019.

Vladimir Vapnik. *Estimation of dependences based on empirical data*. Springer Science & Business Media, 2006.

Menghan Wang, Mingming Gong, Xiaolin Zheng, and Kun Zhang. Modeling dynamic missingness of implicit feedback for recommendation. *Advances in neural information processing systems*, 31, 2018.

Jiaming Xu. Rates of convergence of spectral methods for graphon estimation. In *International Conference on Machine Learning*, pages 5433–5442. PMLR, 2018.

Chengrun Yang, Lijun Ding, Ziyang Wu, and Madeleine Udell. Tenips: Inverse propensity sampling for tensor completion. In *International Conference on Artificial Intelligence and Statistics*, pages 3160–3168. PMLR, 2021.

Christina Lee Yu. Nonparametric matrix estimation with one-sided covariates. *arXiv preprint arXiv:2110.13969*, 2021.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Supplementary Materials for
"Tackling Combinatorial Distribution Shift: A Matrix Completion Perspective"

Contents

Part I
Organization, Related Work, Further Discussion, and the SVD Perturbation bound

Appendix A. Organization of the Appendix

We detail the organization of our appendix as follows. Part I provides the overall organization of the appendix in Appendix A, a detailed related work in Appendix B, especially on matrix completion, and an elaboration on the connection between bilinear combinatorial extrapolation and matrix completion in Appendix C. In Appendix D, we prove Theorem 1, our main SVD perturbation bound.

Part II provides supplementary material regarding our guarantees for the single-stage and double-stage ERM procedures. Appendix E provides the high-level proof of our guarantee for single-stage ERM (Theorem 3), and a more detailed guarantee for double-stage ERM (Theorem 6), deriving Theorem 4 from that more granular result. Appendix F provides the proof of Theorem 6, which in turn contains as the single-stage ERM guarantee used by Theorem 3. These proofs in turn rely on some general (though quite standard) learning-theoretic bounds, which are supplied in Appendix G. Finally, Appendix H performs the computations which instantiates out single- and double-stage ERM guarantees for the spectral decay regimes prescribed by Assumption 2.6.

Part III contains the supplementary results needed for the proof of the meta-theorem (Theorem 2), as well as general-purpose linear algebraic results. Appendix I contains the proof of our main technical endeavor - a bound on the error of factor recovery in low-rank matrix approximation. Appendix J extends the matrix factorization guarantee to its natural generalization to bilinear embeddings, applying suitable limiting arguments to accomodate distributions with infinite/uncountable support. Most supporting linear algebraic results/proofs are deferred to Appendix K; notable, these include the proof of our relative singular-value gap perturbation bound (Theorem 11). Results pertaining to balancing (of both matrices and embeddings) are given in Appendix L. Finally, Appendix M provides the proof of the error decomposition (Proposition 4.1), as well as the derivation of Theorem 2 from Proposition 4.1 and Theorem 8.

Appendix B. Detailed Related Work

This subsection provides a more detailed summary of related work, to the best of our knowledge.

B.1. Matrix completion

To facilitate comparison, we consider a ground-truth matrix $\mathbf{M}^\star \in \mathbb{R}^{M\times N}$ as the matrix to be completed. $\mathbf{M} \in \mathbb{R}^{M\times N}$ is a noisy realization of $\mathbf{M}^\star$ with $\mathbb{E}[\mathbf{M}] = \mathbf{M}^\star$, and we assume that we are given observed matrix $\tilde{\mathbf{M}} \in (\mathbb{R} \cup \{\mathbf{?}\})^{M\times N}$, where '**?**' denotes an unseen entry, such that $\tilde{\mathbf{M}}_{[ab]} = \mathbf{M}_{[ab]}$ unless $\tilde{\mathbf{M}}_{[ab]} = \mathbf{?}$. We let $\mathbf{D} \in \{0,1\}^{M\times N}$ denote the masking matrix of $\tilde{\mathbf{M}}$: $\mathbf{D}_{[ab]} = \mathbb{I}\{\tilde{\mathbf{M}}_{[ab]} \neq \mathbf{?}\}$.

Missing-completely-at-random (MAR) matrix completion. In the MAR setting, it is assumed that the entries of $\mathbf{D}$ are i.i.d. Bernoulli random variables with positive probability $p > 0$ and independent of $\mathbf{M}$; some existing works include Candès and Tao (2010); Recht (2011); Hastie et al. (2015); Mazumder et al. (2010); Koltchinskii et al. (2011). More recent works study settings where

$\mathbf{M}^\star_{[ab]}$ is generated by the bivariate function $h^\star(x_a, y_b) = \langle f^\star(x_a), g^\star(y_b)\rangle$ of features x_a, y_b; in (Xu, 2018), this encodes graphon structure, whereas in Song et al. (2016); Li et al. (2019), $h^\star(x, y)$ is a globally Lipschitz function, which admits learning via matrix completion by considering linearizing expansions. Yu (2021) considers an extension to the "one-sided" covariate setting that is more challenging, where only the first argument of $h^\star$ is observed. A "one-bit" sensing model has also been studied in Davenport et al. (2014), and refined under a latent variable model for features x_a, y_b (Borgs et al., 2017). All aforementioned works consider the MAR setting.

Missing at random. In the missing-at-random setting, it is assumed that there exists a set of observed covariates $\mathcal{O}$ such that $\mathbf{M} \perp \mathbf{D} \mid \mathcal{O}$, and that $\mathbf{D}_{[ab]} \mid \mathcal{O}$ are independent Bernoulli random-variables with possibly different probabilities p_{ab} uniformly bounded below. See e.g., Schnabel et al. (2016); Wang et al. (2018); Liang et al. (2016).

Missing-not-at-random (MNAR) matrix completion. Many works consider generative models, relating missingness of entries to either ground-truth or realized values of the matrix via logistic expressions (Sportisse et al., 2020; Yang et al., 2021). Guarantees obtained from this strategy typically depend on a lower bound on the minimal probability that an entry is revealed (Ma and Chen, 2019), dependence on which is also incurred in an alternative approach due to Bhattacharya and Chatterjee (2022). Note that in our setting, we allow the entries of $\mathbf{M}^\star_{2,2}$ to be *entirely* omitted from $\tilde{\mathbf{M}}_{2,2}$, so these guarantees are vacuous here. Another approach due to Foucart et al. (2020) studies reconstruction from MNAR data under weighting matrices that are suitably calibrated to the pattern of missing entries. Again, in our setting, these results become vacuous.

Two more recent works establish recovery for entries that are indeed missing with probability one. Shah et al. (2020) considers almost precisely our setting, where, motivated by reinforcement learning, one attempt to recover $\mathbf{M}^\star_{2,2}$ by observing the other blocks $\mathbf{M}^\star_{1,2}, \mathbf{M}^\star_{2,1}, \mathbf{M}^\star_{2,2}$. However, their results require that either (a) $\mathbf{M}^\star$ is an exactly low rank, or (b) that $\mathbf{M}^\star$ is an approximately low rank, but that the error between $\mathbf{M}^\star$ and its rank-r SVD is very small entry-wise. This precludes the much more gradual polynomial decay allowed by our main results. A second work, Agarwal et al. (2021), considers far more general patterns of missing entries than we do in this work. However, this comes at the cost of requiring even stronger assumptions on the spectrum (Agarwal et al., 2021, Assumption 6), which again precludes approximately low-rank matrices with spectral decay.

B.2. Learning under distribution shift

In contrast to the well-established and classical statistical learning theory (Bartlett and Mendelson, 2002; Vapnik, 2006), our theoretical understanding of distribution shift is considerably more sparse. Notably, recent work has given precise characterizations of the effects of covariate shift for certain specific function classes, notably kernels (Ma et al., 2022) and Hölder smooth classes (Pathak et al., 2022); still, these works focus on the regimes where the test-distribution has bounded density with respect to the train distribution; in our bilinear combinatorial extrapolation setting, however, this is no longer the case. Resilience to distribution shift has received considerable empirical attention in recent years, see Miller et al. (2021); Taori et al. (2020); Santurkar et al. (2020); Koh et al. (2021); Zhou et al. (2022) for example.

Appendix C. Connection to Matrix Completion

Now we provide a connection of the bilinear combinatorial extrapolation problem to the problem of matrix completion with MNAR data. Consider a bilinear combinatorial extrapolation setting where the support sets $\mathcal{X}$ and $\mathcal{Y}$ have *finite* cardinalities, with elements $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_m\}$. For $i, j \in \{1, 2\}$, define the probabilities $\mathsf{p}_{i,\ell} = \mathbb{P}_{\mathcal{D}_{\mathcal{X},i}}[x = x_\ell]$ and $\mathsf{q}_{j,k} = \mathbb{P}_{\mathcal{D}_{\mathcal{Y},j}}[y = y_k]$. Because of the finite support of the distributions, we can regard any $\mathcal{H}$-embeddings (f, g) (including $(f^\star, g^\star)$) as embeddings into $\mathbb{R}^d$, $d = \max\{n, m\}$, appending zeros if necessary. Then we can define matrices $\mathbf{A}_i(f) \in \mathbb{R}^{n\times d}$ and $\mathbf{B}_j(g) \in \mathbb{R}^{m\times d}$ by assigning the rows to the scaled values of the embeddings $\mathbf{A}_i(f)[\ell, :] = \sqrt{\mathsf{p}_{i,\ell}} f(x_\ell)^\top$, $\mathbf{B}_j(g)[k, :] = \sqrt{\mathsf{q}_{j,k}} g(y_k)^\top$, and the matrices $\mathbf{M}_{i\otimes j}(f, g) = \mathbf{A}_i(f)\mathbf{B}_j(g)^\top$. Each matrix $\mathbf{M}_{i\otimes j}(f, g)$ can be thought of as a look-up table, where $\mathbf{M}_{i\otimes j}(f, g)[\ell, k] = \sqrt{\mathsf{p}_{i,\ell}\mathsf{q}_{j,k}}\langle f(x_\ell), g(y_k)\rangle$ is the prediction of $\langle f, g\rangle$, scaled by the square root probability of x_ℓ and y_k. Then, one can see that the risk of f, g is precisely equal to the Frobenius-norm error difference between the matrices $\mathbf{M}_{i\otimes j}(f, g)$ and $\mathbf{M}_{i\otimes j}(f^\star, g^\star)$. For simplicity, we write them as $\mathbf{M}$ and $\mathbf{M}^\star$ for short, respectively. Such a factorization inherently introduces the bilinear embedding form as defined in Assumption 2.1, and is also illustrated figuratively in Figure 2.

Consider the bilinear combinatorial extrapolation setting where we can sample from the matrix $\mathbf{M}^\star$ in the top three blocks, i.e., the block $\{(1,1), (1,2), (2,1)\}$, where for convenience we partition $\mathbf{M}^\star$ as $\mathbf{M}^\star = \begin{bmatrix}\mathbf{M}^\star_{11} & \mathbf{M}^\star_{12}\\ \mathbf{M}^\star_{21} & \mathbf{M}^\star_{22}\end{bmatrix}$, with $\mathbf{M}^\star_{11} \in \mathbb{R}^{\alpha n\times \beta m}$, $\mathbf{M}^\star_{12} \in \mathbb{R}^{\alpha n\times (1-\beta)m}$, $\mathbf{M}^\star_{21} \in \mathbb{R}^{(1-\alpha)n\times \beta m}$, and $\mathbf{M}^\star_{22} \in \mathbb{R}^{(1-\alpha)n\times(1-\beta)m}$. Here we assume that $\alpha, \beta \in (0, 1)$ are chosen such that the dimensions of these sub-matrices are positive integers. Our goal is to use the data uniformly sampled from blocks $\{\mathbf{M}^\star_{11}, \mathbf{M}^\star_{12}, \mathbf{M}^\star_{21}\}$ to predict and generalize to the uniform distribution supported on the bottom block $\mathbf{M}^\star_{22}$. In this case, we know that

$$\frac{\mathrm{d}\mathcal{D}_{1\otimes 1}}{\mathrm{d}\mathcal{D}_{\mathrm{train}}} = \frac{(1-\alpha)\beta + (1-\beta)\alpha + \alpha\beta}{\alpha\beta}, \quad \frac{\mathrm{d}\mathcal{D}_{1\otimes 2}}{\mathrm{d}\mathcal{D}_{\mathrm{train}}} = \frac{(1-\alpha)\beta + (1-\beta)\alpha + \alpha\beta}{(1-\beta)\alpha} \tag{C.1}$$

$$\frac{\mathrm{d}\mathcal{D}_{2\otimes 1}}{\mathrm{d}\mathcal{D}_{\mathrm{train}}} = \frac{(1-\alpha)\beta + (1-\beta)\alpha + \alpha\beta}{(1-\alpha)\beta}, \quad \frac{\mathrm{d}\mathcal{D}_{\mathrm{test}}}{\sum_{i,j}\mathrm{d}\mathcal{D}_{i\otimes j}} = \alpha\beta, \tag{C.2}$$

where we write $\frac{d\mathcal{D}_1(x,y)}{d\mathcal{D}_2(x,y)}$ as $\frac{d\mathcal{D}_1}{d\mathcal{D}_2}$ for short since they are identical on the support with uniform distributions. Note that Eqs. (C.1) and (C.2) instantiate the constants $\kappa_{\mathrm{trn}} = \frac{(1-\alpha)\beta+(1-\beta)\alpha+\alpha\beta}{\min\{(1-\beta)\alpha,(1-\alpha)\beta,\alpha\beta\}}$ and $\kappa_{\mathrm{tst}} = \alpha\beta$ in Assumption 2.2.

Moreover, suppose that $\mathbf{M}^\star = \mathbf{A}^\star(\mathbf{B}^\star)^\top$ for some $\mathbf{A}^\star = \begin{bmatrix}\mathbf{A}^\star_1\\ \mathbf{A}^\star_2\end{bmatrix}$ and $\mathbf{B}^\star = \begin{bmatrix}\mathbf{B}^\star_1\\ \mathbf{B}^\star_2\end{bmatrix}$ that are balanced, in the sense that $(\mathbf{A}^\star_1)^\top\mathbf{A}^\star_1 = (\mathbf{B}^\star_1)^\top\mathbf{B}^\star_1$. Also, suppose that with this block partition, $(\mathbf{A}^\star_1)^\top\mathbf{A}^\star_1 \succeq \kappa_1(\mathbf{A}^\star_2)^\top\mathbf{A}^\star_2$ and $(\mathbf{B}^\star_1)^\top\mathbf{B}^\star_1 \succeq \kappa_1(\mathbf{B}^\star_2)^\top\mathbf{B}^\star_2$ for some $\kappa_1 > 0$. Then we have that the constant κ_{cov} in Assumption 2.3 satisfies that $\kappa_{\mathrm{cov}} \geq 1/\kappa_1$. In addition, Assumption 2.6 now becomes the spectral decay assumption on the matrix $(\mathbf{A}^\star_1)^\top\mathbf{A}^\star_1$ (and thus also $(\mathbf{B}^\star_1)^\top\mathbf{B}^\star_1$). Finally, note that under this setting, we also have

$$\frac{\mathrm{d}\mathcal{D}_{\mathrm{train}}}{\sum_{i,j}\mathrm{d}\mathcal{D}_{i\otimes j}} = \frac{\alpha\beta(1-\alpha)(1-\beta)}{(1-\alpha)\beta + (1-\beta)\alpha + \alpha\beta} =: \tilde{\kappa}_{\mathrm{trn}},$$

for some $\tilde{\kappa}_{\mathrm{trn}}$. Then, together with Assumption 2.3 with $\kappa_{\mathrm{cov}} = 1/\kappa_1$, we know that Assumption 2.5 is satisfied with

$$\kappa_{\mathrm{apx}} = 4\tilde{\kappa}_{\mathrm{trn}}\kappa_{\mathrm{cov}}^2 = \frac{4\alpha\beta(1-\alpha)(1-\beta)}{\kappa_1^2 \cdot [(1-\alpha)\beta + (1-\beta)\alpha + \alpha\beta]}.$$

Appendix D. Relative-gap perturbations of the SVD

Theorem 1 (Perturbation of SVD Approximation with Relative Gap) *Let $\mathbf{M}^\star, \hat{\mathbf{M}} \in \mathbb{R}^{n\times m}$. Fix a $k \le \min\{n, m\}$ for which $\sigma_k(\mathbf{M}^\star) > 0$ and the relative spectral gap $\delta_k(\mathbf{M}^\star)$ (Eq. (1.1)) is positive. Then, if $\|\mathbf{M}^\star - \hat{\mathbf{M}}\|_{\mathrm{op}} \le \eta\sigma_k(\mathbf{M}^\star)\delta_k(\mathbf{M}^\star)$ for some $\eta \in (0, 1)$, we have that the rank-k SVD approximations of $\mathbf{M}^\star$ and $\hat{\mathbf{M}}$, denoted as $\mathbf{M}^\star_{[k]}$ and $\hat{\mathbf{M}}_{[k]}$, are unique, and satisfy*

$$\|\hat{\mathbf{M}}_{[k]} - \mathbf{M}^\star_{[k]}\|_{\mathrm{F}} \le \frac{9\|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{F}}}{\delta_k(\mathbf{M}^\star)(1-\eta)}.$$

Proof [Proof of Theorem 1] We begin by expanding the Frobenius error:

$$\begin{aligned}\|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{F}}^2 &= \|\hat{\mathbf{M}}_{[k]} - \mathbf{M}^\star_{[k]} + (\hat{\mathbf{M}}_{>k} - \mathbf{M}^\star_{>k})\|_{\mathrm{F}}^2 \\ &= \|\hat{\mathbf{M}}_{[k]} - \mathbf{M}^\star_{[k]}\|_{\mathrm{F}}^2 + \|\hat{\mathbf{M}}_{>k} - \mathbf{M}^\star_{>k}\|_{\mathrm{F}}^2 + 2\langle \hat{\mathbf{M}}_{[k]} - \mathbf{M}^\star_{[k]}, \hat{\mathbf{M}}_{>k} - \mathbf{M}^\star_{>k}\rangle.\end{aligned}$$

Hence,

$$\begin{aligned}\|\hat{\mathbf{M}}_{[k]} - \mathbf{M}^\star_{[k]}\|_{\mathrm{F}}^2 &\le \|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{F}}^2 + 2|\langle \hat{\mathbf{M}}_{[k]} - \mathbf{M}^\star_{[k]}, \hat{\mathbf{M}}_{>k} - \mathbf{M}^\star_{>k}\rangle| \\ &\overset{(i)}{=} \|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{F}}^2 + 2|\langle \hat{\mathbf{M}}_{[k]}, \mathbf{M}^\star_{>k}\rangle - \langle \mathbf{M}^\star_{[k]}, \hat{\mathbf{M}}_{>k}\rangle| \\ &\overset{(ii)}{\le} \|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{F}}^2 + 2|\langle \hat{\mathbf{M}}_{[k]}, \mathbf{M}^\star_{>k}\rangle| + 2|\langle \mathbf{M}^\star_{[k]}, \hat{\mathbf{M}}_{>k}\rangle|, \qquad \text{(D.1)}\end{aligned}$$

where above (i) uses that the range of the rank-k SVD of a matrix and its complement are orthogonal, and (ii) is just the triangle inequality. The following claim bounds the cross terms:

Claim D.1 *Suppose $\sigma_k(\hat{\mathbf{M}}) > \sigma_{k+1}(\mathbf{M}^\star)$. Then,*

$$|\langle \hat{\mathbf{M}}_{[k]}, \mathbf{M}^\star_{>k}\rangle| \le 4\|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{F}}^2 \cdot \left(\left(1 - \frac{\sigma_{k+1}(\mathbf{M}^\star)}{\sigma_k(\hat{\mathbf{M}})}\right)^{-2} + 4\right),$$

Similarly, if $\sigma_k(\mathbf{M}^\star) > \sigma_{k+1}(\hat{\mathbf{M}})$. Then,

$$|\langle \mathbf{M}^\star_{[k]}, \hat{\mathbf{M}}_{>k}\rangle| \le 4\|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{F}}^2 \cdot \left(\left(1 - \frac{\sigma_{k+1}(\hat{\mathbf{M}})}{\sigma_k(\mathbf{M}^\star)}\right)^{-2} + 4\right).$$

The proof of Claim D.1 uses a careful peeling argument, and is deferred to the end. The key idea is to parition the singular values of $\mathbf{M}^\star$ into blocks whose singular values are all within a constant factor, and into one final block such corresponding to singular values $j > k$ of $\mathbf{M}^\star$. We then apply a standard variant of Wedin's theorem (Lemma D.2) to each block. The form of the matrix inner product allows us to weight the contribution of each block by its associated singular value. The upshot is that this leads to gap-free bounds for all but the last-block (as all singular values in these blocks are within a constant of eachother), and a similar argument leaves us only with dependence on the relative singular gap for the final block.

We now specialize the above upper bound when $\|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{op}}$ is sufficiently small.

Claim D.2 *Suppose* $\|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{op}} \le \eta \delta_k^\star \sigma_k(\mathbf{M}^\star)$. *Then,*

$$|\langle \hat{\mathbf{M}}_{[k]}, \mathbf{M}^\star_{>k}\rangle| \vee |\langle \mathbf{M}^\star_{[k]}, \hat{\mathbf{M}}_{>k}\rangle| \le 4\|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{F}}^2 \cdot \left((\delta_k^\star)^{-2}(1-\eta)^{-2} + 4\right),$$

Proof If $\|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{op}} \le \eta \delta_k^\star \sigma_k(\mathbf{M}^\star)$,

$$\begin{aligned}
1 - \frac{\sigma_{k+1}(\mathbf{M}^\star)}{\sigma_k(\hat{\mathbf{M}})} &\ge 1 - (1-\eta\delta_k^\star)^{-1}\frac{\sigma_{k+1}(\mathbf{M}^\star)}{\sigma_k(\mathbf{M}^\star)} \\
&= 1 - (1-\eta\delta_k^\star)^{-1}(1-\delta_k^\star) \\
&= \frac{1-\eta\delta_k^\star - (1-\delta_k^\star)}{1-\eta\delta_k^\star} = \frac{\delta_k^\star(1-\eta)}{1-\eta\delta_k^\star} \ge \delta_k^\star(1-\eta),
\end{aligned}$$

and

$$\begin{aligned}
1 - \frac{\sigma_{k+1}(\hat{\mathbf{M}})}{\sigma_k(\mathbf{M}^\star)} &\ge 1 - \frac{\eta\delta_k^\star\sigma_k(\mathbf{M}^\star) + \sigma_{k+1}(\mathbf{M}^\star)}{\sigma_k(\mathbf{M}^\star)} \\
&= 1 - \eta\delta_k^\star - \frac{\sigma_{k+1}(\mathbf{M}^\star)}{\sigma_k(\mathbf{M}^\star)} = (1-\eta\delta_k^\star) - (1-\delta_k^\star) = \delta_k^\star(1-\eta).
\end{aligned}$$

Hence, in both cases, Claim D.1 yields.

$$|\langle \hat{\mathbf{M}}_{[k]}, \mathbf{M}^\star_{>k}\rangle| \vee |\langle \hat{\mathbf{M}}_{[k]}, \mathbf{M}^\star_{>k}\rangle| \le 4\|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{F}}^2 \cdot \left((\delta_k^\star)^{-2}(1-\eta)^{-2} + 4\right),$$

which completes the proof. ∎

To conclude, we recall Eq. (D.1) and apply the previous claim

$$\begin{aligned}
\|\hat{\mathbf{M}}_{[k]} - \mathbf{M}^\star_{[k]}\|_{\mathrm{F}}^2 &\le \|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{F}}^2 + 2|\langle \hat{\mathbf{M}}_{[k]}, \mathbf{M}^\star_{>k}\rangle| + 2|\langle \mathbf{M}^\star_{[k]}, \hat{\mathbf{M}}_{>k}\rangle| \\
&\le \|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{F}}^2 + 4(|\langle \hat{\mathbf{M}}_{[k]}, \mathbf{M}^\star_{>k}\rangle| \vee |\langle \mathbf{M}^\star_{[k]}, \hat{\mathbf{M}}_{>k}\rangle|) \\
&\le \|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{F}}^2 + \|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{F}}^2 \cdot \left(16(\delta_k^\star)^{-2}(1-\eta)^{-2} + 64\right) \\
&= \|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{F}}^2 \cdot \left(16(\delta_k^\star)^{-2}(1-\eta)^{-2} + 65\right) \\
&\le 81\|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{F}}^2 \cdot \left((\delta_k^\star)^{-2}(1-\eta)^{-2}\right).
\end{aligned}$$

The bound follows. ∎

Proof [Proof of Claim D.1] We prove the first statement of the claim; the second is analogous. Consider a sequence of indices $k_0 > k_1 > \dots k_\ell = 0$ as follows (For convenience, k_i are decreasing, unlike the pivots k_i in the definition of the well-tempered parition Definition I.3).

- $k_0 = k$.

- Given k_i, set $k_{i+1} = \max\{j \geq 1 : \sigma_j(\hat{\mathbf{M}}) \geq 2\sigma_{k_i}(\hat{\mathbf{M}})\}$. If no such j exists, set $i+1 = \ell$ and $k_\ell = 0$.

We also define the index sets and corresponding SVD of $\hat{\mathbf{M}}$ as

$$\mathcal{I}_i := \{j : k_i \geq j > k_{i+1}\}, \quad \hat{\mathbf{M}}_{\mathcal{I}_i} := \hat{\mathbf{U}}_{\mathcal{I}_i}\hat{\mathbf{\Sigma}}_{\mathcal{I}_i}\hat{\mathbf{V}}_{\mathcal{I}_i}^\top,$$

where $\hat{\mathbf{U}}_{\mathcal{I}_i} \in \mathbb{R}^{n\times|\mathcal{I}_i|}$, $\hat{\mathbf{\Sigma}}_{\mathcal{I}_i} \in \mathbb{R}^{|\mathcal{I}_i|\times|\mathcal{I}_i|}$, $\hat{\mathbf{V}}_{\mathcal{I}_i} \in \mathbb{R}^{m\times|\mathcal{I}_i|}$ denote a compact SVD of $\hat{\mathbf{M}}_{\mathcal{I}_i}$ corresponding to singular values/vectors with indices in $\mathcal{I}_i$ (i.e. to the rows of $\hat{\mathbf{U}}$ corresponding to entries $j \in \mathcal{I}_i$, and similarly for $\hat{\mathbf{\Sigma}}_{\mathcal{I}_i}\hat{\mathbf{V}}_{\mathcal{I}_i}^\top$). We then have

$$\sum_{i=0}^{\ell-1} \hat{\mathbf{M}}_{\mathcal{I}_i} = \hat{\mathbf{M}}_{[k]}.$$

Using this decomposition, we write

$$\begin{aligned}
|\langle \hat{\mathbf{M}}_{[k]}, \mathbf{M}^\star_{>k}\rangle| &= \left|\langle \sum_{i=0}^{\ell-1} \hat{\mathbf{M}}_{\mathcal{I}_i}, \mathbf{M}^\star_{>k}\rangle\right| \\
&\leq \sum_{i=0}^{\ell-1} |\langle \hat{\mathbf{M}}_{\mathcal{I}_i}, \mathbf{M}^\star_{>k}\rangle| \\
&= \sum_{i=0}^{\ell-1} |\langle \hat{\mathbf{U}}_{\mathcal{I}_i}\hat{\mathbf{\Sigma}}_{\mathcal{I}_i}\hat{\mathbf{V}}_{\mathcal{I}_i}^\top, \mathbf{U}^\star_{>k}\mathbf{\Sigma}^\star_{>k}(\mathbf{V}^\star_{>k})^\top\rangle| \\
&= \sum_{i=0}^{\ell-1} |\mathrm{tr}(\hat{\mathbf{V}}_{\mathcal{I}_i}\hat{\mathbf{\Sigma}}_{\mathcal{I}_i}\hat{\mathbf{U}}_{\mathcal{I}_i}^\top\mathbf{U}^\star_{>k}\mathbf{\Sigma}^\star_{>k}(\mathbf{V}^\star_{>k})^\top)| \\
&= \sum_{i=0}^{\ell-1} |\mathrm{tr}(\hat{\mathbf{\Sigma}}_{\mathcal{I}_i}\hat{\mathbf{U}}_{\mathcal{I}_i}^\top\mathbf{U}^\star_{>k}\mathbf{\Sigma}^\star_{>k}(\mathbf{V}^\star_{>k})^\top\hat{\mathbf{V}}_{\mathcal{I}_i})| \\
&\leq \sum_{i=0}^{\ell-1} \|\hat{\mathbf{\Sigma}}_{\mathcal{I}_i}\hat{\mathbf{U}}_{\mathcal{I}_i}^\top\mathbf{U}^\star_{>k}\|_{\mathrm{F}}\|\mathbf{\Sigma}^\star_{>k}(\mathbf{V}^\star_{>k})^\top\hat{\mathbf{V}}_{\mathcal{I}_i}\|_{\mathrm{F}} \\
&\leq \sum_{i=0}^{\ell-1} \|\hat{\mathbf{\Sigma}}_{\mathcal{I}_i}\|_{\mathrm{op}}\|\mathbf{\Sigma}^\star_{>k}\|_{\mathrm{op}}\|\hat{\mathbf{U}}_{\mathcal{I}_i}^\top\mathbf{U}^\star_{>k}\|_{\mathrm{F}}\|(\mathbf{V}^\star_{>k})^\top\hat{\mathbf{V}}_{\mathcal{I}_i}\|_{\mathrm{F}}.
\end{aligned} \tag{D.2}$$

Since $\mathcal{I}_i \subseteq [k_i]$, we can bound

$$\begin{aligned}
\|\hat{\mathbf{U}}_{\mathcal{I}_i}^\top\mathbf{U}^\star_{>k}\|_{\mathrm{F}}\|(\mathbf{V}^\star_{>k})^\top\hat{\mathbf{V}}_{\mathcal{I}_i}\|_{\mathrm{F}} &\leq \|\hat{\mathbf{U}}_{[k_i]}^\top\mathbf{U}^\star_{>k}\|_{\mathrm{F}}\|(\mathbf{V}^\star_{>k})^\top\hat{\mathbf{V}}_{[k_i]}\|_{\mathrm{F}} \\
&\leq \frac{1}{2}\left(\|\hat{\mathbf{U}}_{[k_i]}^\top\mathbf{U}^\star_{>k}\|_{\mathrm{F}}^2 + \|(\mathbf{V}^\star_{>k})^\top\hat{\mathbf{V}}_{[k_i]}\|_{\mathrm{F}}^2\right).
\end{aligned}$$

In particular, since $k_i \leq k$, we see that as long as $\sigma_k(\hat{\mathbf{M}}) > \sigma_{k+1}(\mathbf{M}^\star)$, then by a standard variant of Wedin's theorem, Lemma D.2,

$$\|\hat{\mathbf{U}}_{\mathcal{I}_i}^\top\mathbf{U}^\star_{>k}\|_{\mathrm{F}}\|(\mathbf{V}^\star_{>k})^\top\hat{\mathbf{V}}_{\mathcal{I}_i}\|_{\mathrm{F}} \leq \frac{2}{(\sigma_{k_i}(\hat{\mathbf{M}}) - \sigma_{k+1}(\mathbf{M}^\star))^2}\|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{F}}^2.$$

We furthe observe that $\|\boldsymbol{\Sigma}^\star_{>k}\|_{\mathrm{op}} = \sigma_{k+1}(\mathbf{M}^\star)$, and $\|\hat{\boldsymbol{\Sigma}}_{\mathcal{I}_i}\|_{\mathrm{op}} = \sigma_{k_{i+1}-1}(\hat{\mathbf{M}}) \le 2\sigma_{k_i}(\hat{\mathbf{M}})$. Thus, picking up from Eq. (D.2)

$$\begin{aligned}
|\langle \hat{\mathbf{M}}_{[k]}, \mathbf{M}^\star_{>k}\rangle| &\le \|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{F}}^2 \cdot \sum_{i=0}^{\ell-1} \frac{4\sigma_{k+1}(\mathbf{M}^\star)\sigma_{k_i}(\hat{\mathbf{M}})}{(\sigma_{k_i}(\hat{\mathbf{M}}) - \sigma_{k+1}(\mathbf{M}^\star))^2} \\
&= \|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{F}}^2 \cdot \left(\frac{4\sigma_{k+1}(\mathbf{M}^\star)\sigma_{k}(\hat{\mathbf{M}})}{(\sigma_{k}(\hat{\mathbf{M}}) - \sigma_{k+1}(\mathbf{M}^\star))^2} + \sum_{i=1}^{\ell-1} \frac{4\sigma_{k+1}(\mathbf{M}^\star)\sigma_{k_i}(\hat{\mathbf{M}})}{(\sigma_{k_i}(\hat{\mathbf{M}}) - \sigma_{k+1}(\mathbf{M}^\star))^2} \right) \\
&\overset{(i)}{\le} \|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{F}}^2 \cdot \left(\frac{4\sigma_{k+1}(\mathbf{M}^\star)\sigma_{k}(\hat{\mathbf{M}})}{(\sigma_{k}(\hat{\mathbf{M}}) - \sigma_{k+1}(\mathbf{M}^\star))^2} + \sum_{i=1}^{\ell-1} \frac{4\sigma_{k}(\hat{\mathbf{M}})\sigma_{k_i}(\hat{\mathbf{M}})}{(\sigma_{k_i}(\hat{\mathbf{M}}) - \sigma_{k}(\hat{\mathbf{M}}))^2} \right),
\end{aligned}$$

where in (i) we use that $\sigma_k(\hat{\mathbf{M}}) \ge \sigma_{k+1}(\mathbf{M}^\star)$. Using that $\sigma_{k_i}(\hat{\mathbf{M}}) \ge 2\sigma_{k_{i-1}}(\hat{\mathbf{M}}) \ge \dots 2^i\sigma_{k_0}(\hat{\mathbf{M}}) = 2^i\sigma_k(\hat{\mathbf{M}})$, we find

$$\begin{aligned}
\sum_{i=1}^{\ell-1} \frac{4\sigma_{k}(\hat{\mathbf{M}})\sigma_{k_i}(\hat{\mathbf{M}})}{(\sigma_{k_i}(\hat{\mathbf{M}}) - \sigma_{k}(\hat{\mathbf{M}}))^2} &= \sum_{i=1}^{\ell-1} \frac{4\sigma_{k}(\hat{\mathbf{M}})}{(\sigma_{k_i}(\hat{\mathbf{M}}) - \sigma_{k}(\hat{\mathbf{M}}))(1 - \sigma_k(\hat{\mathbf{M}})/\sigma_{k_i}(\hat{\mathbf{M}}))} \\
&\le \sum_{i=1}^{\ell-1} \frac{4\sigma_{k}(\hat{\mathbf{M}})}{(2^i - 1)\sigma_k(\hat{\mathbf{M}})(1 - 2^{-i})} = \sum_{i=1}^{\ell-1} \frac{4}{(2^i - 1)(1 - 2^{-i})} \\
&\le \sum_{i \ge 1} \frac{4}{(2^i - 1)(1 - 2^{-i})} \le 16.
\end{aligned}$$

Hence, we conclude

$$\begin{aligned}
|\langle \hat{\mathbf{M}}_{[k]}, \mathbf{M}^\star_{>k}\rangle| &\le \|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{F}}^2 \cdot \left(\frac{4\sigma_k(\hat{\mathbf{M}})^2}{(\sigma_{k}(\hat{\mathbf{M}}) - \sigma_{k+1}(\mathbf{M}^\star))^2} + 16 \right) \\
&= 4\|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{F}}^2 \cdot \left(\left(1 - \frac{\sigma_{k+1}(\mathbf{M}^\star)}{\sigma_k(\hat{\mathbf{M}})}\right)^{-2} + 4 \right),
\end{aligned}$$

completing the proof.

D.1. Useful Variants of Wedin's Theorem

Lemma D.1 ("Gap-Free" Davis Kahan, Lemma B.3, Allen-Zhu and Li (2016)) *Let $\|\cdot\|_\circ$ denote any Schatten p-norm. Fix $\epsilon > 0$, and suppose that $\mathbf{X}, \tilde{\mathbf{X}}$ are symmetric matrices with $\|\mathbf{X} - \tilde{\mathbf{X}}\|_\circ \le \epsilon$. Given $\mu \ge 0$ and $\tau \ge 0$, let $\mathbf{U}_0$ be an orthonormal matrix with columns being the eigenvectors of $\mathbf{X}$, whose corresponding eigenvalues have absolutely value $\le \mu$, and $\tilde{\mathbf{U}}_1$ be an orthonormal matrix with columns being the eigenvectors of $\tilde{\mathbf{X}}$, whose corresponding eigenvalues have absolutely value $\ge \mu + \tau$. Then, $\|\mathbf{U}_0^\top \tilde{\mathbf{U}}_1\|_\circ \le \frac{\tau}{\epsilon}$.*

Proof We follow the proof of Lemma B.3, Allen-Zhu and Li (2016), originally stated in the operator norm (and for positive semidefinite matrices), to accommodate the Frobenius norm and absolute value eigenvalue magnitudes. Next, write out compact diagonalizations

$$\begin{aligned}
\mathbf{X} &= \mathbf{U}_0\boldsymbol{\Sigma}_0(\mathbf{U}_0)^\top + \mathbf{U}_1\boldsymbol{\Sigma}_1(\mathbf{U}_1)^\top \\
\tilde{\mathbf{X}} &= \tilde{\mathbf{U}}_0\tilde{\boldsymbol{\Sigma}}_0(\tilde{\mathbf{U}}_0)^\top + \tilde{\mathbf{U}}_1\tilde{\boldsymbol{\Sigma}}_1(\tilde{\mathbf{U}}_1)^\top,
\end{aligned}$$

where all entries of $\mathbf{\Sigma}_0$ lie in $[-\mu, \mu]$, and entries of $\mathbf{\Sigma}_1$ lie in $(-\infty, \mu) \cup (\mu, \infty)$, all entries of $\tilde{\mathbf{\Sigma}}_0$ lie in $(-(\mu+\tau), \mu+\tau)$, and entries of $\tilde{\mathbf{\Sigma}}_1$ are in $(-\infty, -(\mu+\tau)] \cup [\mu+\tau, \infty)$. Consider the residual $\mathbf{\Delta} := \mathbf{X} - \tilde{\mathbf{X}}$, we find that

$$\begin{aligned}\mathbf{\Sigma}_0 \mathbf{U}_0^\top &= \mathbf{U}_0^\top \mathbf{X} = \mathbf{U}_0^\top \tilde{\mathbf{X}} + \mathbf{U}_0^\top \mathbf{\Delta} \\ \text{implying } \mathbf{\Sigma}_0 \mathbf{U}_0^\top \tilde{\mathbf{U}}_1 &= \mathbf{U}_0^\top \tilde{\mathbf{X}} \tilde{\mathbf{U}}_1 + \mathbf{U}_0^\top \mathbf{\Delta} \tilde{\mathbf{U}}_1 \\ &= \mathbf{U}_0^\top \tilde{\mathbf{U}}_1 \tilde{\mathbf{\Sigma}}_1 + \mathbf{U}_0^\top \mathbf{\Delta} \tilde{\mathbf{U}}_1.\end{aligned}$$

Taking norms and applying the triangle inequality

$$\|\mathbf{\Sigma}_0 (\mathbf{U}_0)^\top \tilde{\mathbf{U}}_1\|_\circ \geq \|(\mathbf{U}_0)^\top \tilde{\mathbf{U}}_1 \tilde{\mathbf{\Sigma}}_1\|_\circ - \|(\mathbf{U}_0)^\top \mathbf{\Delta} \tilde{\mathbf{U}}_1\|_\circ.$$

Since $(\mathbf{\Sigma}_0)^\top (\mathbf{\Sigma}_0) \preceq \mu^2 \mathbf{I}$, and $\tilde{\mathbf{\Sigma}}_1^\top \tilde{\mathbf{\Sigma}}_1 \succeq (\mu+\tau)^2 \mathbf{I}$, and since $\mathbf{U}_0, \tilde{\mathbf{U}}_1$ are orthogonal, we estimate $\|\mathbf{\Sigma}_0 (\mathbf{U}_0)^\top \tilde{\mathbf{U}}_1\|_\circ \leq \mu \|\mathbf{U}_0^\top \tilde{\mathbf{U}}_1\|_\circ$, that $\|(\mathbf{U}_0)^\top \tilde{\mathbf{U}}_1 \tilde{\mathbf{\Sigma}}_1\|_\circ \geq (\mu+\tau) \|\mathbf{U}_0^\top \tilde{\mathbf{U}}_1\|_\circ$, and $\|(\mathbf{U}_0)^\top \mathbf{\Delta} \tilde{\mathbf{U}}_1\|_\circ \leq \|\mathbf{\Delta}\|_\circ$. Thus

$$\mu \|\mathbf{U}_0^\top \tilde{\mathbf{U}}_1\|_\circ \geq (\mu+\tau) \|\mathbf{U}_0^\top \tilde{\mathbf{U}}_1\|_\circ - \|\mathbf{\Delta}\|_\circ.$$

Rearranging concludes the proof. ■

Lemma D.2 (Variant of Wedin's Theorem) *Suppose that $\mathbf{M}, \tilde{\mathbf{M}} \in \mathbb{R}^{m \times n}$. Given $\mu \geq 0$ and $\tau \geq 0$, let $\mathbf{U}_0, \mathbf{V}_0$ be an orthonormal basis for left (resp. right) singular vectors of $\mathbf{M}$ whose corresponding singular values are $\leq \mu$, and let $\tilde{\mathbf{U}}_1, \tilde{\mathbf{V}}_1$ be the same for singular vectors of $\tilde{\mathbf{M}}$ whose corresponding singular values are $\geq \mu + \tau$. Then,*

$$\left(\|\mathbf{U}_0^\top \tilde{\mathbf{U}}_1\|_{\mathrm{F}}^2 + \|\mathbf{V}_0^\top \tilde{\mathbf{V}}_1\|_{\mathrm{F}}^2 \right)^{\frac{1}{2}} \leq \frac{2\|\mathbf{M} - \tilde{\mathbf{M}}\|_{\mathrm{F}}}{\tau}.$$

The same is true when the Frobenius norm is replaced by the operator norm.[4]

Proof Consider the matrices

$$\mathbf{X} = \begin{bmatrix} 0 & \mathbf{M} \\ \mathbf{M}^\top & 0 \end{bmatrix}, \quad \tilde{\mathbf{X}} = \begin{bmatrix} 0 & \tilde{\mathbf{M}} \\ (\tilde{\mathbf{M}})^\top & 0 \end{bmatrix}.$$

Letting $\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ and $\tilde{\mathbf{M}} = \tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{V}}^\top$, we observe that we can write

$$\mathbf{X} = \mathbf{W}\mathbf{\Lambda}\mathbf{W}^\top, \quad \mathbf{W} := \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{U} & \mathbf{U} \\ \mathbf{V} & -\mathbf{V}, \end{bmatrix}, \quad \mathbf{\Lambda} := \begin{bmatrix} \mathbf{\Sigma} & 0 \\ 0 & -\mathbf{\Sigma} \end{bmatrix}$$

and analogously for $\tilde{\mathbf{X}}$. Letting $\mathbf{M} = \mathbf{U}_0 \mathbf{\Sigma}_0 \mathbf{V}_0^\top + \mathbf{U}_1 \mathbf{\Sigma}_1 \mathbf{V}_1^\top$ decompose into singular values $\leq \mu$ and those $> \mu$, we can write

$$\begin{aligned}&\mathbf{X} = \mathbf{W}_0 \mathbf{\Lambda}_0 \mathbf{W}_0^\top + \mathbf{W}_1 \mathbf{\Lambda}_1 \mathbf{W}_1^\top, \\ &\mathbf{W}_0 = \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{U}_0 & 0 & \mathbf{U}_0 & 0 \\ \mathbf{V}_0 & 0 & -\mathbf{V}_0 & 0 \end{bmatrix}, \quad \mathbf{W}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 & \mathbf{U}_1 & 0 & \mathbf{U}_1 \\ 0 & \mathbf{V}_1 & 0 & -\mathbf{V}_1 \end{bmatrix},\end{aligned}$$

4. A similar bound can be established for arbitrary Schatten p-norms, albeit with a slightly worse constant.

where $\mathbf{\Lambda}_0$ has eigenvalues with absolute value $\le \mu$, and $\mathbf{\Lambda}_1$ eigenvalues with absolute value $> \mu$. Applying a similar decomposition to $\tilde{\mathbf{X}}$, we find that Lemma D.1 yields that, for $\|\cdot\|_\circ$ representing either the operator norm or Frobenius norm,

$$\left\|\mathbf{W}_0^\top \mathbf{W}_1\right\|_\circ \le \frac{\|\mathbf{X}-\tilde{\mathbf{X}}\|_{\mathrm{F}}}{\tau} = \frac{\sqrt{2}\|\mathbf{M}-\tilde{\mathbf{M}}\|_\circ}{\tau}. \tag{D.3}$$

On the other hand, we expand

$$\begin{aligned}
\left\|\mathbf{W}_0^\top \mathbf{W}_1\right\|_\circ &= \frac{1}{2}\left\|\begin{bmatrix}\mathbf{U}_0 & 0 & \mathbf{U}_0 & 0\\ \mathbf{V}_0 & 0 & -\mathbf{V}_0 & 0\end{bmatrix}^\top \begin{bmatrix}0 & \tilde{\mathbf{U}}_1 & 0 & \tilde{\mathbf{U}}_1\\ 0 & \tilde{\mathbf{V}}_1 & 0 & -\tilde{\mathbf{V}}_1\end{bmatrix}\right\|_\circ \\
&= \frac{1}{2}\left\|\begin{bmatrix}0 & \mathbf{U}_0^\top\tilde{\mathbf{U}}_1 + \mathbf{V}_0^\top\tilde{\mathbf{V}}_1 & 0 & \mathbf{U}_0^\top\tilde{\mathbf{U}}_1 - \mathbf{V}_0^\top\tilde{\mathbf{V}}_1\end{bmatrix}\right\|_\circ \\
&= \frac{1}{2}\left\|\begin{bmatrix}\mathbf{A}+\mathbf{B} & \mathbf{A}-\mathbf{B}\end{bmatrix}\right\|_\circ, \quad \mathbf{A} := \mathbf{U}_0^\top\tilde{\mathbf{U}}_1, \;\; \mathbf{B} := \mathbf{V}_0^\top\tilde{\mathbf{V}}_1.
\end{aligned}$$

When $\circ$ denotes the Frobenius norm, we use

$$\begin{aligned}
\left\|\begin{bmatrix}\mathbf{A}+\mathbf{B} & \mathbf{A}-\mathbf{B}\end{bmatrix}\right\|_{\mathrm{F}}^2 &= \langle \mathbf{A}+\mathbf{B}, \mathbf{A}+\mathbf{B}\rangle + \langle \mathbf{A}-\mathbf{B}, \mathbf{A}-\mathbf{B}\rangle \\
&= 2\langle \mathbf{A}, \mathbf{A}\rangle + 2\langle \mathbf{B}, \mathbf{B}\rangle \\
&= 2\left(\|\mathbf{A}\|_{\mathrm{F}}^2 + \|\mathbf{B}\|_{\mathrm{F}}^2\right).
\end{aligned}$$

Similarly, when $\circ$ denotes the operator norm,

$$\begin{aligned}
\left\|\begin{bmatrix}\mathbf{A}+\mathbf{B} & \mathbf{A}-\mathbf{B}\end{bmatrix}\right\|_{\mathrm{op}}^2 &= \max_{\mathbf{v}:\|\mathbf{v}\|=1} \|\mathbf{v}^\top(\mathbf{A}+\mathbf{B})\|_2^2 + \|\mathbf{v}^\top(\mathbf{A}-\mathbf{B})\|_2^2 \\
&= \max_{\mathbf{v}:\|\mathbf{v}\|=1} 2\mathbf{v}^\top\mathbf{A}\mathbf{A}^\top\mathbf{v} + 2\mathbf{v}^\top\mathbf{B}\mathbf{B}^\top\mathbf{v} + 2\mathbf{v}^\top\mathbf{A}\mathbf{A}^\top\mathbf{v} - 2\mathbf{v}^\top\mathbf{A}\mathbf{A}^\top\mathbf{v} \\
&= \max_{\mathbf{v}:\|\mathbf{v}\|=1} 2\mathbf{v}^\top\mathbf{A}\mathbf{A}^\top\mathbf{v} + 2\mathbf{v}^\top\mathbf{B}\mathbf{B}^\top\mathbf{v} \\
&\le 2\left(\|\mathbf{A}\|_{\mathrm{op}}^2 + \|\mathbf{B}\|_{\mathrm{op}}^2\right).
\end{aligned}$$

Thus,

$$\left\|\mathbf{W}_0^\top \mathbf{W}_1\right\|_\circ \le \frac{1}{\sqrt{2}}\left(\|\mathbf{U}_0^\top\tilde{\mathbf{U}}_1\|_\circ^2 + \|\mathbf{V}_0^\top\tilde{\mathbf{V}}_1\|_\circ^2\right)^{\frac{1}{2}}.$$

Plugging this into Eq. (D.3) concludes. ■

Part II
Supplement for Single- and Double-Stage ERM

Appendix E. Addenda for Single- And Double-Stage ERM (Theorem 3)

E.1. Single-stage ERM

Proof [Proof of Theorem 3] The first part of Theorem 3 follows directly from combining Theorem 2 and using a standard statistical training guarantee, Lemma F.1, to bound $\epsilon_{1\otimes 1}$ and ϵ_{trn}; Eq. (3.4) follows from a computation performed in Lemma E.1, below, and whose proof appears in Appendix H. ■

Lemma E.1 (Single Training Bound) *Under Assumption 2.6, we have*

$$\textsc{ApxErr}_{\mathrm{SS}}(r) \lesssim \begin{cases} C^2(1+\gamma^{-1})^2 r^{6-2\gamma} & \text{(polynomial decay)} \\ C^2 r^6(\gamma^{-1}+r)^2 e^{-2\gamma r} & \text{(exponential decay)} \end{cases}.$$

Remark E.1 (Sufficient Spectral Decay for α-Conditioning) . *For sufficiently rapid spectral decay, it* is *possible to ensure $(\hat{f}_{\mathrm{SS}}, \hat{g}_{\mathrm{SS}})$ are well-conditioned. From Lemma F.1, we have that with probability at least $1-\delta$,*

$$\mathcal{R}(\hat{f}_{\mathrm{SS}}, \hat{g}_{\mathrm{SS}}; \mathcal{D}_{1\otimes 1}) \le \kappa_{\mathrm{trn}}(2\kappa_{\mathrm{apx}}\mathsf{tail}_2^\star(r) + \frac{352B^4(\mathcal{M}_r + \log\frac{2}{\delta})}{n}).$$

In particular, if for a given $\alpha \ge 1$ it holds that

$$2\kappa_{\mathrm{trn}}\kappa_{\mathrm{apx}}\mathsf{tail}_2^\star(r) \le (1-\alpha^{-1})(\boldsymbol{\sigma}_r^\star)^2, \tag{E.1}$$

then, by letting $n \ge 352\alpha B^4(\mathcal{M}_r + \log\frac{2}{\delta})\boldsymbol{\sigma}_r^\star$, we can take $\epsilon_{1\otimes 1}^2 = \mathcal{R}(\hat{f}_{\mathrm{SS}}, \hat{g}_{\mathrm{SS}}; \mathcal{D}_{1\otimes 1}) \le (1-(2\alpha)^{-1})(\boldsymbol{\sigma}_r^\star)^2$. By Theorem 2, this implies that $(\hat{f}_{\mathrm{SS}}, \hat{g}_{\mathrm{SS}})$ are 2α-conditioned. Thus, when the tail of the spectrum at r is considerably smaller than $(\boldsymbol{\sigma}_r^\star)^2$, we can ensure that $(\hat{f}_{\mathrm{SS}}, \hat{g}_{\mathrm{SS}})$ are well-conditioned.

Eq. (E.1) requires rather rapid spectral decay, and will not hold for polynomially decaying singular values (e.g. $\boldsymbol{\sigma}_r^\star = r^{-(1+\gamma)}$). Under the exponential decay regime of Assumption 2.6 (for all n, $\boldsymbol{\sigma}_n^\star \le Ce^{-\gamma n}$), Lemma H.1 implies that $\mathsf{tail}_2^\star(r) \le C^2(1+\gamma^{-1})e^{-2\gamma(r+1)}$ (which is more-or-less tight in the worst case). Thus, Eq. (E.1) holds as soon as

$$2\kappa_{\mathrm{trn}}\kappa_{\mathrm{apx}}(1+\gamma^{-1})e^{-2\gamma} \le (1-\alpha^{-1})e^{2\gamma r}\left(\frac{\sigma^\star}{C}\right)^2. \tag{E.2}$$

Now assume that a lower bound for spectral decay also holds: for some other constant c, we have $\boldsymbol{\sigma}_r^\star \ge ce^{-r\gamma}$. Then, Eq. (E.2) holds as soon as

$$2\kappa_{\mathrm{trn}}\kappa_{\mathrm{apx}}(1+\gamma^{-1})e^{-2\gamma} \le (1-\alpha^{-1})\left(\frac{c}{C}\right)^2, \tag{E.3}$$

which is true once $\gamma > \log(\frac{2\sqrt{2}C\kappa_{\mathrm{trn}}\kappa_{\mathrm{apx}}}{c(1-\alpha^{-1})})$. In summary, we can ensure well-conditioned $(\hat{f}_{\mathrm{SS}}, \hat{g}_{\mathrm{SS}})$ when (a) there is rapid, exponential spectral decay and (b) a lower bound on the spectral decay as well.

E.2. Double-stage ERM (Theorem 4)

Here, we present Theorem 6, a more detailed version of Theorem 4 which specifies the necessary setting of algorithm parameters. We then specialize Theorem 6 to Theorem 4 at the end of the section. These two aforementioned conditions are specified in the following two conditions.

Condition E.1 (Algorithm Parameters) *Let c_1 be some unspecified parameter satisfying $1 \le c_1 \lesssim_\star 1$. We stipulate that the algorithm parameters $(\sigma_{\mathrm{cut}}, r_{\mathrm{cut}}, p)$ satisfy*

(a) $r_{\mathrm{cut}} \ge c_1$ *and* $\mathsf{tail}_2^\star(r_{\mathrm{cut}}) \le \frac{1}{c_1} r_{\mathrm{cut}}^2 (\sigma_{\mathrm{cut}})^2$;

(b) $\mathsf{tail}_2^\star(p) \le \frac{1}{c_1}\frac{\sigma_{\mathrm{cut}}^2}{r_{\mathrm{cut}}^5}$;

(c) $\sigma_{\mathrm{cut}} \in [2\boldsymbol{\sigma}_{r_{\mathrm{cut}}}^\star, \frac{2}{3e}\boldsymbol{\sigma}_1^\star]$.

Condition E.2 (Sample Size Conditions) *Let c_2 be some unspecified parameter satisfying $c_2 \lesssim_\star 1$. We stipulate that, given $\delta \in (0, 1)$,*

- *The supervised sample sizes of n_1, n_3 satisfy*

$$n_1 \ge p + B^4 c_2(\mathcal{M}_p + \log\frac{1}{\delta}) r_{\mathrm{cut}}^4 \sigma_{\mathrm{cut}}^{-2}, \quad n_3 \ge c_2 B^4(\mathcal{M}_{r_{\mathrm{cut}}} + \log\frac{1}{\delta})\sigma_{\mathrm{cut}}^{-2}$$

- *The unsupervised sample sizes n_2, n_4 satisfy*

$$n_2 \ge 722 r_{\mathrm{cut}}^2 n_1^9 \log(24p/\delta), \quad n_4 \ge r_{\mathrm{cut}}^4 n_1 n_3.$$

Note that when $r_{\mathrm{cut}} \le p$ (and hence $\mathcal{M}_{r_{\mathrm{cut}}} \le \mathcal{M}_p$), it suffices take $n_i \lesssim_\star \mathrm{poly}(p, \mathcal{M}_p, \log(1/\delta), B, \sigma_{\mathrm{cut}}^{-2})$.

Our main detailed theorem is as follows, and its proof is given in Appendix F.

Theorem 6 *Suppose Algorithm 1 is run with parameters $\sigma_{\mathrm{cut}}, r_{\mathrm{cut}}, p$, sample sizes $n_1, \dots, n_4$, and $\lambda = r_{\mathrm{cut}}^4$, $\mu = B^2/n_1$ and fix a probability of error $\delta \in (0, 1)$. Then, as long $\sigma_{\mathrm{cut}}, r_{\mathrm{cut}}, p$ satisfy Condition E.1 and $n_{1:4}$ satisfy Condition E.2, it holds with probability at least $1 - \delta$,*

$$\mathcal{R}(\hat{f}, \hat{g}; \mathcal{D}_{\mathrm{test}}) \lesssim_\star \mathrm{ERR}_{\mathrm{DT}}(r_{\mathrm{cut}}, \sigma_{\mathrm{cut}}) := r_{\mathrm{cut}}^2\sigma_{\mathrm{cut}}^2 + \mathsf{tail}_1^\star(r_{\mathrm{cut}})^2 + \frac{\mathsf{tail}_2^\star(r_{\mathrm{cut}})^2}{(\sigma_{\mathrm{cut}})^2}.$$

In Appendix H.2, we prove the following lemma. It gives an upper bound $\mathrm{ERR}_{\mathrm{DT}}(r_{\mathrm{cut}}, \sigma_{\mathrm{cut}})$, as well as sufficient conditions for Condition E.1, under the spectral decay assumption in Assumption 2.6.

Lemma E.2 (Double Training Decay Bounds) *Suppose Assumption 2.6 holds, and that the algorithm parameters $\sigma_{\mathrm{cut}}, r_{\mathrm{cut}}, p$ satisfy $\sigma_{\mathrm{cut}} \le \frac{2}{3e}\boldsymbol{\sigma}_1^\star$, and the following (feasible) constraints*

$$r_{\mathrm{cut}} \ge c_1 \vee \begin{cases} \frac{3eC}{\boldsymbol{\sigma}_1^\star} \\ \sqrt{c_1(1+\frac{1}{\gamma})} \vee \frac{1}{\gamma}\log(\frac{3eC}{\boldsymbol{\sigma}_1^\star}) \end{cases} \quad p \ge \begin{cases} c_1^{-\frac{1}{1+2\gamma}} r_{\mathrm{cut}}^{\frac{7+5\gamma}{1+2\gamma}} \\ 2r_{\mathrm{cut}} \vee \frac{1}{\gamma}\log(r_{\mathrm{cut}}^5 c_1) \end{cases} \quad \sigma_{\mathrm{cut}} \ge \begin{cases} 2Cr_{\mathrm{cut}}^{-(1+\gamma)} \\ 2Ce^{-\gamma r_{\mathrm{cut}}} \end{cases},$$

where the top-case correponds to the polynomial-decay regime, and the bottom to exponential-decay. Then, Condition E.1 holds and

$$\mathrm{ERR}_{\mathrm{DT}}(r_{\mathrm{cut}}, \sigma_{\mathrm{cut}}) \lesssim \sigma_{\mathrm{cut}}^2 r_{\mathrm{cut}}^2 + C^2(1+\gamma^{-2}) \begin{cases} r_{\mathrm{cut}}^{-2\gamma} & \textit{(polynomial decay)} \\ e^{-2\gamma r_{\mathrm{cut}}} & \textit{(exponential decay)} \end{cases}. \tag{E.4}$$

Proof [Proof of Theorem 4] For the target accuracy ϵ, set $\sigma_{\mathrm{cut}} = \max\{2Cr_{\mathrm{cut}}^{-(1+\gamma)}, \epsilon/r_{\mathrm{cut}}\}$ under polynomial spectral decay, and $\sigma_{\mathrm{cut}} = \max\{2Ce^{-\gamma r_{\mathrm{cut}}}, \epsilon/r_{\mathrm{cut}}\}$. From Eq. (E.5) and absorbing absolute constants into $\lesssim$, it then follows that

$$\mathrm{ERR}_{\mathrm{DT}}(r_{\mathrm{cut}}, \sigma_{\mathrm{cut}}) \lesssim \epsilon^2 + C^2(1+\gamma^{-2}) \begin{cases} r_{\mathrm{cut}}^{-2\gamma} & \text{(polynomial decay)} \\ e^{-2\gamma r_{\mathrm{cut}}} & \text{(exponential decay)} \end{cases}. \tag{E.5}$$

From Lemma E.2, Condition E.1 holds as soon as $r_{\mathrm{cut}} \gtrsim_\star \mathrm{poly}(C/\boldsymbol{\sigma}_1^\star, \gamma^{-1})$ and $p \lesssim_\star (r_{\mathrm{cut}})^c$ for a universal $c > 0$ (note that, in the constraint on p in polynomial case, the ratio $\frac{7+5\gamma}{1+2\gamma}$ is at most 7). Moreover, there exist sample sizes $n_1, n_2, n_3, n_4 \lesssim_\star \mathrm{poly}(p, \mathcal{M}_p, \log(1/\delta), B, \epsilon^{-2})$ which ensure Condition E.2. The result now follows from Theorem 6 above. ■

E.3. Generalizing unsupervised access to $\mathcal{D}_{1\otimes 1}$ (Assumption 3.1)

In this section, we argue that if we replace $\mathcal{D}_{\mathcal{X},1}, \mathcal{D}_{\mathcal{Y},1}$ with any other distribution $\tilde{\mathcal{D}}_{\mathcal{X},1}, \tilde{\mathcal{D}}_{\mathcal{Y},1}$ satisfying for some $\tilde{\kappa} \ge 1$ the inequalities

$$\tilde{\kappa}^{-1} \le \frac{\mathrm{d}\tilde{\mathcal{D}}_{\mathcal{X},1}(x)}{\mathrm{d}\mathcal{D}_{\mathcal{X},1}(x)} \le \tilde{\kappa}, \quad \tilde{\kappa}^{-1} \le \frac{\mathrm{d}\tilde{\mathcal{D}}_{\mathcal{Y},1}(y)}{\mathrm{d}\mathcal{D}_{\mathcal{Y},1}(y)} \le \tilde{\kappa}, \tag{E.6}$$

and if the function classes $\mathcal{F}_k, \mathcal{G}_k$ are sufficiently expressive, then all of our problem assumptions remain true, up to multiplicative constants in $\tilde{\kappa}$. In particular, this means that, for any target distributions $\mathcal{D}_{\mathcal{X},1}, \mathcal{D}_{\mathcal{Y},1}$, we can replace the oracle in Assumption 3.1 with the one that samples from $\tilde{\mathcal{D}}_{1\otimes 1} := \tilde{\mathcal{D}}_{\mathcal{X},1} \otimes \tilde{\mathcal{D}}_{\mathcal{Y},1}$. We now go through each assumption in sequence.

- First, Assumption 2.1 is unaffected.
- Second, let us consider the covariance $\boldsymbol{\Sigma}_{f^\star} = \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X},1}}\left[f^\star(x)f^\star(x)^\top\right]$ and $\boldsymbol{\Sigma}_{g^\star} = \mathbb{E}_{y\sim\mathcal{D}_{\mathcal{Y},1}}\left[g^\star(y)g^\star(y)^\top\right]$. Uder assumption Assumption 2.4, $\boldsymbol{\Sigma}_{f^\star} = \boldsymbol{\Sigma}_{g^\star}$, and $\|f^\star(x)\|_{\mathcal{H}} \vee \|g^\star(x)\|_{\mathcal{H}} \le B$. Introduce as well $\tilde{\boldsymbol{\Sigma}}_{f^\star} = \mathbb{E}_{x\sim\tilde{\mathcal{D}}_{\mathcal{X},1}}\left[f^\star(x)f^\star(x)^\top\right]$ and $\tilde{\boldsymbol{\Sigma}}_{g^\star} = \mathbb{E}_{y\sim\tilde{\mathcal{D}}_{\mathcal{Y},1}}\left[g^\star(y)g^\star(y)^\top\right]$. Then, Eq. (E.6) implies that

  $$\tilde{\kappa}^{-1}\boldsymbol{\Sigma}_{f^\star} \preceq \tilde{\boldsymbol{\Sigma}}_{f^\star} \preceq \tilde{\kappa}\boldsymbol{\Sigma}_{f^\star}, \quad \tilde{\kappa}^{-1}\boldsymbol{\Sigma}_{g^\star} \preceq \tilde{\boldsymbol{\Sigma}}_{g^\star} \preceq \tilde{\kappa}\boldsymbol{\Sigma}_{g^\star}. \tag{E.7}$$

  Using $\boldsymbol{\Sigma}_{f^\star} = \boldsymbol{\Sigma}_{g^\star}$, we have

  $$\tilde{\kappa}^{-2}\tilde{\boldsymbol{\Sigma}}_{f^\star} \preceq \tilde{\boldsymbol{\Sigma}}_{g^\star} \preceq \tilde{\kappa}^2\tilde{\boldsymbol{\Sigma}}_{f^\star}$$

  By generalizing Lemma L.1(i&iv) to linear operators, we can construct a transformation an invertible $\mathbf{W}$ such that $\tilde{\kappa}^{-1}\mathbf{I} \preceq \mathbf{W} \preceq \tilde{\kappa}\mathbf{I}$ and

  $$\mathbf{W}\tilde{\boldsymbol{\Sigma}}_{f^\star}\mathbf{W} = \tilde{\boldsymbol{\Sigma}}_{g^\star}$$

  Hence, if we define the operator $\mathbf{T} = \mathbf{W}^{1/2}$ and set

  $$\tilde{f}^\star := \mathbf{T}f^\star, \quad \tilde{g}^\star := \mathbf{T}g^\star,$$

then $(\tilde{f}^\star, \tilde{g}^\star)$ are balanced:

$$\tilde{\mathbf{\Sigma}}_{1\otimes 1} := \mathbb{E}_{x\sim\tilde{\mathcal{D}}_{\mathcal{X},1}}\left[f^\star(x)f^\star(x)^\top\right] = \mathbb{E}_{y\sim\tilde{\mathcal{D}}_{\mathcal{Y},1}}\left[\tilde{g}^\star(y)\tilde{g}^\star(y)^\top\right].$$

Moreover, as $\tilde{\kappa}^{-1}\mathbf{I} \preceq \mathbf{W} \preceq \tilde{\kappa}\mathbf{I}$, $\tilde{\kappa}^{-1/2}\mathbf{I} \preceq \mathbf{T} \preceq \tilde{\kappa}^{1/2}\mathbf{I}$, so that

$$\sup_{x,y} \|\tilde{f}^\star(x)\|_{\mathcal{H}} \vee \|\tilde{g}^\star(x)\|_{\mathcal{H}} \le \sqrt{\tilde{\kappa}} \sup_{x,y} \|f^\star(x)\|_{\mathcal{H}} \vee \|g^\star(x)\|_{\mathcal{H}} \le \sqrt{\tilde{\kappa}}B;$$

that is, Assumption 2.4 holds with upper bound $\tilde{B} = \sqrt{\tilde{\kappa}}B$.

- One can directly check from Eq. (E.6) that replacing $\mathcal{D}_{\mathfrak{X},1} \leftarrow \tilde{\mathcal{D}}_{\mathcal{X},1}$ and $\mathcal{D}_{\mathfrak{Y},1} \leftarrow \tilde{\mathcal{D}}_{\mathcal{Y},1}$ ensures Assumption 2.2 holds with $\tilde{\kappa}_{\mathrm{trn}} = \tilde{\kappa}^2\kappa_{\mathrm{trn}}$ and $\tilde{\kappa}_{\mathrm{tst}} = \tilde{\kappa}^2\kappa_{\mathrm{tst}}$.

- Similarly, one can check that Assumption 2.3 with $\tilde{\kappa}_{\mathrm{cov}} \leftarrow \tilde{\kappa}\kappa_{\mathrm{cov}}$.

- The construction of $\tilde{\mathbf{\Sigma}}_{1\otimes 1}$ and Lemma L.1 (vii) imply

$$\lambda_i(\tilde{\mathbf{\Sigma}}_{1\otimes 1}) \le \sigma_i(\tilde{\mathbf{\Sigma}}_{f^\star}^{1/2}\tilde{\mathbf{\Sigma}}_{g^\star}^{1/2}) \le \sqrt{\lambda_i(\tilde{\mathbf{\Sigma}}_{f^\star})\lambda_i(\tilde{\mathbf{\Sigma}}_{g^\star})}.$$

  Using Eq. (E.7) to bound $\lambda_i(\tilde{\mathbf{\Sigma}}_{f^\star}) \le \tilde{\kappa}\lambda_i(\mathbf{\Sigma}_{f^\star}) = \lambda_i(\mathbf{\Sigma}^\star_{1\otimes 1})$ and similarly for $\tilde{\mathbf{\Sigma}}_{g^\star}$, we find

$$\lambda_i(\tilde{\mathbf{\Sigma}}_{1\otimes 1}) \le \tilde{\kappa}\lambda_i(\mathbf{\Sigma}^\star_{1\otimes 1}).$$

  Thus, Assumption 2.6 holds after inflacting the constant C by a factor of $\tilde{\kappa}$.

- In can be directly checked that Assumption 2.5 holds after replacing κ_{apx} with $\tilde{\kappa}_{\mathrm{apx}} := \tilde{\kappa}^2\kappa_{\mathrm{apx}}$.

- The last assumption, Assumption 2.7 needs to be modified so as to ensure the function classes $\mathcal{F}_k, \mathcal{G}_k$ are rich enough to express the rank-k projections $\tilde{f}^\star_k, \tilde{g}^\star_k$ (the analogues of $f^\star_k, g^\star_k$ defined in Section 2).

Appendix F. Analysis of the Algorithms

In this section, we provide analyses for the training algorithms we proposed. Appendix F.1 gives guarantees for a single stage of supervised ERM. Appendix F.2 establishes our main guarantee for Algorithm 1, Theorem 6, via a technical proposition Proposition F.1, whose proof is divided between the subsequent three sections.

F.1. Statistical guarantee for single-stage ERM

We present an analysis of a single phase of empirical risk minimization, which we use both to analyze the single-stage ERM, and to serve as the first step in our analysis of double-stage ERM. The following is proved in Appendix G.3, using a standard analysis of empirical risk minimization with the squared loss.

Lemma F.1 *Let $(\tilde{f}, \tilde{g}) \in \mathcal{F}_p \times \mathcal{G}_p$ be empirical risk minimizers on n_1 i.i.d. samples $(x_i, y_i, z_i) \sim \mathcal{D}_{\mathrm{train}}$. Then, for any $\delta \in (0,1)$, the followings hold with probability at least $1-\delta$:*

$$\mathcal{R}(\tilde{f}, \tilde{g}; \mathcal{D}_{\mathrm{train}}) \leq 2\mathcal{R}(f_p^\star, g_p^\star; \mathcal{D}_{\mathrm{train}}) + \frac{352B^4(\mathcal{M}_p + \log\frac{2}{\delta})}{n_1}$$
$$\mathcal{R}(\tilde{f}, \tilde{g}; \mathcal{D}_{\mathrm{train}}) \leq 2\kappa_{\mathrm{apx}}\mathsf{tail}_2^\star(p) + \frac{352B^4(\mathcal{M}_p + \log\frac{2}{\delta})}{n_1}$$
$$\mathcal{R}(\tilde{f}, \tilde{g}; \mathcal{D}_{1\otimes 1}) \leq \kappa_{\mathrm{trn}}(2\kappa_{\mathrm{apx}}\mathsf{tail}_2^\star(p) + \frac{352B^4(\mathcal{M}_p + \log\frac{2}{\delta})}{n_1}).$$

F.2. Proof overview of Theorem 6

To prove Theorem 6, we first demonstrate that a certain technical proposition Proposition F.1 which shows that (a) a good spectral event $\mathcal{E}_{\mathrm{spec}}$ holds, under which the rank $\hat{r}$ chosen by Algorithm 2 satisfies various convenient spectral properties, and (b) that the regularized risk optimized in the last line of Algorithm 1 is small. More precisely, we define:

Definition F.1 (Good Spectral Event) *For parameters $(\sigma_{\mathrm{cut}}, r_{\mathrm{cut}})$ used in Algorithm 1, we define $\mathcal{E}_{\mathrm{spec}}(\hat{r}, \sigma_{\mathrm{cut}}, r_{\mathrm{cut}})$ as the event that the following inequalities hold:*

$$\boldsymbol{\sigma}_{\hat{r}}^\star \geq \frac{3}{4}\sigma_{\mathrm{cut}}, \quad \boldsymbol{\sigma}_{\hat{r}+1}^\star \leq 3\sigma_{\mathrm{cut}}, \quad \boldsymbol{\sigma}_{\hat{r}}^\star - \boldsymbol{\sigma}_{\hat{r}+1}^\star \geq \frac{\boldsymbol{\sigma}_{\hat{r}}^\star}{3r_{\mathrm{cut}}}$$
$$\mathsf{tail}_2^\star(\hat{r}) \leq \mathsf{tail}_2^\star(r_{\mathrm{cut}}) + 9\sigma_{\mathrm{cut}}^2 r_{\mathrm{cut}}, \quad \mathsf{tail}_1^\star(\hat{r})^2 \leq 18r_{\mathrm{cut}}^2\sigma_{\mathrm{cut}}^2 + 2\mathsf{tail}_1^\star(r_{\mathrm{cut}})^2.$$

Our technical proposition is as follows.

Proposition F.1 *Suppose that the parameters in Algorithm 1 are chosen as $\mu = B^2/n_1$, and other parameters $(p, \sigma_{\mathrm{cut}}, r_{\mathrm{cut}})$, the sample sizes $n_1, \ldots, n_4$, and $\lambda > 0$ satisfy that for some $C \lesssim_\star 1$,*

- $\sigma_{\mathrm{cut}} \in [2\boldsymbol{\sigma}_{r_{\mathrm{cut}}}^\star, \frac{2}{3e}\boldsymbol{\sigma}_1^\star]$, $\mathsf{tail}_2^\star(p) \leq \frac{\sigma_{\mathrm{cut}}^2}{Cr_{\mathrm{cut}}^2}$, *and* $p \geq 2$;
- $n_1 \geq p + C\sigma_{\mathrm{cut}}^{-2}r_{\mathrm{cut}}^2 \max\{1, B^4\}(\mathcal{M}_p + \log\frac{1}{\delta})$, $n_2 \geq 722r_{\mathrm{cut}}^2 n_1^9 \log(24p/\delta)$, $n_4 \geq \lambda n_1 n_3$.

Then, with probability at least $1-\delta$, the event $\mathcal{E}_{\mathrm{spec}}(\hat{r}, \sigma_{\mathrm{cut}}, r_{\mathrm{cut}})$ holds and

$$\begin{aligned}&\mathcal{R}(\hat{f}_{\mathrm{DS}}, \hat{f}_{\mathrm{DS}}; \mathcal{D}_{\mathrm{train}}) + \lambda\mathcal{R}_{[r]}(\hat{f}_{\mathrm{DS}}, \hat{f}_{\mathrm{DS}}; \mathcal{D}_{1\otimes 1})\\ &\quad\lesssim_\star \mathsf{tail}_2^\star(r_{\mathrm{cut}}) + r_{\mathrm{cut}}\sigma_{\mathrm{cut}}^2 + \lambda r_{\mathrm{cut}}^2\mathsf{tail}_2^\star(p) + \frac{B^4(\mathcal{M}_{r_{\mathrm{cut}}} + \log(1/\delta))}{n_3} + \frac{\lambda r_{\mathrm{cut}}^2 B^4(\mathcal{M}_p + \log(1/\delta))}{n_1}.\end{aligned}$$

We will prove Proposition F.1 in Appendices F.3 and F.4, addressing the first and second phases of training in Algorithm 1 respectively. Using this result, we prove Theorem 6.

Proof [Proof of Theorem 6] Recall the statement of Theorem 2. It states that if $(\hat{f}_{\mathrm{DS}}, \hat{f}_{\mathrm{DS}})$ are $(\epsilon_{\mathrm{trn}}, \epsilon_{1\otimes 1})$-accurate, that $\epsilon_{1\otimes 1} \leq \min\{\boldsymbol{\sigma}_1^\star/40\hat{r}, \boldsymbol{\sigma}_{\hat{r}}^\star/4\}$, then we can bound $\alpha \leq 2$ and therefore bound

$$\mathcal{R}(\hat{f}_{\mathrm{DS}}, \hat{f}_{\mathrm{DS}}; \mathcal{D}_{\mathrm{test}}) \lesssim_\star \left\{\hat{r}^4\epsilon_{1\otimes 1}^2 + \mathsf{tail}_1^\star(\hat{r})^2 + \hat{r}^2(\boldsymbol{\sigma}_{\hat{r}+1}^\star)^2\right\} + \left\{\frac{(\hat{r}^3\epsilon_{1\otimes 1}^2 + \epsilon_{\mathrm{trn}}^2 + \mathsf{tail}_2^\star(\hat{r}))^2}{(\boldsymbol{\sigma}_{\hat{r}}^\star)^2}\right\}.$$

In particular, recall we select $\hat{r} \le r_{\mathrm{cut}}$ and $\lambda = r_{\mathrm{cut}}^4$. Then it suffices that $\epsilon_{1\otimes 1} \le \min\{\boldsymbol{\sigma}_1^\star/40 r_{\mathrm{cut}}, \boldsymbol{\sigma}_{r_{\mathrm{cut}}}^\star/4\}$ to ensure

$$\mathcal{R}(\hat{f}_{\mathrm{DS}}, \hat{f}_{\mathrm{DS}}; \mathcal{D}_{\mathrm{test}}) \lesssim_\star \left\{\lambda\epsilon_{1\otimes 1}^2 + \mathbf{tail}_1^\star(\hat{r})^2 + \hat{r}^2(\boldsymbol{\sigma}_{\hat{r}+1}^\star)^2\right\} + \left\{\frac{(\lambda\epsilon_{1\otimes 1}^2 + \epsilon_{\mathrm{trn}}^2 + \mathbf{tail}_2^\star(\hat{r}))^2}{(\boldsymbol{\sigma}_{\hat{r}}^\star)^2}\right\}.$$

On the event $\mathcal{E}_{\mathrm{spec}}(r, \sigma_{\mathrm{cut}}, r_{\mathrm{cut}})$, we can then bound

$$\begin{aligned}
\mathcal{R}(\hat{f}_{\mathrm{DS}}, \hat{f}_{\mathrm{DS}}; \mathcal{D}_{\mathrm{test}}) &\lesssim_\star \left\{\lambda\epsilon_{1\otimes 1}^2 + r_{\mathrm{cut}}^2\sigma_{\mathrm{cut}}^2 + \mathbf{tail}_1^\star(r_{\mathrm{cut}})^2 + \hat{r}^2\sigma_{\mathrm{cut}}^2\right\} \\
&\quad + \left\{\frac{(\lambda\epsilon_{1\otimes 1}^2 + \epsilon_{\mathrm{trn}}^2 + r_{\mathrm{cut}}(\sigma_{\mathrm{cut}})^2 + \mathbf{tail}_2^\star(r_{\mathrm{cut}}))^2}{(\sigma_{\mathrm{cut}})^2}\right\} \\
&\lesssim \left\{\lambda\epsilon_{1\otimes 1}^2 + r_{\mathrm{cut}}^2\sigma_{\mathrm{cut}}^2 + \mathbf{tail}_1^\star(r_{\mathrm{cut}})^2\right\} && (\hat{r} \le r_{\mathrm{cut}}) \\
&\quad + \left\{\frac{(\lambda\epsilon_{1\otimes 1}^2 + \epsilon_{\mathrm{trn}}^2 + r_{\mathrm{cut}}(\sigma_{\mathrm{cut}})^2 + \mathbf{tail}_2^\star(r_{\mathrm{cut}}))^2}{(\sigma_{\mathrm{cut}})^2}\right\}.
\end{aligned}$$

Next, we set $\epsilon_{\mathrm{trn}}^2 := \mathcal{R}(\hat{f}_{\mathrm{DS}}, \hat{f}_{\mathrm{DS}}; \mathcal{D}_{\mathrm{train}})$ and $\epsilon_{1\otimes 1}^2 := \mathcal{R}_{[r]}(\hat{f}_{\mathrm{DS}}, \hat{f}_{\mathrm{DS}}; \mathcal{D}_{1\otimes 1})$. Then, on the event of the conclusion of Proposition F.1, and using $\lambda = r_{\mathrm{cut}}^4$, we have

$$\begin{aligned}
&\lambda\epsilon_{1\otimes 1}^2 \le \lambda\epsilon_{1\otimes 1}^2 + \epsilon_{\mathrm{trn}}^2 \\
&\lesssim_\star \mathbf{tail}_2^\star(r_{\mathrm{cut}}) + r_{\mathrm{cut}}\sigma_{\mathrm{cut}}^2 + r_{\mathrm{cut}}^6\mathbf{tail}_2^\star(p) + \underbrace{\tfrac{B^4(\mathcal{M}_{r_{\mathrm{cut}}} + \log(1/\delta))}{n_3} + \tfrac{r_{\mathrm{cut}}^6 B^4(\mathcal{M}_p + \log(1/\delta))}{n_1}}_{\le 2\sigma_{\mathrm{cut}}^2 \text{ under Condition E.2}} \\
&\lesssim \mathbf{tail}_2^\star(r_{\mathrm{cut}}) + r_{\mathrm{cut}}\sigma_{\mathrm{cut}}^2 + r_{\mathrm{cut}}^6\mathbf{tail}_2^\star(p).
\end{aligned}$$

Plugging the former display into the one before it, and suppressing constants, we have

$$\begin{aligned}
\mathcal{R}(\hat{f}_{\mathrm{DS}}, \hat{f}_{\mathrm{DS}}; \mathcal{D}_{\mathrm{test}}) \lesssim_\star &\left\{r_{\mathrm{cut}}^6\mathbf{tail}_2^\star(p) + \mathbf{tail}_2^\star(r_{\mathrm{cut}}) + r_{\mathrm{cut}}\sigma_{\mathrm{cut}}^2 + r_{\mathrm{cut}}^2\sigma_{\mathrm{cut}}^2 + \mathbf{tail}_1^\star(r_{\mathrm{cut}})^2\right\} \\
&+ \left\{\frac{(r_{\mathrm{cut}}^6\mathbf{tail}_2^\star(p) + r_{\mathrm{cut}}(\sigma_{\mathrm{cut}})^2 + \mathbf{tail}_2^\star(r_{\mathrm{cut}}))^2}{(\sigma_{\mathrm{cut}})^2}\right\}.
\end{aligned}$$

In particular, if in addition it holds that

$$\mathbf{tail}_2^\star(p) \le \frac{(\sigma_{\mathrm{cut}})^2}{r^5}, \tag{F.1}$$

then

$$\begin{aligned}
\mathcal{R}(\hat{f}_{\mathrm{DS}}, \hat{f}_{\mathrm{DS}}; \mathcal{D}_{\mathrm{test}}) &\lesssim_\star \mathbf{tail}_2^\star(r_{\mathrm{cut}}) + r_{\mathrm{cut}}\sigma_{\mathrm{cut}}^2 + r_{\mathrm{cut}}^2\sigma_{\mathrm{cut}}^2 + \mathbf{tail}_1^\star(r_{\mathrm{cut}})^2\frac{(r_{\mathrm{cut}}(\sigma_{\mathrm{cut}})^2 + \mathbf{tail}_2^\star(r_{\mathrm{cut}}))^2}{(\sigma_{\mathrm{cut}})^2} \\
&\lesssim \mathbf{tail}_2^\star(r_{\mathrm{cut}}) + r_{\mathrm{cut}}^2\sigma_{\mathrm{cut}}^2 + \mathbf{tail}_1^\star(r_{\mathrm{cut}})^2 + \frac{\mathbf{tail}_2^\star(r_{\mathrm{cut}})^2}{(\sigma_{\mathrm{cut}})^2} \\
&\lesssim r_{\mathrm{cut}}^2\sigma_{\mathrm{cut}}^2 + \mathbf{tail}_1^\star(r_{\mathrm{cut}})^2 + \frac{\mathbf{tail}_2^\star(r_{\mathrm{cut}})^2}{(\sigma_{\mathrm{cut}})^2},
\end{aligned}$$

where in the last step, we use $\mathbf{tail}_1^\star(r_{\mathrm{cut}})^2 = (\sum_{i > r_{\mathrm{cut}}} \boldsymbol{\sigma}_i^\star)^2 \ge \sum_{i > r_{\mathrm{cut}}} (\boldsymbol{\sigma}_i^\star)^2 = \mathbf{tail}_2^\star(r_{\mathrm{cut}})$. so that, if the conditions on $n_{1:4}$ and p of Proposition F.1 are met, and if $\epsilon_{1\otimes 1} \le \min\{\boldsymbol{\sigma}_1^\star/40\hat{r}, \boldsymbol{\sigma}_{\hat{r}}^\star/4\}$, and if Eq. (F.1) holds, then with probability at least $1 - \delta$,

$$\mathcal{R}(\hat{f}_{\mathrm{DS}}, \hat{f}_{\mathrm{DS}}; \mathcal{D}_{\mathrm{test}}) \lesssim_\star r_{\mathrm{cut}}^2\sigma_{\mathrm{cut}}^2 + \mathbf{tail}_1^\star(r_{\mathrm{cut}})^2 + \frac{\mathbf{tail}_2^\star(r_{\mathrm{cut}})^2}{(\sigma_{\mathrm{cut}})^2}.$$

Checking the appropriate conditions. For Proposition F.1 to hold with $\lambda = r_{\mathrm{cut}}^4$, and for we need that for some $c_0 \lesssim_\star 1$,

$$n_1 \geq p + c_0\sigma_{\mathrm{cut}}^{-2} r_{\mathrm{cut}}^2 B^4(\mathcal{M}_p + \log\frac{1}{\delta}), \quad p \geq 2, \quad \mathsf{tail}_2^\star(p) \leq \frac{\sigma_{\mathrm{cut}}^2}{c_0 r_{\mathrm{cut}}^5}, \tag{F.2}$$

as well as

$$n_2 \geq 722 r_{\mathrm{cut}}^2 n_1^9 \log(24p/\delta), \quad n_4 \geq r_{\mathrm{cut}}^4 n_1 n_3, \qquad \mu = B^2/n_1. \tag{F.3}$$

All these conditions are ensured by Conditions E.1 and E.2.

Let us conclude by making explicit conditions under which $\epsilon_{1\otimes 1} \leq \min\{\boldsymbol{\sigma}_1^\star/40\hat{r}, \boldsymbol{\sigma}_{\hat{r}}^\star/4\}$ holds, provided the high-probablity event of Proposition F.1 holds. As $\hat{r} \leq r_{\mathrm{cut}}$, on the $\mathcal{E}_{\mathrm{spec}}(\hat{r}, \sigma_{\mathrm{cut}}, r_{\mathrm{cut}})$, it is enough that, for some small universal constant c,

$$\epsilon_{1\otimes 1}^2 \leq c \min\left\{\frac{(\boldsymbol{\sigma}_1^\star)^2}{r_{\mathrm{cut}}^2}, \sigma_{\mathrm{cut}}^2\right\}. \tag{F.4}$$

On the event of Proposition F.1, we would like to have

$$\epsilon_{1\otimes 1}^2 \lesssim_\star \frac{\mathsf{tail}_2^\star(r_{\mathrm{cut}})}{r_{\mathrm{cut}}^4} + \frac{\sigma_{\mathrm{cut}}^2}{r_{\mathrm{cut}}^3} + r_{\mathrm{cut}}^2 \mathsf{tail}_2^\star(p) + \frac{B^4(\mathcal{M}_{r_{\mathrm{cut}}} + \log(1/\delta))}{n_3 r_{\mathrm{cut}}^4} + \frac{r_{\mathrm{cut}}^2 B^4(\mathcal{M}_p + \log(1/\delta))}{n_1}.$$

By modifying $c_0 \lesssim_\star 1$ below if necessary, it suffices that for Eq. (F.4) that

$$\max\{\tfrac{\sigma_{\mathrm{cut}}^2}{r_{\mathrm{cut}}^3}, \tfrac{\mathsf{tail}_2^\star(r_{\mathrm{cut}})}{r_{\mathrm{cut}}^4}, r_{\mathrm{cut}}^2\mathsf{tail}_2^\star(p), \tfrac{B^4(\mathcal{M}_{r_{\mathrm{cut}}}+\log(1/\delta))}{n_3 r_{\mathrm{cut}}^4}, \tfrac{r_{\mathrm{cut}}^2 B^4(\mathcal{M}_p+\log(1/\delta))}{n_1}\} \leq \frac{1}{c_0}\min\left\{\frac{(\boldsymbol{\sigma}_1^\star)^2}{r_{\mathrm{cut}}^2}, \sigma_{\mathrm{cut}}^2\right\}.$$

We handle each term in sequence,

1. As $\sigma_{\mathrm{cut}} \leq \boldsymbol{\sigma}_1^\star$, we have $\frac{\sigma_{\mathrm{cut}}^2}{r_{\mathrm{cut}}^3} \leq \frac{1}{c_0}\min\left\{\frac{(\boldsymbol{\sigma}_1^\star)^2}{r_{\mathrm{cut}}^2}, \sigma_{\mathrm{cut}}^2\right\}$ as soon as $r_{\mathrm{cut}} \geq c_0$.
2. The term $\frac{\mathsf{tail}_2^\star(r_{\mathrm{cut}})}{r_{\mathrm{cut}}^4}$ is appropriately bounded as soon as $\mathsf{tail}_2^\star(r_{\mathrm{cut}}) \leq \frac{1}{c_0}\min\left\{r_{\mathrm{cut}}^2(\boldsymbol{\sigma}_1^\star)^2, r_{\mathrm{cut}}^4\sigma_{\mathrm{cut}}^2\right\}$. Under the condition that $\sigma_{\mathrm{cut}} \leq \boldsymbol{\sigma}_1^\star$, it suffices that $\mathsf{tail}_2^\star(r_{\mathrm{cut}}) \leq \frac{1}{c_0} r_{\mathrm{cut}}^2\sigma_{\mathrm{cut}}^2$.
3. The term $r_{\mathrm{cut}}^2\mathsf{tail}_2^\star(p)$ is appropriately bounded as soon as $\mathsf{tail}_2^\star(p) \leq \frac{1}{c_0}\min\left\{\frac{(\boldsymbol{\sigma}_1^\star)^2}{r_{\mathrm{cut}}^4}, \sigma_{\mathrm{cut}}^2/r_{\mathrm{cut}}^2\right\}$. As $\sigma_{\mathrm{cut}} \leq \boldsymbol{\sigma}_1^\star$, this holds when $\mathsf{tail}_2^\star(p) \leq \frac{\sigma_{\mathrm{cut}}^2}{c_0 r_{\mathrm{cut}}^5}$.
4. The term $\frac{B^4(\mathcal{M}_{r_{\mathrm{cut}}}+\log(1/\delta))}{n_3 r_{\mathrm{cut}}^4}$ is appropriately bounded as soon as
$$n_3 \geq c_0 B^4(\mathcal{M}_{r_{\mathrm{cut}}} + \log(1/\delta))\left\{\frac{1}{(\boldsymbol{\sigma}_1^\star)^2 r_{\mathrm{cut}}^2} + \frac{1}{\sigma_{\mathrm{cut}}^2 r_{\mathrm{cut}}^4}\right\}.$$
5. Similarly, term $\frac{r_{\mathrm{cut}}^2 B^4(\mathcal{M}_p+\log(1/\delta))}{n_1}$ is appropriately bounded as soon as (adding an additive p for convenience),
$$n_1 \geq p + B^4 c_0(\mathcal{M}_p + \log(1/\delta))\left\{\frac{r_{\mathrm{cut}}^4}{(\boldsymbol{\sigma}_1^\star)^2} + \frac{r_{\mathrm{cut}}^2}{\sigma_{\mathrm{cut}}^2}\right\}.$$
For which, using $\sigma_{\mathrm{cut}} \leq \boldsymbol{\sigma}_1^\star$, it suffices that
$$n_1 \geq p + B^4 c_0(\mathcal{M}_p + \log(1/\delta))\frac{r_{\mathrm{cut}}^4}{\sigma_{\mathrm{cut}}^2}.$$

All such bounds hold under Conditions E.1 and E.2. This completes the proof of Theorem 6. ■

F.3. Analysis of the first phase of double-stage ERM

We begin with a precise analysis of the first phase of the double-stage ERM Algorithm 1. Recall that $(\tilde{f}, \tilde{g})$ are the empirical risk minimizers on n_1 i.i.d. samples $(x_i, y_i, z_i) \sim \mathcal{D}_{\mathrm{train}}$, and $\hat{\mathbf{Q}}_{\hat{r}}$ is the balancing projection on the top r eigenvectors of $\hat{\mathbf{\Sigma}}_{\tilde{g}}$. We define the following effective error term.

$$\tilde{\epsilon}(p, n_1, \delta)^2 := \kappa_{\mathrm{trn}} \left(2\kappa_{\mathrm{apx}} \textsf{tail}_2^\star(p) + \frac{354B^4(\mathcal{M}_p + \log \frac{6}{\delta})}{n_1} \right). \tag{F.5}$$

We first show that $(\tilde{f}, \hat{\mathbf{Q}}_{\hat{r}}\tilde{g})$ has small risk on the top block.

Proposition F.2 (Guarantee for Double-Training, First-Phase) *Suppose $\sigma_{\mathrm{cut}} \in [2\boldsymbol{\sigma}^\star_{r_{\mathrm{cut}}}, \frac{2}{3e}\boldsymbol{\sigma}^\star_1]$, $n_1 \geq p \geq 2$, $\mu = B^2/n_1$, and both $n_1 \geq B^2/\sigma_{\mathrm{cut}}^2$ and $n_2 \geq 722 r_{\mathrm{cut}}^2 n_1^9 \log(24p/\delta)$. Further, suppose*

$$\tilde{\epsilon}(p, n_1, \delta)^2 \leq \sigma_{\mathrm{cut}}^2/(64 r_{\mathrm{cut}}^2). \tag{F.6}$$

Then, with probability at least $1 - \frac{2}{3}\delta$, we have

$$\mathcal{R}_{[\hat{r}]}(\tilde{f}, \hat{\mathbf{Q}}_{\hat{r}}\tilde{g}; \mathcal{D}_{1\otimes 1}) \leq 3000 r_{\mathrm{cut}}^2 \tilde{\epsilon}(p, n_1, \delta)^2.$$

Moreover, on this same event, both $\sup_{x,y} |\langle \tilde{f}(x), \hat{\mathbf{Q}}_{\hat{r}}\tilde{g}(y)\rangle| \leq \sqrt{2n_1}B^2$ and $\mathcal{E}_{\mathrm{spec}}(\hat{r}, \sigma_{\mathrm{cut}}, r_{\mathrm{cut}})$, defined in Definition F.1, holds.

F.3.1. PROOF OVERVIEW

Our first step is to verify the performance of the overparametrized $(\tilde{f}, \tilde{g})$ on the nominal distribution $\mathcal{D}_{1\otimes 1}$. For convenience, we upper bound a slightly augmented quantity which absorbs errors from regularizing the balancing covariances.

Lemma F.2 *Recall $\tilde{\epsilon}(\cdot)$ defined in Eq. (F.5). With probability at least $1 - \frac{1}{3}\delta$, it holds that $\mathcal{R}(\tilde{f}, \tilde{g}; \mathcal{D}_{1\otimes 1}) + \frac{2B^4}{n_1} \leq \tilde{\epsilon}(p, n_1, \delta)^2$.*

The above lemma is a direct consequence of the last line of Lemma F.1.

Our next goal is to find a good rank-$\hat{r}$ projection of the functions $(\tilde{f}, \tilde{g})$ which enjoys good performance on $\mathcal{D}_{1\otimes 1}$. This projection is best computed in a coordinate system in which $\tilde{f}, \tilde{g}$ are balanced in the sense of Definition 4.1: that is, under a transformation $\mathbf{T}$ such that $f = \mathbf{T}^{-\top}\tilde{f}$ and $g = \mathbf{T}\tilde{g}$, it holds that $\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[ff^\top] = \mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}[gg^\top]$. To compute this transformation, we first introduce sample and population covariance matrices.

Definition F.2 (Covariance Matrices) *Let $\{(x_{2,i}, x_{2,i})\}_{i=1}^{n_2} \overset{\text{i.i.d}}{\sim} \mathcal{D}_{1\otimes 1}$, we define the population covariance matrices $\mathbf{\Sigma}_{\tilde{f}} = \mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[\tilde{f}\tilde{f}^\top]$, $\quad \mathbf{\Sigma}_{\tilde{g}} = \mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}[\tilde{g}\tilde{g}^\top]$, and their finite sample analogues using the n_2 samples:*

$$\hat{\mathbf{\Sigma}}_{\tilde{f}} = \frac{1}{n_2}\sum_{i=1}^{n_2} \tilde{f}(x_{2,i})\tilde{f}(x_{2,i})^\top, \quad \hat{\mathbf{\Sigma}}_{\tilde{g}} = \frac{1}{n_2}\sum_{i=1}^{n_2} \tilde{g}(x_{2,i})\tilde{g}(x_{2,i})^\top.$$

Balancing then finds a transformation $\mathbf{T}$ for which $\mathbf{T}^{-\top}\mathbf{\Sigma}_{\tilde{f}}\mathbf{T}^{-1} = \mathbf{T}\mathbf{\Sigma}_{\tilde{g}}\mathbf{T}^{\top}$. It is challenging to establish a lower bound on $\lambda_{\min}(\mathbf{\Sigma}_{\tilde{f}})$ and $\lambda_{\min}(\mathbf{\Sigma}_{\tilde{g}})$, say when $\mathbf{\Sigma}^{\star}_{1\otimes 1}$ has rapid spectral decay. The matter only becomes worse when solving for $\mathbf{T}$ using the finite sample covariance matrices $\hat{\mathbf{\Sigma}}_{\tilde{f}}$ and $\hat{\mathbf{\Sigma}}_{\tilde{g}}$. As a consequence, we instead consider regularized covariance matrices, defined as follows:

Definition F.3 (Regularized Covariance Matrices) *Let* $\mu > 0$. *Define*

$$\mathbf{\Sigma}_{\tilde{f},\mu} = \mathbf{\Sigma}_{\tilde{f}} + \mu\mathbf{I}_p \quad \mathbf{\Sigma}_{\tilde{g},\mu} := \mathbf{\Sigma}_{\tilde{g}} + \mu\mathbf{I}_p \quad \hat{\mathbf{\Sigma}}_{\tilde{f},\mu} = \hat{\mathbf{\Sigma}}_{\tilde{f}} + \mu\mathbf{I}_p \quad \hat{\mathbf{\Sigma}}_{\tilde{g},\mu} = \hat{\mathbf{\Sigma}}_{\tilde{g}} + \mu\mathbf{I}_p.$$

Leveraging standard finite sample concentration inequality of matrices (see Lemma G.2 in the appendix), we ensure that the empirical and population covariance matrices concentrate.

Lemma F.3 *Let* $\{(x_{2,i}, y_{2,i})\}_{i=1}^{n_2} \overset{\text{i.i.d}}{\sim} \mathcal{D}_{1\otimes 1}$, *and define the following empirical and population covariance operators. Then, with probability at least* $1 - \frac{1}{3}\delta$, *we have*

$$\max\left\{\|\mathbf{\Sigma}_{\tilde{g},\mu} - \hat{\mathbf{\Sigma}}_{\tilde{g},\mu}\|_{\mathrm{op}}, \|\mathbf{\Sigma}_{\tilde{f},\mu} - \hat{\mathbf{\Sigma}}_{\tilde{f},\mu}\|_{\mathrm{op}}\right\} \le \epsilon_{\Sigma}(n_2, \delta) := B^2\sqrt{\frac{2\log(24p/\delta)}{n_2}}.$$

Moreover, for any $\mathbf{\Sigma} \in \{\mathbf{\Sigma}_{\tilde{g},\mu}, \hat{\mathbf{\Sigma}}_{\tilde{g},\mu}, \mathbf{\Sigma}_{\tilde{f},\mu}, \hat{\mathbf{\Sigma}}_{\tilde{f},\mu}\}$, *we have* $\mu\mathbf{I}_p \preceq \mathbf{\Sigma} \preceq B^2 + \mu\mathbf{I}_p$.

The above bound is proved for the non-regularized covariances, and follows by adding and subtracting $\mu\mathbf{I}_p$. The remainder of the proof has three components, each of which we give its own subsection below.

(a) We first show that the regularized covariance matrices can be thought of as *unregularized* covariance matrices corresponding to convolving the embeddings $(\tilde{f}, \tilde{g})$ with isotropic noise. We argue that the excess risk of these noisy embeddings, denoted by $(\tilde{f}_\mu, \tilde{g}_\mu)$, is $\mathcal{O}(\mu)$, and always upper bounds the risk of the noise-free embeddings. Hence, we can analyze balancing and projecting the noisy-embeddings as a proxy for the noise-free ones.

(b) We then analyze the performance of a balanced projection of the embeddings $(\tilde{f}, \tilde{g})$, and that of the projections of noisy embeddings $(\tilde{f}_\mu, \tilde{g}_\mu)$.

(c) We analyze the empirical balancing operator obtained via samples, and conclude the proof of Proposition F.2 by combining the above results.

F.3.2. INTERPRETING REGULARIZATION AS CONVOLUTION WITH NOISE

In this part of the proof, we illustrate how the *regularized* covariance matrices of $(\tilde{f}, \tilde{g})$ correspond to *unregularized* covariance matrices obtained by convolving $(\tilde{f}, \tilde{g})$ with noise. Let $\mathcal{K}_p := \{-1, 1\}^p$ denote the p-dimensional (boolean, centered) hypercube. We can augment $\mathcal{D}_{\mathfrak{X},1}$ and $\mathcal{D}_{\mathfrak{Y},1}$ to form distributions $\bar{\mathcal{D}}_{\mathcal{X},1}$ and $\bar{\mathcal{D}}_{\mathcal{Y},1}$ over $\mathcal{Y} \times \mathcal{K}_p$ and $\mathfrak{Y} \times \mathcal{K}_p$, where

$$\begin{aligned} \bar{x} &= (x, \check{\mathbf{x}}) \sim \bar{\mathcal{D}}_{\mathcal{X},1} \overset{\text{dist}}{=} x \sim \mathcal{D}_{\mathfrak{X},1} \perp \check{\mathbf{x}} \sim \mathsf{Unif}[\mathcal{K}_p] \\ \bar{y} &= (y, \check{\mathbf{y}}) \sim \bar{\mathcal{D}}_{\mathcal{Y},1} \overset{\text{dist}}{=} y \sim \mathcal{D}_{\mathfrak{Y},1} \perp \check{\mathbf{y}} \sim \mathsf{Unif}[\mathcal{K}_p]. \end{aligned} \tag{F.7}$$

On these augmented distributions, we define

$$\tilde{f}_\mu(x, \check{\mathbf{x}}) := \tilde{f}(x) + \sqrt{\mu}\check{\mathbf{x}}, \quad \tilde{g}_\mu(y, \check{\mathbf{y}}) = \tilde{g}(y) + \sqrt{\mu}\check{\mathbf{y}}.$$

We can readily verify that

$$\boldsymbol{\Sigma}_{\tilde{f},\mu} := \mathbb{E}_{\bar{\mathcal{D}}_{\mathcal{X},1}}[\tilde{f}_\mu \tilde{f}_\mu^\top] = \boldsymbol{\Sigma}_{\tilde{f}} + \mu \mathbf{I}_p, \quad \boldsymbol{\Sigma}_{\tilde{g},\mu} := \mathbb{E}_{\bar{\mathcal{D}}_{\mathcal{Y},1}}[\tilde{g}_\mu \tilde{g}_\mu^\top] = \boldsymbol{\Sigma}_{\tilde{g}} + \mu \mathbf{I}_p.$$

Two other observations are useful. In both, let $\bar{\mathcal{D}}_{1\otimes 1} := \bar{\mathcal{D}}_{\mathcal{X},1} \otimes \bar{\mathcal{D}}_{\mathcal{Y},1}$ (by analogy to $\mathcal{D}_{1\otimes 1}$), so that $\mathcal{R}(\tilde{f}_\mu, \tilde{g}_\mu; \bar{\mathcal{D}}_{1\otimes 1}) = \mathbb{E}[(\langle \tilde{f}_\mu(\bar{x}), \tilde{g}_\mu(\bar{y})\rangle - h^\star(x,y))^2]$. Then, the following bounds the excess risk of the regularized functions $\tilde{f}_\mu, \tilde{g}_\mu$ in terms of that of $\tilde{f}, \tilde{g}$:

Lemma F.4 *The following holds for any B-bounded $\tilde{f}, \tilde{g}$ and associated $\tilde{f}_\mu, \tilde{g}_\mu$:*

$$\mathcal{R}(\tilde{f}_\mu, \tilde{g}_\mu; \bar{\mathcal{D}}_{1\otimes 1}) \le \tilde{\epsilon}_\mu^2 := p\mu^2 + \mu B^2 + \mathcal{R}(\tilde{f}, \tilde{g}; \mathcal{D}_{1\otimes 1}).$$

In particular, if $n_1 \ge p$, then for $\mu \le B^2/n_1$, the functions $(\tilde{f}, \tilde{g})$ as in Lemma F.2 satisfy

$$\mathcal{R}(\tilde{f}_\mu, \tilde{g}_\mu; \bar{\mathcal{D}}_{1\otimes 1}) \le \tilde{\epsilon}_\mu^2 \le \tilde{\epsilon}(p, n_1, \delta)^2,$$

with probability at least $1 - \delta/3$.

Proof [Proof of Lemma F.4] Using independence of $x, y, \check{\mathbf{x}}, \check{\mathbf{y}}$ under $\bar{\mathcal{D}}_{1\otimes 1}$, and and $\mathbb{E}[\check{\mathbf{x}}\check{\mathbf{x}}^\top] = \mathbb{E}[\check{\mathbf{y}}\check{\mathbf{y}}^\top] = \mathbf{I}_p$, we have

$$\begin{aligned}
&\mathbb{E}_{\bar{\mathcal{D}}_{1\otimes 1}}[(\langle \tilde{f}_\mu(\bar{x}), \tilde{g}_\mu(\bar{y})\rangle - h^\star(x,y))^2] \\
&= \mathbb{E}_{\bar{\mathcal{D}}_{1\otimes 1}}[(\mu\langle \check{\mathbf{x}}, \check{\mathbf{y}}\rangle + \sqrt{\mu}\langle \check{\mathbf{x}}, \tilde{g}(y)\rangle + \sqrt{\mu}\langle \tilde{f}(x), \check{\mathbf{y}}\rangle + \langle \tilde{f}(x), \tilde{g}(y)\rangle - h^\star(x,y))^2] \\
&= \mu^2 \mathbb{E}_{\bar{\mathcal{D}}_{1\otimes 1}}[\langle \check{\mathbf{x}}, \check{\mathbf{y}}\rangle^2] + \mu \mathbb{E}_{\bar{\mathcal{D}}_{1\otimes 1}}[\langle \check{\mathbf{x}}, \tilde{g}(y)\rangle^2 + \langle \check{\mathbf{y}}, \tilde{f}(x)\rangle^2] + \mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle \tilde{f}(x), \tilde{g}(y)\rangle - h^\star(x,y))^2] \\
&= \mathrm{tr}(\mathbf{I}_p)\mu^2 + \mu \mathbb{E}_{\mathcal{D}_{1\otimes 1}}[\|f(x)\|^2 + \|g(y)\|^2] + \mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle \tilde{f}(x), \tilde{g}(y)\rangle - h^\star(x,y))^2] \\
&= p\mu^2 + \mu \mathbb{E}_{\mathcal{D}_{1\otimes 1}}[\|f(x)\|^2 + \|g(y)\|^2] + \mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle \tilde{f}(x), \tilde{g}(y)\rangle - h^\star(x,y))^2] \\
&\le p\mu^2 + \mu B^2 + \underbrace{\mathbb{E}[(\langle \tilde{f}(x), \tilde{g}(y)\rangle - h^\star(x,y))^2]}_{\mathcal{R}(\tilde{f}, \tilde{g}; \mathcal{D}_{1\otimes 1})}.
\end{aligned}$$

The second statement of the lemma follows from selecting $\mu \le B^2/n_1$, using the assumption that $n_1 \ge p$, and invoking Lemma F.2. ■

The second fact shows that weighted inner products involving the regularized functions are always worse predictors than the corresponding unregularized functions:

Lemma F.5 *The following inequality holds for any $\tilde{f}, \tilde{g}$ and associated $\tilde{f}_\mu, \tilde{g}_\mu$, matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$, and $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$:*

$$\mathbb{E}_{\bar{\mathcal{D}}_{1\otimes 1}}[(\langle \tilde{f}_\mu(\bar{x}), \mathbf{A}\tilde{g}_\mu(\bar{y})\rangle - h(x,y))^2] \ge \mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle \tilde{f}(x), \mathbf{A}\tilde{g}(y)\rangle - h(x,y))^2].$$

The lemma is a direct consequence of Jensen's inequality, and the fact that $\mathbb{E}_{\bar{\mathcal{D}}_{1\otimes 1}}[\langle \tilde{f}_\mu(\bar{x}), \mathbf{A}\tilde{g}_\mu(\bar{y})\rangle \mid x, y] = \langle \tilde{f}(x), \mathbf{A}\tilde{g}(y)\rangle$ for any $\mathbf{A} \in \mathbb{R}^{p\times p}$.

F.3.3. ANALYSIS UNDER AN EXACT BALANCED PROJECTION

We now analyze the performance of an idealized balanced projection of $(\tilde{f}, \tilde{g})$, and as a corollary, state a guarantee for deviations from this idealized projection. We accomplish this by analyzing the performance of the projections of noisy embeddings $(\tilde{f}_\mu, \tilde{g}_\mu)$ as a proxy, and then applying Lemma F.5 to return to the noise-free embeddings.

It is useful for us to formalize balancing as a general operation on matrices.

Definition F.4 (Balancing Operator) *Let $\mathbf{X}, \mathbf{Y} \in \mathbb{S}^p_>$. We define the balancing operator*

$$\Psi_{\mathrm{bal}}(\mathbf{Y}; \mathbf{X}) := \mathbf{X}^{\frac{1}{2}}(\mathbf{X}^{\frac{1}{2}}\mathbf{Y}\mathbf{X}^{\frac{1}{2}})^{-\frac{1}{2}}\mathbf{X}^{\frac{1}{2}} \in \mathbb{S}^p_>.$$

It is shown in Lemma L.1 that $\mathbf{W} = \Psi_{\mathrm{bal}}(\mathbf{Y}; \mathbf{X})$ is the unique positive definite operator satisfying $\mathbf{X} = \mathbf{W}\mathbf{Y}\mathbf{W}$. As a consequence, given $(\tilde{f}_\mu, \tilde{g}_\mu)$, the functions $(\tilde{f}_{\mu,\mathrm{bal}}, \tilde{g}_{\mu,\mathrm{bal}})$ defined as

$$\tilde{f}_{\mu,\mathrm{bal}} = \mathbf{W}_{\mathrm{bal},\mu}^{-\frac{1}{2}}\tilde{f}_\mu, \quad \tilde{g}_{\mu,\mathrm{bal}} = \mathbf{W}_{\mathrm{bal},\mu}^{\frac{1}{2}}\tilde{g}_\mu, \quad \mathbf{W}_{\mathrm{bal},\mu} := \Psi_{\mathrm{bal}}(\mathbf{\Sigma}_{\tilde{g},\mu}; \mathbf{\Sigma}_{\tilde{f},\mu})$$

satisfy (using $\mathbf{W}_{\mathrm{bal},\mu} = \mathbf{W}_{\mathrm{bal},\mu}^\top$)

$$\mathbb{E}_{\bar{\mathcal{D}}_{\mathcal{X},1}}[\tilde{f}_{\mu,\mathrm{bal}}(\tilde{f}_{\mu,\mathrm{bal}})^\top] = \underbrace{\mathbf{W}_{\mathrm{bal},\mu}^{-\frac{1}{2}}\mathbf{\Sigma}_{\tilde{f},\mu}\mathbf{W}_{\mathrm{bal},\mu}^{-\frac{1}{2}} = \mathbf{W}_{\mathrm{bal},\mu}^{\frac{1}{2}}\mathbf{\Sigma}_{\tilde{g},\mu}\mathbf{W}_{\mathrm{bal},\mu}^{\frac{1}{2}}}_{:=\mathbf{\Sigma}_{\mathrm{bal},\mu}} = \mathbb{E}_{\bar{\mathcal{D}}_{\mathcal{Y},1}}[\tilde{g}_{\mu,\mathrm{bal}}(\tilde{g}_{\mu,\mathrm{bal}})^\top], \tag{F.8}$$

as well as trivially $\langle \tilde{f}_{\mu,\mathrm{bal}}, \tilde{g}_{\mu,\mathrm{bal}} \rangle \equiv \langle \tilde{f}_\mu, \tilde{g}_\mu \rangle$. That is, the transformation

$$(\tilde{f}_\mu, \tilde{g}_\mu) \mapsto (\mathbf{W}_{\mathrm{bal},\mu}^{-\frac{1}{2}}\tilde{f}_\mu, \mathbf{W}_{\mathrm{bal},\mu}^{\frac{1}{2}}\tilde{g}_\mu)$$

balances $(\tilde{f}_\mu, \tilde{g}_\mu)$. We now introduce an operator expressing the covariance matrix of the balanced functions (in our case, $\mathbf{\Sigma}_{\mathrm{bal},\mu}$ above).

Definition F.5 (Balanced Covariance) *Given $\mathbf{X}, \mathbf{Y} \in \mathbb{S}^p_>$, we define*

$$\mathsf{CovBal}(\mathbf{X}, \mathbf{Y}) = \Psi_{\mathrm{bal}}(\mathbf{Y}; \mathbf{X})^{\frac{1}{2}} \cdot \mathbf{Y} \cdot \Psi_{\mathrm{bal}}(\mathbf{Y}; \mathbf{X})^{\frac{1}{2}}.$$

We remark that $\mathsf{CovBal}(\mathbf{X}, \mathbf{Y}) = \mathsf{CovBal}(\mathbf{Y}, \mathbf{X})$, as illustrated in Eq. (F.8). In particular,

$$\mathbf{\Sigma}_{\mathrm{bal},\mu} = \mathsf{CovBal}(\mathbf{\Sigma}_{\tilde{g},\mu}, \mathbf{\Sigma}_{\tilde{f},\mu}) = \mathsf{CovBal}(\mathbf{\Sigma}_{\tilde{f},\mu}, \mathbf{\Sigma}_{\tilde{g},\mu}).$$

We can now define our main object of interest: the operator which performs a singular value decomposition of the factorization $\tilde{f}_\mu, \tilde{g}_\mu$ in the coordinate system in which they are balanced.

Definition F.6 (Balancing Projection) *Given $\mathbf{X}, \mathbf{Y} \in \mathbb{S}^p_>$, for any $r \in [p]$, we define*

$$\mathsf{Proj}_{\mathrm{bal}}(r, \mathbf{X}, \mathbf{Y}) := \mathbf{W}^{-\frac{1}{2}}\mathbf{P}_r\mathbf{W}^{\frac{1}{2}},$$

where $\mathbf{W} = \Psi_{\mathrm{bal}}(\mathbf{Y}; \mathbf{X})$, and $\mathbf{P}_{\hat{r}}$ is the orthogonal projection onto the top-r eigenvectors of $\mathsf{CovBal}(\mathbf{Y}, \mathbf{X}) = \mathbf{W}^{\frac{1}{2}}\mathbf{Y}\mathbf{W}^{\frac{1}{2}}$. We say that $\mathbf{Q}_{\hat{r}} = \mathsf{Proj}_{\mathrm{bal}}(r, \mathbf{X}, \mathbf{Y})$ is unique *if the aforementioned projection $\mathbf{P}_{\hat{r}}$ is unique, that is, if $\sigma_r(\mathsf{CovBal}(\mathbf{Y}, \mathbf{X})) > \sigma_{r+1}(\mathsf{CovBal}(\mathbf{Y}, \mathbf{X}))$. Note that when $r = p$, this projection is trivially unique.*

In particular, suppose we consider $\mathbf{Q}_{\hat{r}} := \mathsf{Proj}_{\mathrm{bal}}(\hat{r}, \mathbf{\Sigma}_{\tilde{f},\mu}, \mathbf{\Sigma}_{\tilde{g},\mu})$. This performs a rank-$\hat{r}$ projection in the coordinates in which $\tilde{f}_\mu, \tilde{g}_\mu$ are balanced, and transforming $\langle \tilde{f}_\mu, \tilde{g}_\mu \rangle$ to $\langle \tilde{f}_\mu, \mathbf{Q}_{\hat{r}} \tilde{g}_\mu \rangle$ is equivalent to computing a rank-$\hat{r}$ SVD of the matrices. Thus, the error between $\langle \tilde{f}_\mu, \mathbf{Q}_{\hat{r}} \tilde{g}_\mu \rangle$ and $\langle f^\star_{\hat{r}}, g^\star_{\hat{r}} \rangle$ can be analyzed in terms of the error between the rank-$\hat{r}$ SVD approximations of two matrices which are close by. We use this insight to prove a perturbation bound, which we describe below.

The following lemma establishes three useful bounds: (a) an ℓ_2-deviation bound between the spectrum of $\mathbf{\Sigma}_{\mathrm{bal},\mu}$ and the spectrum of $\mathbf{\Sigma}^\star_{1\otimes 1}$; (b) a suboptimality guarantee for applying the *exact* balanced projection $\mathbf{Q}_{\hat{r}} = \mathsf{Proj}_{\mathrm{bal}}(r, \mathbf{\Sigma}_{\tilde{f},\mu}, \mathbf{\Sigma}_{\tilde{g},\mu})$ to $(\tilde{f}_\mu, \tilde{g}_\mu)$, where $(\tilde{f}_\mu, \tilde{g}_\mu)$ are the noise-convolved functions defined in the previous section; and (c) a perturbation inequality for applying an approximation $\mathbf{Q}'$ of $\mathbf{Q}_{\hat{r}}$ to $(\tilde{f}_\mu, \tilde{g}_\mu)$, and the subsequent guarantee when applying this projection to the original (non-noisy) functions $(\tilde{f}, \tilde{g})$.

Lemma F.6 (Accuracy of Balancing Projections) *Recall the definition of $\tilde{\epsilon}^2_\mu$ from Lemma F.4. Then,*

(a) It holds that $\sum_{i \ge 1} (\sigma_i(\mathbf{\Sigma}_{\mathrm{bal},\mu}) - \boldsymbol{\sigma}^\star_i)^2 \le \tilde{\epsilon}^2_\mu$.

(b) Given a given $\hat{r} \in \mathbb{N}$ for which $\boldsymbol{\sigma}^\star_{\hat{r}} > 0$, define $\boldsymbol{\delta}^\star_{\hat{r}} := 1 - \frac{\boldsymbol{\sigma}^\star_{\hat{r}+1}}{\boldsymbol{\sigma}^\star_{\hat{r}}}$. If $\tilde{\epsilon}_\mu \le \eta \boldsymbol{\sigma}^\star_{\hat{r}} \boldsymbol{\delta}^\star_{\hat{r}}$ for a given $\eta \in [0, 1)$, then

$$\mathcal{R}_{[\hat{r}]}(\tilde{f}_\mu, \mathbf{Q}_{\hat{r}} \cdot \tilde{g}_\mu; \bar{\mathcal{D}}_{1\otimes 1}) \le \frac{81 \tilde{\epsilon}^2_\mu}{(\boldsymbol{\delta}^\star_{\hat{r}}(1-\eta))^2},$$

where we define $\mathbf{Q}_{\hat{r}} = \mathsf{Proj}_{\mathrm{bal}}(\hat{r}, \mathbf{\Sigma}_{\tilde{f},\mu}, \mathbf{\Sigma}_{\tilde{g},\mu})$.

(c) Under the assumptions of (b), if $\hat{\mathbf{Q}} \in \mathbb{R}^{p \times p}$ is any other matrix, then, assuming $\mu \le B^2/p$,

$$\mathcal{R}_{[\hat{r}]}(\tilde{f}, \hat{\mathbf{Q}} \cdot \tilde{g}; \mathcal{D}_{1\otimes 1}) \le \mathcal{R}_{[\hat{r}]}(\tilde{f}_\mu, \hat{\mathbf{Q}} \cdot \tilde{g}_\mu; \bar{\mathcal{D}}_{1\otimes 1}) \le 8B^2 \|\hat{\mathbf{Q}} - \mathbf{Q}_{\hat{r}}\|_{\mathrm{op}} + \frac{162 \tilde{\epsilon}^2_\mu}{(\boldsymbol{\delta}^\star_{\hat{r}}(1-\eta))^2}.$$

Proof [Proof of Lemma F.6] The functions $\tilde{f}_{\mu,\mathrm{bal}}, \tilde{g}_{\mu,\mathrm{bal}}$ are balanced under $\bar{\mathcal{D}}_{1\otimes 1}$: $\mathbb{E}_{\bar{\mathcal{D}}_{\mathcal{X},1}}[\tilde{f}_{\mu,\mathrm{bal}}(\tilde{f}_{\mu,\mathrm{bal}})^\top] = \mathbb{E}_{\bar{\mathcal{D}}_{\mathcal{Y},1}}[\tilde{g}_{\mu,\mathrm{bal}}(\tilde{g}_{\mu,\mathrm{bal}})^\top] = \mathbf{\Sigma}_{\mathrm{bal},\mu}$. Moreover, by Lemma F.4,

$$\mathcal{R}(\tilde{f}_{\mu,\mathrm{bal}}, \tilde{g}_{\mu,\mathrm{bal}}; \bar{\mathcal{D}}_{1\otimes 1}) = \mathcal{R}(\tilde{f}_\mu, \tilde{g}_\mu; \bar{\mathcal{D}}_{1\otimes 1}) \le \tilde{\epsilon}^2_\mu.$$

Further, we have

$$\langle \tilde{f}_\mu, \mathbf{Q}_{\hat{r}} \tilde{g}_\mu \rangle = \langle \tilde{f}_{\mu,\mathrm{bal}}, \mathbf{P}_{\hat{r}} \tilde{g}_{\mu,\mathrm{bal}} \rangle,$$

where $\mathbf{P}_{\hat{r}}$ is the projection onto the top $\hat{r}$ eigenvectors of $\mathbf{\Sigma}_{\mathrm{bal},\mu}$. Hence, we can invoke[5] Theorem 11 to find both (a) $\sum_{i \ge 1} (\sigma_i(\mathbf{\Sigma}_{\mathrm{bal},\mu}) - \boldsymbol{\sigma}^\star_i)^2 \le \tilde{\epsilon}^2_\mu$ and (b) $\mathbb{E}_{\bar{\mathcal{D}}_{1\otimes 1}}[(\langle \tilde{f}_\mu, \mathbf{Q}_{\hat{r}} \cdot \tilde{g}_\mu \rangle - \langle f^\star_{\hat{r}}, g^\star_{\hat{r}} \rangle)^2] \le \frac{81 \tilde{\epsilon}^2_\mu}{(\boldsymbol{\delta}^\star_{\hat{r}}(1-\eta))^2}$. For part (c), the first inequality is a special case of Lemma F.5. Moreover,

$$\begin{aligned} \mathbb{E}_{\bar{\mathcal{D}}_{1\otimes 1}}[(\langle \tilde{f}_\mu, \hat{\mathbf{Q}} \cdot \tilde{g}_\mu \rangle - \langle f^\star_{\hat{r}}, g^\star_{\hat{r}} \rangle)^2] &= \mathbb{E}_{\bar{\mathcal{D}}_{1\otimes 1}}[(\langle \tilde{f}_\mu, (\hat{\mathbf{Q}} - \mathbf{Q}_{\hat{r}}) \cdot \tilde{g}_\mu \rangle + \langle \tilde{f}_\mu, \mathbf{Q}_{\hat{r}} \cdot \tilde{g}_\mu \rangle - \langle f^\star_{\hat{r}}, g^\star_{\hat{r}} \rangle)^2] \\ &\le 2\mathbb{E}_{\bar{\mathcal{D}}_{1\otimes 1}}[\langle \tilde{f}_\mu, (\hat{\mathbf{Q}} - \mathbf{Q}_{\hat{r}}) \cdot \tilde{g}_\mu \rangle^2] + 2\mathbb{E}_{\bar{\mathcal{D}}_{1\otimes 1}}(\langle \tilde{f}_\mu, \mathbf{Q}_{\hat{r}} \cdot \tilde{g}_\mu \rangle - \langle f^\star_{\hat{r}}, g^\star_{\hat{r}} \rangle)^2]. \end{aligned}$$

5. We note that while Theorem 11 is stated in terms of the non-augmented distribution $\mathcal{D}_{1\otimes 1}$, it holds for $\bar{\mathcal{D}}_{1\otimes 1}$ as well, as the augmented distribution preserves the covariance and balancing of the ground truth embeddings.

As the second term above is controlled by part (b) of the lemma, it remains to bound $\mathbb{E}_{\bar{\mathcal{D}}_{1\otimes 1}}[\langle \tilde{f}_\mu, (\hat{\mathbf{Q}} - \mathbf{Q}_{\hat{r}}) \cdot \tilde{g}_\mu \rangle^2]$. Using independence of $x, y, \check{\mathbf{x}}, \check{\mathbf{y}}$ under $\bar{\mathcal{D}}_{1\otimes 1}$, and and $\mathbb{E}[\check{\mathbf{x}}\check{\mathbf{x}}^\top] = \mathbb{E}[\check{\mathbf{y}}\check{\mathbf{y}}^\top] = \mathbf{I}_p$,

$$
\begin{aligned}
&\mathbb{E}_{\bar{\mathcal{D}}_{1\otimes 1}}[\langle \tilde{f}_\mu, (\hat{\mathbf{Q}} - \mathbf{Q}_{\hat{r}}) \cdot \tilde{g}_\mu \rangle^2] \\
&= \mathbb{E}_{\bar{\mathcal{D}}_{1\otimes 1}}[\langle \tilde{f}(x) + \sqrt{\mu}\check{\mathbf{x}}, (\hat{\mathbf{Q}} - \mathbf{Q}_{\hat{r}})(\tilde{g}(y) + \sqrt{\mu}\check{\mathbf{y}}) \rangle^2] \\
&= \mathrm{tr}(\mathbb{E}_{\bar{\mathcal{D}}_{1\otimes 1}}[(\tilde{f}(x) + \sqrt{\mu}\check{\mathbf{x}})(f(x) + \sqrt{\mu}\check{\mathbf{x}})^\top (\hat{\mathbf{Q}} - \mathbf{Q}_{\hat{r}})(g(x) + \sqrt{\mu}\check{\mathbf{y}})(\tilde{g}(y) + \sqrt{\mu}\check{\mathbf{y}})^\top (\hat{\mathbf{Q}} - \mathbf{Q}_{\hat{r}})^\top] \\
&= \mathrm{tr}((\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[\tilde{f}(x)\tilde{f}(x)^\top] + \mu \mathbf{I}_p)(\hat{\mathbf{Q}} - \mathbf{Q}_{\hat{r}})(\mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}[\tilde{g}(y)\tilde{g}(y)^\top] + \mu \mathbf{I}_p)(\hat{\mathbf{Q}} - \mathbf{Q}_{\hat{r}})^\top] \\
&= \mathrm{tr}(\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[\tilde{f}(x)\tilde{f}(x)^\top](\hat{\mathbf{Q}} - \mathbf{Q}_{\hat{r}})\mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}[\tilde{g}(y)\tilde{g}(y)^\top](\hat{\mathbf{Q}} - \mathbf{Q}_{\hat{r}})) \\
&\quad + \mu \mathrm{tr}((\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[\tilde{f}(x)\tilde{f}(x)^\top] + \mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}[\tilde{g}(y)\tilde{g}(y)^\top])(\hat{\mathbf{Q}} - \mathbf{Q}_{\hat{r}})(\hat{\mathbf{Q}} - \mathbf{Q}_{\hat{r}})^\top) \\
&\quad + \mu^2 \mathrm{tr}((\hat{\mathbf{Q}} - \mathbf{Q}_{\hat{r}})(\hat{\mathbf{Q}} - \mathbf{Q}_{\hat{r}})^\top).
\end{aligned}
$$

Using $\mathrm{tr}(\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[\tilde{f}(x)\tilde{f}(x)^\top]) \vee \mathrm{tr}(\mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}[\tilde{g}(y)\tilde{g}(y)^\top]) \le B^2$ due to $\tilde{f} \in \mathcal{F}_p, \tilde{g} \in \mathcal{G}_p$ and Assumption 2.7, (and using various standard trace inequalities), the above is atmost

$$
\begin{aligned}
&B^4 \|\hat{\mathbf{Q}} - \mathbf{Q}_{\hat{r}}\|_{\mathrm{op}}^2 + 2\mu^2 B^2 \|\hat{\mathbf{Q}} - \mathbf{Q}_{\hat{r}}\|_{\mathrm{op}}^2 + \mu^4 \mathrm{tr}((\hat{\mathbf{Q}} - \mathbf{Q}_{\hat{r}})(\hat{\mathbf{Q}} - \mathbf{Q}_{\hat{r}})^\top) \\
&\le (B^4 + 2\mu B^2 + \mu^2 p)\|\hat{\mathbf{Q}} - \mathbf{Q}_{\hat{r}}\|_{\mathrm{op}}^2 \le 4B^2 \|\hat{\mathbf{Q}} - \mathbf{Q}_{\hat{r}}\|_{\mathrm{op}}^2
\end{aligned}
$$

where the last inequality takes $\mu \le B^2/p$. ■

F.3.4. ANALYSIS OF EMPIRICAL BALANCING OPERATOR

Definition F.7 *Given $\mathbf{\Sigma} \in \mathbb{S}_{\ge}^p$, $r_0 \in [p]$, $\sigma > 0$, the* separated-rank *at (r_0, σ) (if it exists) is*

$$
\textsf{sep-rank}(r_0, \sigma; \mathbf{\Sigma}) := \max \left\{ r \in [r_0] : \sigma_r(\mathbf{\Sigma}) \ge \sigma, \sigma_r(\mathbf{\Sigma}) - \sigma_{r+1}(\mathbf{\Sigma}) \ge \frac{\sigma_r(\mathbf{\Sigma})}{r_0} \right\}. \qquad (F.9)
$$

We say the separated-rank is well-defined *if the above maximum exists.*

We next provide the result on the perturbation of the balancing projections, whose proof is deferred to Appendix L.5.

Proposition F.3 (Perturbation of Balancing Projections) *Let $r_0 \in \mathbb{N}$, matrices $\mathbf{X}, \mathbf{X}', \mathbf{Y}, \mathbf{Y}' \in \mathbb{S}_{>}^p$, and positive numbers $\sigma > 0$ and $(\bar{\sigma}_i)_{i \in [r_0+1]}$ satisfy the following conditions:*

(a) For any $\mathbf{A} \in \{\mathbf{X}, \mathbf{X}', \mathbf{Y}, \mathbf{Y}'\}$, $\mu \mathbf{I}_p \preceq \mathbf{A} \preceq M \mathbf{I}_p$.

(b) $\max\{\|\mathbf{X} - \mathbf{X}'\|_{\mathrm{op}}, \|\mathbf{Y} - \mathbf{Y}'\|_{\mathrm{op}}\} \le \Delta$, where $\Delta \le \frac{\mu}{32 r_0}(\mu/M)^2$.

(c) $\max_{i \in [r_0+1]} |\bar{\sigma}_i - \sigma_i(\mathbf{\Sigma})| \le \sigma/8r_0$, where $\mathbf{\Sigma} = \textsf{CovBal}(\mathbf{X}, \mathbf{Y})$.

(d) $\sigma \in [\max\{\mu, 2\bar{\sigma}_{\hat{r}_0}\}, \frac{2}{3e}\bar{\sigma}_1]$.

Define $\mathbf{\Sigma}' = \textsf{CovBal}(\mathbf{X}', \mathbf{Y}')$, $r = \textsf{sep-rank}(r_0, \sigma; \mathbf{\Sigma}')$, $\mathbf{Q} = \textsf{Proj}_{\mathrm{bal}}(r; \mathbf{X}, \mathbf{Y})$ and $\mathbf{Q}' = \textsf{Proj}_{\mathrm{bal}}(r; \mathbf{X}', \mathbf{Y}')$. Then, r is well defined, $\mathbf{Q}$ and $\mathbf{Q}'$ are unique, and the following bounds hold:

$$
\|\mathbf{Q}' - \mathbf{Q}\|_{\mathrm{op}} \le \frac{19 r_0 (M/\mu)^{5/2} \Delta}{\mu}, \quad \max\{\|\mathbf{Q}\|_{\mathrm{op}}, \|\mathbf{Q}'\|_{\mathrm{op}}\} \le \sqrt{M/\mu}.
$$

Moreover, $\bar{\sigma}_r \ge \frac{3}{4}\sigma$, $\bar{\sigma}_{r+1} \le 3\sigma$, and $\bar{\sigma}_r - \bar{\sigma}_{r+1} \ge \frac{\bar{\sigma}_{\hat{r}}}{3r_0}$.

F.3.5. CONCLUDING THE PROOF OF PROPOSITION F.2

Proof [Proof of Proposition F.2] Throughout suppose that the high probability events of Lemma F.3 and Lemma F.4 hold, with have a total failure probability of $2\delta/3$. We instantiate Proposition F.3 with

(1) $\mathbf{X} = \mathbf{\Sigma}_{\tilde{f},\mu}$, $\mathbf{X}' = \hat{\mathbf{\Sigma}}_{\tilde{f},\mu}$, $\mathbf{Y} = \mathbf{\Sigma}_{\tilde{g},\mu}$, and $\mathbf{Y}' = \hat{\mathbf{\Sigma}}_{\tilde{g},\mu}$.

(2) $r_0 \leftarrow r_{\mathrm{cut}}$, $\sigma \leftarrow \sigma_{\mathrm{cut}}$, $\hat{r} \leftarrow \mathsf{sep\text{-}rank}(r_{\mathrm{cut}}, \sigma_{\mathrm{cut}}; \hat{\mathbf{\Sigma}}_{\mathrm{bal},\mu})$, and

$$\mathbf{Q}' \leftarrow \hat{\mathbf{Q}}_{\hat{r}} := \mathsf{Proj}_{\mathrm{bal}}\left(r; \hat{\mathbf{\Sigma}}_{\tilde{f},\mu}, \hat{\mathbf{\Sigma}}_{\tilde{g},\mu}\right), \quad \mathbf{Q} \leftarrow \mathbf{Q}_{\hat{r}} := \mathsf{Proj}_{\mathrm{bal}}\left(r; \mathbf{\Sigma}_{\tilde{f},\mu}, \mathbf{\Sigma}_{\tilde{g},\mu}\right).$$

(3) $\mu \leftarrow B^2/n_1$ and $M \leftarrow 2B^2$. By assumption, $n_1 \geq p$, so $\mu = B^2/p$ satisfies the conditions of Lemma F.6.

(4) On the event of Lemma F.3, we have $\max\{\|\mathbf{X}-\mathbf{X}'\|_{\mathrm{op}}, \|\mathbf{Y}-\mathbf{Y}'\|_{\mathrm{op}}\} \leq \Delta$ for $\Delta = \epsilon_\Sigma(n_2) = B^2\sqrt{2\frac{\log(24p/\delta)}{n_2}}$. This holds with probability at least $1-\delta/3$.

(5) $\bar{\sigma}_i \leftarrow \boldsymbol{\sigma}_i^\star$, and $\sigma_i = \sigma_i(\mathbf{\Sigma}_{\mathrm{bal},\mu}) = \sigma_i(\mathsf{CovBal}(\mathbf{\Sigma}_{\tilde{f},\mu}, \mathbf{\Sigma}_{\tilde{g},\mu}))$.

We now check that the conditions (a)-(d) of Proposition F.3 are met.

(a) The PSD inequality holds by Lemma F.3 .

(b) $\Delta \leq \frac{\mu}{32r_0}(\mu/M)^2$ holds for $n_2 \geq r_{\mathrm{cut}}^2 2^{11}(n_1)^6 \log(24p/\delta)$, on the event of Lemma F.3.

(c) By Lemma F.6(a), it is enough that $\tilde{\epsilon}_\mu^2 \leq \frac{\sigma_{\mathrm{cut}}^2}{64r_{\mathrm{cut}}^2}$. On the event of Lemma F.4, it is enough that $\tilde{\epsilon}(p, n_1, \delta)^2 \leq \frac{\sigma_{\mathrm{cut}}^2}{64r_{\mathrm{cut}}^2}$.

(d) Substituting in $\mu = B^2/n_1$, $\sigma \leftarrow \sigma_{\mathrm{cut}}$ and $\bar{\sigma}_i \leftarrow \boldsymbol{\sigma}_i^\star$ the condition $\sigma \in [\max\{\mu, 2\bar{\sigma}_{\hat{r}_0}\}, \frac{2}{3e}\bar{\sigma}_1]$ holds for $n_1 \geq B^2/\sigma_{\mathrm{cut}}^2$ and $\sigma_{\mathrm{cut}} \in [2\boldsymbol{\sigma}_{r_{\mathrm{cut}}}^\star, \frac{2}{3e}\boldsymbol{\sigma}_1^\star]$.

Note that the suffcient conditions in (b)-(d) are all guaranteed by Proposition F.2. With the above substitutions, we achieve

(i) $\boldsymbol{\sigma}_{\hat{r}+1}^\star \leq 3\sigma_{\mathrm{cut}}$, $\boldsymbol{\sigma}_{\hat{r}}^\star \geq 3\sigma_{\mathrm{cut}}/4$, and $\boldsymbol{\delta}_{\hat{r}}^\star = \frac{\boldsymbol{\sigma}_{\hat{r}}^\star - \boldsymbol{\sigma}_{\hat{r}+1}^\star}{\boldsymbol{\sigma}_{\hat{r}}^\star} \geq \frac{1}{3r_{\mathrm{cut}}}$, and thus $\boldsymbol{\delta}_{\hat{r}}^\star \boldsymbol{\sigma}_{\hat{r}}^\star \geq \sigma_{\mathrm{cut}}/(4r_{\mathrm{cut}})$.

(ii) The upper bound on $\|\hat{\mathbf{Q}}_{\hat{r}} - \mathbf{Q}_{\hat{r}}\|_{\mathrm{op}}$ is given by

$$\|\hat{\mathbf{Q}}_{\hat{r}} - \mathbf{Q}_{\hat{r}}\|_{\mathrm{op}} \leq \frac{19r_{\mathrm{cut}}(M/\mu)^{5/2}\Delta}{\mu} = 19\sqrt{2\log(24p/\delta)}r_{\mathrm{cut}} \cdot \sqrt{n_1^7/n_2} \leq \frac{1}{n_1}$$

for $n_2 \geq 722r_{\mathrm{cut}}^2 n_1^9 \log(24p/\delta)$ (achieved under the proposition).

From Lemma F.6 with $\eta = 1/8$, we have that as long as $\tilde{\epsilon}_\mu \leq \sigma_{\mathrm{cut}}/(16r_{\mathrm{cut}}) \leq \boldsymbol{\sigma}_{\hat{r}}^\star \boldsymbol{\delta}_{\hat{r}}^\star/4$,

$$\begin{aligned}\mathcal{R}_{[\hat{r}]}(\tilde{f}, \hat{\mathbf{Q}}_{\hat{r}} \cdot \tilde{g}; \mathcal{D}_{1\otimes 1}) &\leq 4B^2\|\hat{\mathbf{Q}}_{\hat{r}} - \mathbf{Q}_{\hat{r}}\|_{\mathrm{op}} + \frac{324\tilde{\epsilon}_\mu^2}{(\boldsymbol{\delta}_{\hat{r}}^\star)^2} \\ &\leq \frac{4B^2}{n_1} + 2898r_{\mathrm{cut}}^2\tilde{\epsilon}_\mu^2.\end{aligned}$$

where the last line follows by invoking items (i) and (ii) above. On the event of Lemma F.4, we may upper bound $\tilde{\epsilon}_\mu^2$ by $\tilde{\epsilon}(p, n_1, \delta)^2$, as in Lemma F.2, giving

$$\mathcal{R}_{[\hat{r}]}(\tilde{f}, \hat{\mathbf{Q}}_{\hat{r}} \cdot \tilde{g}; \mathcal{D}_{1\otimes 1}) \le \frac{4B^2}{n_1} + 2898 r_{\mathrm{cut}}^2 \tilde{\epsilon}(p, n_1, \delta)^2 \le 3000 r_{\mathrm{cut}}^2 \tilde{\epsilon}(p, n_1, \delta)^2$$

We conclude by checking the two statements in the last line of Proposition F.2. To show the first, we note that, due to Proposition F.3, we find $\|\hat{\mathbf{Q}}_{\hat{r}}\|_{\mathrm{op}} \le \sqrt{M/\mu} = \sqrt{2n_1}$. Using Assumption 2.7 and the fact that $\tilde{f} \in \mathcal{F}_p$ and $\tilde{g} \in \mathcal{G}_p$ concludes that

$$|\langle \tilde{f}(x), \hat{\mathbf{Q}}_{\hat{r}} \tilde{g}(y) \rangle| \le B^2 \cdot \|\hat{\mathbf{Q}}_{\hat{r}}\|_{\mathrm{op}} \le \sqrt{2n_1} B^2.$$

To show the second, we note that, due to Proposition F.3, $\boldsymbol{\sigma}^\star_{\hat{r}+1} \le 3\sigma_{\mathrm{cut}}$, from which the inequalities $\mathbf{tail}_2^\star(\hat{r}) \le \mathbf{tail}_2^\star(r_{\mathrm{cut}}) + 9\sigma_{\mathrm{cut}}^2 r_{\mathrm{cut}}$ and $\mathbf{tail}_1^\star(\hat{r})^2 \le 18 r_{\mathrm{cut}}^2 \sigma_{\mathrm{cut}}^2 + 2\mathbf{tail}_1^\star(r_{\mathrm{cut}})^2$ are straightforward to verify. Together with Proposition F.3, these verify that the event $\mathcal{E}_{\mathrm{spec}}(r, \sigma_{\mathrm{cut}}, r_{\mathrm{cut}})$, defined in Definition F.1, holds. ■

F.4. Analysis of the second stage of double-stage ERM

The following lemma, which is established in Appendix G.4, handles the error on the second phase of double-stage ERM in terms of the first. Recall that we choose

$$(\hat{f}_{\mathrm{DS}}, \hat{f}_{\mathrm{DS}}) \in \operatorname*{arg\,min}_{(f,g) \in \mathcal{F}_{\hat{r}} \times \mathcal{G}_{\hat{r}}} \hat{L}_{(3)}(f, g) + \lambda \hat{L}_{(4)}(f, g)$$

$$\hat{L}_{(3)}(f, g) = \frac{1}{n_3} \sum_{i=1}^{n_3} (\langle f(x_{3,i}), g(y_{3,i}) \rangle - z_{3,i})^2$$

$$\hat{L}_{(4)}(f, g) = \frac{1}{n_4} \sum_{i=1}^{n_4} (\langle f(x_{4,i}), g(y_{4,i}) \rangle - \langle \tilde{f}(x_{4,i}), \hat{\mathbf{Q}}_{\hat{r}} \cdot \tilde{g}(y_{4,i}) \rangle)^2.$$

Lemma F.7 *Suppose it holds that $\|\hat{\mathbf{Q}}_{\hat{r}}\|_{\mathrm{op}} \le \sqrt{2n_1}$, as in the proof of Proposition F.2. Then, with probability at least $1 - \delta/3$ over the samples collected in Line 5 of Algorithm 1,*

$$\begin{aligned} &\mathcal{R}(\hat{f}_{\mathrm{DS}}, \hat{f}_{\mathrm{DS}}; \mathcal{D}_{\mathrm{train}}) + \frac{\lambda}{2} \mathcal{R}_{[r]}(\hat{f}_{\mathrm{DS}}, \hat{f}_{\mathrm{DS}}; \mathcal{D}_{1\otimes 1}) \\ &\quad \le 2\kappa_{\mathrm{apx}} \mathbf{tail}_2^\star(\hat{r}) + 3\lambda \mathcal{R}_{[r]}(\tilde{f}, \hat{\mathbf{Q}}_{\hat{r}} \cdot \tilde{g}; \mathcal{D}_{1\otimes 1}) + 352 \left(1 + \frac{\lambda n_1 n_3}{n_4}\right) \frac{B^4(\mathcal{M}_{\hat{r}} + \log(12/\delta))}{n_3}. \end{aligned}$$

We can now conclude the proof of our main theorem for double-stage ERM as follows.

Proof [Proof of Proposition F.1] First, we bound the regularized risk $\mathcal{R}(\hat{f}_{\mathrm{DS}}, \hat{f}_{\mathrm{DS}}; \mathcal{D}_{\mathrm{train}}) + \lambda \mathcal{R}_{[r]}(\hat{f}_{\mathrm{DS}}, \hat{f}_{\mathrm{DS}}; \mathcal{D}_{1\otimes 1})$. Using Proposition F.2 in Lemma F.7, we have

$$\begin{aligned} &\mathcal{R}(\hat{f}_{\mathrm{DS}}, \hat{f}_{\mathrm{DS}}; \mathcal{D}_{\mathrm{train}}) + \lambda \mathcal{R}_{[r]}(\hat{f}_{\mathrm{DS}}, \hat{f}_{\mathrm{DS}}; \mathcal{D}_{1\otimes 1}) \\ &\lesssim \kappa_{\mathrm{apx}}(\mathbf{tail}_2^\star(\hat{r}) + \lambda \kappa_{\mathrm{trn}} r_{\mathrm{cut}}^2 \mathbf{tail}_2^\star(p)) + \left(1 + \frac{\lambda \kappa_{\mathrm{trn}} n_1 n_3}{n_4}\right) \frac{B^4(\mathcal{M}_{\hat{r}} + \log(1/\delta))}{n_3} + \frac{\lambda \kappa_{\mathrm{trn}} r_{\mathrm{cut}}^2 B^4(\mathcal{M}_p + \log(1/\delta))}{n_1} \\ &\lesssim_\star \mathbf{tail}_2^\star(r_{\mathrm{cut}}) + r_{\mathrm{cut}} \sigma_{\mathrm{cut}}^2 + \lambda r_{\mathrm{cut}}^2 \mathbf{tail}_2^\star(p) + \left(1 + \frac{\lambda n_1 n_3}{n_4}\right) \frac{B^4(\mathcal{M}_{\hat{r}} + \log(1/\delta))}{n_3} + \frac{\lambda r_{\mathrm{cut}}^2 B^4(\mathcal{M}_p + \log(1/\delta))}{n_1} \\ &\le \mathbf{tail}_2^\star(\hat{r}) + r_{\mathrm{cut}} \sigma_{\mathrm{cut}}^2 + \lambda r_{\mathrm{cut}}^2 \mathbf{tail}_2^\star(p) + \left(1 + \frac{\lambda n_1 n_3}{n_4}\right) \frac{B^4(\mathcal{M}_{r_{\mathrm{cut}}} + \log(1/\delta))}{n_3} + \frac{\lambda r_{\mathrm{cut}}^2 B^4(\mathcal{M}_p + \log(1/\delta))}{n_1}, \end{aligned}$$

where in the second to last line, we use $\lesssim_\star$ to suprress polynomials in problem dependend constants, and in the last line, we use the assumptions that $K \mapsto \mathcal{M}_K$ is non-decreasing (see Assumption 2.7). For our choice of $n_4 \geq \lambda n_1 n_3$, the above simplifies further to

$$\mathcal{R}(\hat{f}_{\mathrm{DS}}, \hat{f}_{\mathrm{DS}}; \mathcal{D}_{\mathrm{train}}) + \lambda \mathcal{R}_{[r]}(\hat{f}_{\mathrm{DS}}, \hat{f}_{\mathrm{DS}}; \mathcal{D}_{1\otimes 1})$$
$$\lesssim_\star \mathsf{tail}_2^\star(r_{\mathrm{cut}}) + r_{\mathrm{cut}}\sigma_{\mathrm{cut}}^2 + \lambda r_{\mathrm{cut}}^2 \mathsf{tail}_2^\star(p) + \frac{B^4(\mathcal{M}_{r_{\mathrm{cut}}} + \log(1/\delta))}{n_3} + \frac{\lambda r_{\mathrm{cut}}^2 B^4(\mathcal{M}_p + \log(1/\delta))}{n_1}.$$

That the good spectral event $\mathcal{E}_{\mathrm{spec}}$ holds also follows from Proposition F.2. Lastly, we gather the necessary conditions in order for the conclusion of Proposition F.2 to hold, $\mu = B^2/n_1$, $n_1 \geq \max\{p, B^2/\sigma_{\mathrm{cut}}^2\}$ $n_2 \geq 722 r_{\mathrm{cut}}^2 n_1^9 \log(24p/\delta)$, and finally, we require Eq. (F.6). Stated succinctly, this last condition stipulates that for some constant $C \lesssim_\star 1$,

$$\mathsf{tail}_2^\star(p) + \frac{B^4(\mathcal{M}_p + \log\frac{1}{\delta})}{n_1} \leq \frac{\sigma_{\mathrm{cut}}^2}{C r_{\mathrm{cut}}^2}.$$

Doubling C by a factor 2, it is enough that $\mathsf{tail}_2^\star(p) \leq \frac{\sigma_{\mathrm{cut}}^2}{C r_{\mathrm{cut}}^2}$ and $n_1 \geq \frac{r_{\mathrm{cut}}^2 B^4(\mathcal{M}_p + \log\frac{1}{\delta})}{\sigma_{\mathrm{cut}}^2}$. The bound follows. ■

Appendix G. Learning Theory and Proofs in Appendix F

In this section, we review some fundamental while important results from learning theory, and related proofs in Appendix F.

G.1. Concentration inequalities

We begin with Bernstein's inequality (see e.g., (Boucheron et al., 2005, Chapter 2)).

Lemma G.1 (Bernstein Inequality) *Let $Z_1, \dots, Z_n \in \mathbb{R}$ be i.i.d. random variables with $|Z_i| \leq M$ and $\mathrm{Var}[Z_i] \leq \sigma^2$. Then, with probability at least $1-\delta$,*

$$\left|\frac{1}{n}\sum_{i=1}^n Z_i - \mathbb{E}[Z_i]\right| \leq \sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}} + \frac{M\log(1/\delta)}{3n}.$$

The following is a simplification of (Mackey et al., 2014, Corollary 4.2).

Lemma G.2 (Matrix Hoeffding) *Let $\mathbf{Y}_1, \dots, \mathbf{Y}_n \in \mathbb{R}^{d\times d}$ be i.i.d. symmetric matrices with $\mathbb{E}[\mathbf{Y}_i] = 0$ and $\|\mathbf{Y}\|_{\mathrm{op}}^2 \leq M$. Then, with probability at least $1-\delta$,*

$$\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{Y}_i\right\|_{\mathrm{op}} \leq M\sqrt{\frac{2\log(2d/\delta)}{n}}.$$

G.2. Learning with finite function classes

Lemma G.3 *Let Φ be a finite class of functions $\phi : \mathcal{W} \to \mathbb{R}^k$, and let $\phi_\star(w)$ be a nominal function, possibly not in Φ. Let $M > 0$ be a constant such that $\sup_{w\in\mathcal{W}} \max_{\phi\in\Phi} \|(\phi - \phi^\star)(w)\|_2 \le M$, and let $\mathcal{D}$ be a distribution over pairs $(w, \mathbf{z}) \in \mathcal{W} \times \mathbb{R}^k$ such that $\|\mathbf{z} - \phi^\star(w)\|_2 \le M$ and $\mathbb{E}[\mathbf{z} \mid w] = \phi^\star(w)$. Define $R(\phi) := \mathbb{E}_{w\sim\mathcal{D}}[\|\phi(w) - \phi^\star(w)\|^2]$, $\hat{L}_n(\phi) := \frac{1}{n}\sum_{i=1}^n \|\phi(w_i) - \phi^\star(w_i)\|^2$, and set $\hat{R}_n(\phi) = \hat{L}_n(\phi) - \hat{L}_n(\phi_\star)$. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$:*

- *The following guarantee holds simultaneously for all $\phi \in \Phi$ and all $\alpha > 0$:*

$$|R(\phi) - \hat{R}_n(\phi)| \le \frac{\alpha R(\phi)}{2} + \left(\frac{9}{\alpha} + 1\right) \cdot \frac{M^2 \log(2|\Phi|/\delta)}{n}.$$

- *All empirical risk minimizers $\hat{\phi} \in \arg\min_{\phi\in\Phi} \hat{L}_n(\phi) = \arg\min_{\phi\in\Phi} \hat{R}_n(\phi)$ satisfy*

$$R(\hat{\phi}) \le 2 \inf_{\phi'\in\Phi} \mathbb{E}_{\mathcal{D}}[(\phi'(w) - \phi^\star(w))^2] + \frac{78M^2 \log(2|\Phi|/\delta)}{n}.$$

Proof Throughout, all expectations are taken under spaces from $\mathcal{D}$. We expand

$$\hat{R}_n(\phi) = \frac{1}{n}\sum_{i=1}^n Z_i(\phi), \quad Z_i(\phi) := \|\phi(w_i) - \mathbf{z}_i\|^2 - \|\phi^\star(w_i) - \mathbf{z}_i\|^2.$$

By expanding

$$Z_i(\phi) := \|(\phi - \phi^\star)(w_i)\|^2 + 2\langle(\phi - \phi^\star)(w_i), (\phi^\star(w_i) - \mathbf{z}_i)\rangle,$$

we see that

$$\forall\phi, \quad \mathbb{E}[Z_i(\phi)] = R(\phi), \quad \text{w.p. } 1, \quad |Z_i(\phi)| \le 3M^2.$$

Furthermore, for all ϕ,

$$\begin{aligned}
\mathbb{E}[Z_i(\phi)^2] &= \mathbb{E}[(\|(\phi - \phi^\star)(w_i)\|^2 + 2\langle(\phi - \phi^\star)(w_i), (\phi^\star(w_i) - \mathbf{z}_i)\rangle)^2] \\
&\le \mathbb{E}[(\|(\phi - \phi^\star)(w_i)\|^2 + 2\|(\phi - \phi^\star)(w_i)\|\|\phi^\star(w_i) - \mathbf{z}_i\|)^2] \\
&\le \mathbb{E}[(3M\|(\phi - \phi^\star)(w_i)\|)^2] = 9M^2 R(\phi).
\end{aligned}$$

Thus, by Bernstein's inequality (Lemma G.1) and a union bound over all $\phi \in \Phi$, the following holds with probability at least $1 - \delta$:

$$\forall\phi \in \Phi, \quad |R(\phi) - \hat{R}_n(\phi)| \le \sqrt{\frac{18M^2 R(\phi) \log(2|\Phi|/\delta)}{n}} + \frac{M^2 \log(2|\Phi|/\delta)}{n}.$$

Therefore, by AM-GM inequality, the following holds for all fixed $\alpha > 0$:

$$\forall\phi \in \Phi, \quad |R(\phi) - \hat{R}_n(\phi) - \hat{R}_n(\phi_\star)| \le \frac{\alpha R(\phi)}{2} + \left(\frac{9}{\alpha} + 1\right) \cdot \frac{M^2 \log(2|\Phi|/\delta)}{n}.$$

This establishes the first statement of the lemma.

To prove the second statement, let $\tilde{\phi} \in \arg\min_{\phi \in \Phi} R(\phi)$. Then, we have that on the event of the previous display,

$$\begin{aligned} R(\hat{\phi}) - R(\tilde{\phi}) &= R(\hat{\phi}) - R_n(\hat{\phi}) + \underbrace{\hat{R}_n(\hat{\phi}) - \hat{R}_n(\tilde{\phi})}_{\le 0} + \hat{R}_n(\tilde{\phi}) - R(\tilde{\phi}) \\ &\le \frac{\alpha}{2}(R(\hat{\phi}) + R(\tilde{\phi})) + 2\left(\frac{9}{\alpha} + 1\right) \cdot \frac{M^2 \log(2|\Phi|/\delta)}{n} \\ &\le \alpha R(\hat{\phi}) + 2\left(\frac{9}{\alpha} + 1\right) \cdot \frac{M^2 \log(2|\Phi|/\delta)}{n}. \end{aligned}$$

Selecting $\alpha = 1/2$ and rearranging

$$\frac{1}{2} R(\hat{\phi}) \le R(\tilde{\phi}) + 2(18+1) \cdot \frac{M^2 \log(2|\Phi|/\delta)}{n}.$$

The bound follows. ■

G.3. Proof of Lemma F.1

The first inequality is a direct consequence of Lemma G.3. Here, we take the function class $\Phi = \mathcal{F}_p \times \mathcal{G}_p$, so $\log|\Phi| = \mathcal{M}_p$. Moreover, by Assumption 2.7, we can take

$$M = \sup_{\mathcal{F}_p \in \mathcal{F}, g \in \mathcal{G}_p} \sup_{x,y} (\langle f(x), g(y) \rangle - \langle f^\star(x), g^\star(y) \rangle) \le 2B^2.$$

The second inequality uses Assumption 2.5 to bound $\mathcal{R}(f_p^\star, g_p^\star; \mathcal{D}_{\mathrm{train}}) \le \kappa_{\mathrm{apx}} \mathcal{R}(f_p^\star, g_p^\star; \mathcal{D}_{1\otimes 1})$, and noting the fact that $\mathcal{R}(f_p^\star, g_p^\star; \mathcal{D}_{1\otimes 1}) = \mathbf{tail}_2^\star(p)$ by Lemma M.4. The third inequality uses Assumption 2.2, incurring an addition factor of κ_{trn}. ■

G.4. Proof of Lemma F.7

Let $\Phi := \{(x,y) \mapsto \langle f(x), g(y) \rangle, (f,g) \in \mathcal{F}_r \times \mathcal{G}_r\}$. Further, define

$$\phi_{3,\star} := \langle f^\star(x), g^\star(y) \rangle, \quad \phi_{4,\star} := \langle \tilde{f}(x), \hat{\mathbf{Q}}_r \tilde{g}(y) \rangle.$$

We define $\mathcal{D}_3$ as the distribution of $(x,y,z) \sim \mathcal{D}_{\mathrm{train}}$, and $\mathcal{D}_4$ as the distribution of (x', y', z'), where $(x', y') \sim \mathcal{D}_{1\otimes 1}$ and $z' = \phi_{4,\star}(x', y')$. We compute that, using Assumptions 2.4 and 2.7, and the last statement of Proposition F.2,

$$\begin{aligned} &\sup_{x,y} \max_{\phi} \|\phi(x,y) - \phi_{3,\star}(x,y)\| \le 2B^2 \\ &\sup_{x,y} \max_{\phi} \|\phi(x,y) - \phi_{4,\star}(x,y)\| \le (1 + \sqrt{2n_1})B^2, \end{aligned}$$

and

$$\log|\Phi| = \log|\mathcal{F}_r||\mathcal{G}_r| = \mathcal{M}_r.$$

For $i \in \{3,4\}$, let R_i and $\hat{L}_{i,n_i}, \hat{R}_{i,n_i}$ denote the corresponding excess risks as in Lemma G.3, the following holds with probability at least $1-\delta/3$ for all $\phi \in \Phi$

$$|R_3(\phi) - \hat{R}_{3,n_3}(\phi)| \le \frac{1}{4}R_3(\phi) + (19\cdot 4)\frac{B^4(\mathcal{M}_r + \log(12/\delta))}{n_3}$$
$$|R_4(\phi) - \hat{R}_{4,n_4}(\phi)| \le \frac{1}{4}R_4(\phi) + (19\cdot(2+2n_1))\frac{B^4(\mathcal{M}_r + \log(12/\delta))}{n_4},$$

where we set $\alpha = 1/2$ in the first statement of Lemma G.3. Set $R_\lambda(\phi) = R_3(\phi) + \lambda R_4(\phi)$. Then if $\hat{\phi} \in \arg\min_{\phi\in\Phi} \hat{L}_{3,n_3}(\phi) + \lambda \hat{L}_{4,n_4}(\phi) = \arg\min_{\phi\in\Phi} \hat{R}_{3,n_3}(\phi) + \lambda\hat{R}_{4,n_4}(\phi)$, we see that for any other $\tilde{\phi} \in \arg\min_{\phi\in\Phi} R_\lambda(\phi)$,

$$\begin{aligned}
R_\lambda(\hat{\phi}) - R_\lambda(\tilde{\phi}) &\le \frac{1}{4}(R_\lambda(\hat{\phi}) + R_\lambda(\tilde{\phi})) + 2(19\cdot 4)\frac{B^4(\mathcal{M}_r + \log(12/\delta))}{n_3} + 2\lambda(19\cdot(2+2n_1))\frac{B^4(\mathcal{M}_r + \log(12/\delta))}{n_4} \\
&\le \frac{1}{2}R_\lambda(\hat{\phi}) + 176\left(1 + \frac{\lambda n_1 n_3}{n_4}\right)\frac{B^4(\mathcal{M}_r + \log(12/\delta))}{n_3}.
\end{aligned}$$

Rearranging,

$$R_\lambda(\hat{\phi}) \le 2R_\lambda(\tilde{\phi}) + 352\left(1 + \frac{\lambda n_1 n_3}{n_4}\right)\frac{B^4(\mathcal{M}_r + \log(12/\delta))}{n_3}.$$

To conclude, we handle the terms $R_\lambda(\hat{\phi})$ and $R_\lambda(\tilde{\phi})$. First,

$$\begin{aligned}
R_\lambda(\tilde{\phi}) &= \inf_{\phi\in\Phi} R_\lambda(\phi) \\
&= \inf_{(f,g)\in\mathcal{F}_r\times\mathcal{G}_r} \mathcal{R}(f,g;\mathcal{D}_{\mathrm{train}}) + \lambda\mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle f,g\rangle - \langle \tilde{f}, \hat{\mathbf{Q}}_r\cdot\tilde{g}\rangle)^2] \\
&\le \mathcal{R}(f_r^\star, g_r^\star;\mathcal{D}_{\mathrm{train}}) + \lambda\mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle f_r^\star, g_r^\star\rangle - \langle \tilde{f}, \hat{\mathbf{Q}}_r\cdot\tilde{g}\rangle)^2] && ((f_r^\star, g_r^\star)\in\mathcal{F}_r\times\mathcal{G}_r) \\
&\le \kappa_{\mathrm{apx}}\mathcal{R}(f_r^\star, g_r^\star;\mathcal{D}_{1\otimes 1}) + \lambda\mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle f_r^\star, g_r^\star\rangle - \langle \tilde{f}, \hat{\mathbf{Q}}_r\cdot\tilde{g}\rangle)^2] && (\text{Assumption 2.5}) \\
&= \kappa_{\mathrm{apx}}\textsf{tail}_2^\star(r) + \lambda\mathcal{R}_{[r]}(\tilde{f}, \hat{\mathbf{Q}}_r\cdot\tilde{g};\mathcal{D}_{1\otimes 1}).
\end{aligned}$$

Second,

$$\begin{aligned}
&\mathcal{R}(\hat{f},\hat{g};\mathcal{D}_{\mathrm{train}}) + \frac{\lambda}{2}\mathcal{R}_{[r]}(\hat{f},\hat{g};\mathcal{D}_{1\otimes 1}) \\
&= \mathcal{R}(\hat{f},\hat{g};\mathcal{D}_{\mathrm{train}}) + \frac{\lambda}{2}\mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle \hat{f},\hat{g}\rangle - \langle f_r^\star, g_r^\star\rangle)^2] \\
&\le \underbrace{\mathcal{R}(\hat{f},\hat{g};\mathcal{D}_{\mathrm{train}})}_{=R_3(\hat{\phi})} + \lambda\underbrace{\mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle \tilde{f}, \hat{\mathbf{Q}}_r\cdot\tilde{g}\rangle - \langle \hat{f},\hat{g}\rangle)^2]}_{R_4(\hat{\phi})} + \lambda\underbrace{\mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle \tilde{f}, \hat{\mathbf{Q}}_r\cdot\tilde{g}\rangle - \langle f_r^\star, g_r^\star\rangle)^2]}_{=\mathcal{R}_{[r]}(\tilde{f},\hat{\mathbf{Q}}_r\cdot\tilde{g};\mathcal{D}_{1\otimes 1})} \\
&= R_\lambda(\hat{\phi}) + \lambda\mathcal{R}_{[r]}(\tilde{f}, \hat{\mathbf{Q}}_r\cdot\tilde{g};\mathcal{D}_{1\otimes 1}).
\end{aligned}$$

In sum, we conclude

$$\begin{aligned}
&\mathcal{R}(\hat{f},\hat{g};\mathcal{D}_{\mathrm{train}}) + \frac{\lambda}{2}\mathcal{R}_{[r]}(\hat{f},\hat{g};\mathcal{D}_{1\otimes 1}) \\
&\quad \le 2\kappa_{\mathrm{apx}}\textsf{tail}_2^\star(r) + 3\lambda\mathcal{R}_{[r]}(\tilde{f}, \hat{\mathbf{Q}}_r\cdot\tilde{g};\mathcal{D}_{1\otimes 1}) + 352\left(1 + \frac{\lambda n_1 n_3}{n_4}\right)\frac{B^4(\mathcal{M}_r + \log(12/\delta))}{n_3},
\end{aligned}$$

which completes the proof. ■

Appendix H. Proof of Rate Instantiations

This section gives the proofs of Lemmas E.1 and E.2, the instantiations of our error bounds under the spectral decay assumptions stipulated in Assumption 2.6. We begin by establishing the following two spectral decay bounds.

Lemma H.1 ($\mathsf{tail}_q^\star$ Bounds) *Suppose Assumption 2.6 holds. Then,*

$$\mathsf{tail}_1^\star(r) \le \begin{cases} C(1+\gamma^{-1})(r+1)^{-\gamma} & \textit{(polynomial decay)} \\ C(1+\gamma^{-1})e^{-\gamma(r+1)} & \textit{(exponential decay)} \end{cases}$$

$$\mathsf{tail}_2^\star(r) \le \begin{cases} 2C^2(r+1)^{-1-2\gamma} & \textit{(polynomial decay)} \\ \mathsf{tail}_2^\star(r) \le C^2(1+\gamma^{-1})e^{-2\gamma(r+1)} & \textit{(exponential decay)} \end{cases}$$

Lemma H.2 *Suppose Assumption 2.6 holds, and $\boldsymbol{\sigma}_r^\star > 0$. Then,*

$$\frac{\mathsf{tail}_2^\star(r)^2}{(\boldsymbol{\sigma}_r^\star)^2} \le \begin{cases} 3C^2 r^{-2\gamma} & \textit{(polynomial decay)} \\ C^2(1+\gamma^{-1}+r)^2 e^{-2\gamma r} & \textit{(exponential decay)} \end{cases}.$$

The above lemmas are proved in Lemma H.1 and Appendix H.4 respectively. We give the proof of Lemmas E.1 and E.2 in the following two sections.

H.1. Proof of Lemma E.1

In both decay regimes, we apply Lemmas H.1 and H.2. Under the polynomial decay, we have

$$\begin{aligned} \textsc{ApxErr}_{\mathrm{ss}}(r) &:= r^4 \cdot \mathsf{tail}_2^\star(r) + \mathsf{tail}_1^\star(r)^2 + r^2(\boldsymbol{\sigma}_{r+1}^\star)^2 + \frac{r^6 \cdot \mathsf{tail}_2^\star(r)^2}{(\sigma_r^\star)^2} \\ &\le 2C^2 r^4 r^{-1-2\gamma} + C^2(1+\gamma^{-1})^2 r^{-2\gamma} + C^2 r^{2-2(1+\gamma)} + 3C^2 r^{6-2\gamma} \\ &\lesssim C^2(1+\gamma^{-1})^2 r^{6-2\gamma}. \end{aligned}$$

In the exponential case, a similar argument applies.

H.2. Proof of Lemma E.2

Again, let $\psi(r)$ be equal to $\psi(r) = Cr^{-(1+\gamma)}$ for polynomial decay, and $\psi(r) = Ce^{-r\gamma}$ for exponential decay; thus, under Assumption 2.6, $\psi(r) \ge \boldsymbol{\sigma}_r^\star$.

Claim H.1 *Suppose we take $\sigma_{\mathrm{cut}} \ge 2\psi(r_{\mathrm{cut}})$. Then, if $\psi(r_{\mathrm{cut}}) \le \frac{1}{3e}\boldsymbol{\sigma}^\star$, Condition E.1*(c) *holds.*

Proof [Proof of Claim H.1] Observe that, if we select $\sigma_{\mathrm{cut}} = 2\psi(r_{\mathrm{cut}})$, then if $\psi(r_{\mathrm{cut}}) \le \frac{1}{3e}\boldsymbol{\sigma}^\star$, thn

$$2\boldsymbol{\sigma}_{r_{\mathrm{cut}}}^\star \le 2\psi(r_{\mathrm{cut}}) = \sigma_{\mathrm{cut}}, \quad \sigma_{\mathrm{cut}} = 2\psi(r_{\mathrm{cut}}) \le \frac{2}{3e}\boldsymbol{\sigma}^\star.$$

Therefore, Condition E.1(c) holds. ■

Polynomial Decay. From Lemma H.1, we have $\mathbf{tail}_1^\star(r) \le C(1+\gamma^{-1})(r_{\mathrm{cut}}+1)^{-\gamma}$ and $\mathbf{tail}_2^\star(r) \le 2C^2(r_{\mathrm{cut}}+1)^{-(1+2\gamma)}$, and by definition of ψ, $\sigma_{\mathrm{cut}} \ge 2\psi(r_{\mathrm{cut}}) = 2Cr_{\mathrm{cut}}^{-(1+\gamma)}$. We then bound

$$\begin{aligned}\mathrm{ERR}_{\mathrm{DT}}(r_{\mathrm{cut}},\sigma_{\mathrm{cut}}) &:= r_{\mathrm{cut}}^2\sigma_{\mathrm{cut}}^2 + \mathbf{tail}_1^\star(r_{\mathrm{cut}})^2 + \frac{\mathbf{tail}_2^\star(r_{\mathrm{cut}})^2}{(\sigma_{\mathrm{cut}})^2}\\ &\le r_{\mathrm{cut}}^2\sigma_{\mathrm{cut}}^2 + C^2(1+\gamma^{-1})^2 r_{\mathrm{cut}}^{-2\gamma} + \frac{4C^2 r_{\mathrm{cut}}^{-2(1+2\gamma)}}{4C^2 r_{\mathrm{cut}}^{-2(1+\gamma)}}\\ &\le r_{\mathrm{cut}}^2\sigma_{\mathrm{cut}}^2 + C^2(1+(1+\gamma^{-1})^2)r_{\mathrm{cut}}^{-2\gamma} \lesssim r_{\mathrm{cut}}^2\sigma_{\mathrm{cut}}^2 + C^2(1+\gamma^{-2})r_{\mathrm{cut}}^{-2\gamma}.\end{aligned}$$

Let us check each of the conditions of Condition E.1.

Claim H.2 *Suppose that* $r_{\mathrm{cut}} \ge \max\{c_1, \frac{3eC}{\boldsymbol{\sigma}_1^\star}\}$ *and* $p \ge c_1^{-\frac{1}{1+2\gamma}} r_{\mathrm{cut}}^{\frac{7+5\gamma}{1+2\gamma}}$. *Then, Condition E.1 holds, and the interval* $[2Cr_{\mathrm{cut}}^{-(1+\gamma)}, \frac{2}{3e}\boldsymbol{\sigma}^\star]$ *is nonempty.*

Proof [Proof of Claim H.2] For Condition E.1(a), we need $r_{\mathrm{cut}} \ge c_1$, and $\mathbf{tail}_2^\star(r_{\mathrm{cut}}) \le \frac{1}{c_1} r_{\mathrm{cut}}^2\sigma_{\mathrm{cut}}^2 = \frac{4}{c_1} r_{\mathrm{cut}}^2(\sigma_{\mathrm{cut}}^2)$. It suffices that $2C^2(r_{\mathrm{cut}}+1)^{-(1+2\gamma)} \le \frac{4C^2}{c_1} r_{\mathrm{cut}}^{2-2(1+\gamma)} = \frac{4C^2}{c_1} r_{\mathrm{cut}}^{-2\gamma}$. As $r_{\mathrm{cut}}+1 \ge r_{\mathrm{cut}}$, it is enough that $1 \le (2/c_1) r_{\mathrm{cut}}^{-2\gamma+1+2\gamma} = (2/c_t) r_{\mathrm{cut}}$, which holds for $r_{\mathrm{cut}} \ge c_1$. For Condition E.1(b), we need $\mathbf{tail}_2^\star(p) \le \frac{1}{c_1}\frac{\sigma_{\mathrm{cut}}^2}{r_{\mathrm{cut}}^5}$. We have $\mathbf{tail}_2^\star(p) \le 2C^2 p^{-(1+2\gamma)}$, and $\frac{1}{c_1}\cdot\frac{\sigma_{\mathrm{cut}}^2}{r_{\mathrm{cut}}^5} = \frac{1}{c_1}\cdot 4C^2 r_{\mathrm{cut}}^{-2(1+\gamma)-5} = \frac{1}{c_1}\cdot 4C^2 r_{\mathrm{cut}}^{-(7+5\gamma)}$. Hence, it is enough that $p^{-(1+2\gamma)} \le r_{\mathrm{cut}}^{-(7+5\gamma)}$, i.e. $p \ge c_1^{-\frac{1}{1+2\gamma}} r_{\mathrm{cut}}^{\frac{7+5\gamma}{1+2\gamma}}$. For Condition E.1(c), Claim H.1 requires the choice of $\psi(r_{\mathrm{cut}}) \le \frac{1}{3e}\boldsymbol{\sigma}_1^\star$, i.e. $Cr_{\mathrm{cut}}^{-(1+\gamma)} \le \frac{1}{3e}\boldsymbol{\sigma}_1^\star$. For this, it is enough that $r_{\mathrm{cut}} \ge \frac{3eC}{\boldsymbol{\sigma}_1^\star}$. ■

Exponential Decay. From Lemma H.1, we have $\mathbf{tail}_1^\star(r) \le C(1+\gamma^{-1})e^{-\gamma r}$, $\mathbf{tail}_2^\star(r) \le C^2(1+\gamma^{-1})e^{-2\gamma r}$, and by definition of ψ, $\sigma_{\mathrm{cut}} \ge 2\psi(r_{\mathrm{cut}}) = 2Ce^{-\gamma r}$. Then,

$$\begin{aligned}\mathrm{ERR}_{\mathrm{DT}}(r_{\mathrm{cut}},\sigma_{\mathrm{cut}}) &:= r_{\mathrm{cut}}^2\sigma_{\mathrm{cut}}^2 + \mathbf{tail}_1^\star(r_{\mathrm{cut}})^2 + \frac{\mathbf{tail}_2^\star(r_{\mathrm{cut}})^2}{(\sigma_{\mathrm{cut}})^2}\\ &\le \sigma_{\mathrm{cut}}^2 r_{\mathrm{cut}}^2 + C^2(1+\gamma^{-1})^2 e^{-2\gamma r_{\mathrm{cut}}} + \frac{C^4(1+\gamma^{-1})^2 e^{-4\gamma r_{\mathrm{cut}}}}{4C^2 e^{-2\gamma r_{\mathrm{cut}}}}\\ &\le \sigma_{\mathrm{cut}}^2 r_{\mathrm{cut}}^2 + C^2(1+\gamma^{-1})^2 e^{-2\gamma r_{\mathrm{cut}}} + \frac{C^4(1+\gamma^{-1})^2 e^{-4\gamma r_{\mathrm{cut}}}}{4C^2 e^{-2\gamma r_{\mathrm{cut}}}}\\ &\lesssim \sigma_{\mathrm{cut}}^2 r_{\mathrm{cut}}^2 + C^2(1+\gamma^{-2})e^{-2\gamma r_{\mathrm{cut}}}\end{aligned}$$

We conclude by checking Condition E.1

Claim H.3 *Suppose that* $r_{\mathrm{cut}} \ge \max\{c_1, \sqrt{c_1(1+\gamma^{-1})}, \frac{1}{\gamma}\log(\frac{3eC}{\boldsymbol{\sigma}_1^\star})\}$ *and* $p \ge \max\{2r_{\mathrm{cut}}, \frac{1}{\gamma}\log(r_{\mathrm{cut}}^5 c_1)\}$. *Then, Condition E.1 holds, and the interval* $[2Ce^{-\gamma r_{\mathrm{cut}}}, \frac{2}{3e}\boldsymbol{\sigma}^\star]$ *is nonempty.*

Proof [Proof of Claim H.3] For Condition E.1(a), we need $r_{\mathrm{cut}} \ge c_1$, and $\mathbf{tail}_2^\star(r_{\mathrm{cut}}) \le \frac{1}{c_1} r_{\mathrm{cut}}^2\sigma_{\mathrm{cut}}^2 = \frac{4}{c_1} r_{\mathrm{cut}}^2(\sigma_{\mathrm{cut}}^2)$. It suffices that $C^2(1+\gamma^{-1})e^{-2\gamma r_{\mathrm{cut}}} \le \frac{4C^2}{c_1} r_{\mathrm{cut}}^2 e^{-2\gamma r_{\mathrm{cut}}}$. For this, it suffices that $r_{\mathrm{cut}} \ge \sqrt{c_1(1+\gamma^{-1}}$.

For Condition E.1 (b), we need $\mathsf{tail}_2^\star(p) \le \frac{1}{c_1}\frac{\sigma_{\mathrm{cut}}^2}{r_{\mathrm{cut}}^5}$. We have $\mathsf{tail}_2^\star(p) \le (1+\gamma^{-1})C^2 e^{-2\gamma p}$, and $\frac{1}{c_1}\cdot\frac{\sigma_{\mathrm{cut}}^2}{r_{\mathrm{cut}}^5} = \frac{1}{r_{\mathrm{cut}}^5 c_1}\cdot 4C^2 e^{-2\gamma r_{\mathrm{cut}}}$. Hence, it is enough that $e^{-2\gamma(p-r_{\mathrm{cut}})} \le \frac{4}{r_{\mathrm{cut}}^5 c_1}$. For $p \ge 2r_{\mathrm{cut}}$, it is enough that $e^{-\gamma p} \le \frac{4}{r_{\mathrm{cut}}^5 c_1}$. Thus, it suffices that $p \ge \max\{2r_{\mathrm{cut}}, \frac{1}{\gamma}\log(r_{\mathrm{cut}}^5 c_1)\}$. For Condition E.1 (c), Claim H.1 requires the choice of $\psi(r_{\mathrm{cut}}) \le \frac{1}{3e}\boldsymbol{\sigma}_1^\star$, i.e. $Ce^{-\gamma r} \le \frac{1}{3e}\boldsymbol{\sigma}_1^\star$. For this, it is enough that $r_{\mathrm{cut}} \ge \frac{1}{\gamma}\log(\frac{3eC}{\boldsymbol{\sigma}_1^\star})$. ■

This concludes the proof. ■

H.3. Proof of Lemma H.1

We begin with the polynomial decay case, where $\boldsymbol{\sigma}_r^\star \le Cr^{-(1+\gamma)}$. We compute

$$\begin{aligned}\mathsf{tail}_1^\star(r) &= C\sum_{n>r}^{\infty} n^{-(1+\gamma)}\\ &\le C(r+1)^{-(1+\gamma)} + \int_{x=r+1}^{\infty} x^{-(1+\gamma)}\mathrm{d}x\\ &\le C(1+\gamma^{-1})(r+1)^{-\gamma}.\end{aligned}$$

and

$$\begin{aligned}\mathsf{tail}_2^\star(r) &= C^2\sum_{n>r}^{\infty} n^{-2(1+\gamma)}\\ &\le C^2(r+1)^{-2(1+\gamma)} + C^2\int_{x=r+1}^{\infty} x^{-2(1+\gamma)}\mathrm{d}x\\ &\le C^2(1+\frac{1}{1+2\gamma})(r+1)^{-1-2\gamma} \le 2C^2(r+1)^{-1-2\gamma}.\end{aligned}$$

We now turn to the exponential decay case, where $\boldsymbol{\sigma}_r^\star \le C\exp(-\gamma r)$. We have

$$\begin{aligned}\mathsf{tail}_1^\star(r) &= C\sum_{n>r}^{\infty} e^{-\gamma n}\\ &\le Ce^{-\gamma(r+1)} + C\int_{x=r+1}^{\infty} e^{-\gamma x}\mathrm{d}x\\ &\le C(1+\gamma^{-1})e^{-\gamma(r+1)}.\end{aligned}$$

and

$$\begin{aligned}\mathsf{tail}_2^\star(r) &= C^2\sum_{n>r}^{\infty} e^{-2\gamma n}\\ &\le (Ce^{-\gamma(r+1)})^2 + C\int_{x=r+1}^{\infty} e^{-2\gamma x}\mathrm{d}x\\ &\le C^2(1+\frac{1}{2}\gamma^{-1})(e^{-\gamma(r+1)})^2\\ &\le C^2(1+\gamma^{-1})(e^{-\gamma(r+1)})^2.\end{aligned}$$

■

H.4. Proof of Lemma H.2

Let $\psi(r) := Cr^{-(1+\gamma)}$ under polynomial decay, and $\psi(r) = Ce^{-\gamma r}$ under exponential decay. We start with a useful claim, and then turn to the polynomial and exponential decay regimes in sequence. Going forward, set $\Delta = \boldsymbol{\sigma}_r^\star$, and let $\bar{r} := \inf\{i \in \mathbb{N} : \psi(i) \le \Delta\}$.

Claim H.4 $\mathsf{tail}_2^\star(r) \le \bar{r}\Delta^2 + \mathsf{tail}_2^\star(\bar{r})$ *and* $\mathsf{tail}_1^\star(r) \le (\bar{r}-1)\Delta + \mathsf{tail}_1^\star(\bar{r})$.

Proof [Proof of Claim H.4] $\psi(r) \ge \boldsymbol{\sigma}_r^\star = \Delta$ implies $r \le \bar{r}$.

$$
\begin{aligned}
\mathsf{tail}_2^\star(r) &= \sum_{n>r} (\sigma_r^\star)^2 = \sum_{n=r+1}^{\bar{r}} (\sigma_r^\star)^2 + \sum_{n>\bar{r}} (\sigma_r^\star)^2 \\
&\le (\bar{r}-1)(\sigma_{r+1}^\star)^2 + \sum_{n>\bar{r}+1} (\sigma_r^\star)^2 = \bar{r}(\sigma_r^\star)^2 + \mathsf{tail}_2^\star(\bar{r}) \\
&\le (\bar{r}-1)\Delta^2 + \mathsf{tail}_2^\star(\bar{r}),
\end{aligned}
$$

where we use that $\sigma_{r+1}^\star \le \Delta$. ■

Polynomial decay. For polynomial decay, we consider $\psi(i) = Ci^{-(1+\gamma)}$. Then $\bar{r}+1 = 1. \inf\{i : C(i)^{-(1+\gamma)} \le \Delta\} = 1 + \inf\{i : i \ge (\Delta/C)^{-\frac{1}{1+\gamma}}\}$. Hence, $\bar{r} \ge (\Delta/C)^{-\frac{1}{1+\gamma}}$ and $\bar{r}+1 \le (\Delta/C)^{-\frac{1}{1+\gamma}}$. By Lemma H.1, we have

$$
\mathsf{tail}_2^\star(\bar{r}+1) \le 2C^2(1 + \frac{1}{1+2\gamma})(\bar{r}+1)^{-1-2\gamma} \le 2C^2(\Delta/C)^{\frac{1+2\gamma}{1+\gamma}}. \tag{H.1}
$$

Thus, by Claim H.4, the above display, and the bound $\bar{r} \le (\Delta/C)^{-\frac{1}{1+\gamma}}$,

$$
\begin{aligned}
\mathsf{tail}_2^\star(r) &\le (\bar{r}-1)\Delta^2 + 2C^2(\Delta/C)^{\frac{1+2\gamma}{1+\gamma}} \\
&\le (\Delta/C)^{-\frac{1}{1+\gamma}}\Delta^2 + 2C^2(\Delta/C)^{\frac{1+2\gamma}{1+\gamma}} \\
&= \Delta^{\frac{1+2\gamma}{1+\gamma}} C^{\frac{1}{1+\gamma}} + +2C^2(\Delta/C)^{\frac{1+2\gamma}{1+\gamma}} \\
&= 3C^2(\Delta/C)^{\frac{1+2\gamma}{1+\gamma}} := 3C^2(\boldsymbol{\sigma}_r^\star/C)^{\frac{1+2\gamma}{1+\gamma}}
\end{aligned}
$$

Thus, using the above display and $\boldsymbol{\sigma}_r^\star \le Cr^{-(1+\gamma)}$.

$$
\frac{\mathsf{tail}_2^\star(r)^2}{(\boldsymbol{\sigma}_r^\star)^2} \le 3C^2(\boldsymbol{\sigma}_r^\star/C)^{\frac{2(1+2\gamma)}{1+\gamma}-2} = 3C^2(\boldsymbol{\sigma}_r^\star/C)^{\frac{2\gamma}{1+\gamma}} \le 3C^2 r^{-2\gamma}
$$

Exponential decay. For polynomial decay, we consider $\psi(i) = Ce^{-\gamma i}$. Then $\bar{r} = \inf\{i : Ce^{-\gamma i} \le \Delta\} = \inf\{i : i \ge \gamma^{-1}\log\frac{C}{\Delta}\}$. Hence,

$$
\bar{r} \ge \gamma^{-1}\log\frac{C}{\Delta}, \quad \bar{r}-1 \le \gamma^{-1}\log\frac{C}{\Delta}.
$$

Then, by Lemma H.1,

$$
\mathsf{tail}_2^\star(\bar{r}) \le C^2(1+\gamma^{-1})(e^{-\gamma(\bar{r}+1)})^2 \le (1+\gamma^{-1})\Delta^2 \le C^2(1+\gamma^{-1})\Delta^2. \tag{H.2}
$$

Thus by Claim H.4,

$$\mathsf{tail}^\star_2(r) \le C^2(1+\gamma^{-1}+\log\frac{C}{\Delta})\Delta^2 = C^2(1+\gamma^{-1}+\gamma^{-1}\log\frac{C}{\boldsymbol{\sigma}^\star_r})(\boldsymbol{\sigma}^\star_r)^2$$

Hence,

$$\frac{\mathsf{tail}^\star_2(r)^2}{(\boldsymbol{\sigma}^\star_r)^2} \le C^2(1+\gamma^{-1}+\gamma^{-1}\log\frac{C}{\boldsymbol{\sigma}^\star_r})^2(\boldsymbol{\sigma}^\star_r)^2$$

As $x\log(1/x)$ is increasing in x, and as $\boldsymbol{\sigma}^\star_r \le Ce^{-\gamma r}$, the above is at most

$$\frac{\mathsf{tail}^\star_2(r)^2}{(\boldsymbol{\sigma}^\star_r)^2} \le C^2(1+\gamma^{-1}+r)^2e^{-2\gamma r}.$$

■

Part III

Supplement for the Meta-Theorem

Appendix I. Factor Recovery for Matrix Factorization

We recall the setup for matrix factor recovery; its relation to the bilinear embeddings is described in Section 4.4. For matrices $\mathbf{A}^\star, \hat{\mathbf{A}} \in \mathbb{R}^{n\times d}$, $\mathbf{B}^\star, \hat{\mathbf{B}} \in \mathbb{R}^{m\times d}$, and matrices $\mathbf{M}^\star, \hat{\mathbf{M}}$, and for orthogonal matrices $\mathbf{R} \in \mathbb{O}(d)$, consider the error terms

$$\Delta_0(\mathbf{R},k) = \|(\mathbf{A}^\star_{[k]} - \hat{\mathbf{A}}\mathbf{R})(\mathbf{B}^\star_{[k]})^\top\|_{\mathrm{F}}^2 \vee \|\mathbf{A}^\star_{[k]}(\mathbf{B}^\star_{[k]} - \hat{\mathbf{B}}\mathbf{R})^\top\|_{\mathrm{F}}^2 \tag{I.1}$$

$$\Delta_1(\mathbf{R},k) = \|\mathbf{A}^\star_{[k]} - \hat{\mathbf{A}}\mathbf{R}\|_{\mathrm{F}}^2 \vee \|\mathbf{B}^\star_{[k]} - \hat{\mathbf{B}}\mathbf{R}\|_{\mathrm{F}}^2, \tag{I.2}$$

where $\mathbf{A}^\star_{[k]}$ and $\mathbf{B}^\star_{[k]}$ are the rank-k approximations of $\mathbf{A}^\star$ and $\mathbf{B}^\star$; formally[6]

$$\mathbf{A}^\star_{[k]} = \mathbf{A}^\star\mathbf{P}^\star_{[k]}, \quad \mathbf{B}^\star_{[k]} = \mathbf{B}^\star\mathbf{P}^\star_{[k]}, \quad \mathbf{P}^\star_{[k]} \in \text{projection on top-}k\text{ eigenspace of } (\mathbf{A}^\star)^\top\mathbf{A}^\star = (\mathbf{B}^\star)^\top\mathbf{B}^\star. \tag{I.3}$$

Our guarantee for controlling these matrix error terms is perhaps the most challenging technical ingredient of the paper. We state the following theorem, of which Theorem 5 is a specialization.

Theorem 7 *Let $\mathbf{A}^\star, \hat{\mathbf{A}} \in \mathbb{R}^{n\times d}$, $\mathbf{B}^\star, \hat{\mathbf{B}} \in \mathbb{R}^{m\times d}$, and suppose $(\mathbf{A}^\star, \mathbf{B}^\star)$ and $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ are balanced factorizations of $\mathbf{M}^\star = \mathbf{A}^\star(\mathbf{B}^\star)^\top$, and $\hat{\mathbf{M}} = \hat{\mathbf{A}}\hat{\mathbf{B}}^\top$. Let $r = \mathrm{rank}(\hat{\mathbf{M}})$. Fix $\epsilon > 0$ and $s \in \mathbb{N}$ such that $s > 1$, $\epsilon \ge \|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{F}}$, and $\epsilon \le \frac{\|\mathbf{M}^\star\|_{\mathrm{op}}}{40s}$. Also, for $q \ge 1$, let $\mathsf{tail}_q(\mathbf{M};k) := \sum_{i>k}\sigma_i(\mathbf{M})^q$. Then,*

6. While $\mathbf{P}^\star_{[k]}$ is non-unique in general, Theorem 7 ensures that there is a spectral gap at rank k, ensuring $\mathbf{P}^\star_{[k]}$ is indeed unique.

(a) *There exists an index* $k \in [\min\{r, s-1\}]$ *and an orthogonal matrix* $\mathbf{R} \in \mathbb{O}(d)$ *such that*

$$\text{(weighted error)} \quad \Delta_0(\mathbf{R}, k) \lesssim \epsilon^2 \cdot s^2 \ell_\star(\epsilon, s) \tag{I.4a}$$

$$\text{(unweighted error)} \quad \Delta_1(\mathbf{R}, k) \lesssim (\sqrt{r} + s^2)\epsilon + s\sigma_s(\mathbf{M}^\star) + \mathsf{tail}_1(\mathbf{M}^\star; s), \tag{I.4b}$$

where we define $\ell_\star(\epsilon, s) := \min\left\{1 + \log \frac{\|\mathbf{M}^\star\|_{\mathrm{op}}}{40 s\epsilon},\ s\right\}$.

(b) *Moreover, the index* k *satisfies*

$$\mathsf{tail}_2(\mathbf{M}^\star; k) \lesssim s^3\epsilon^2 + s(\sigma_s(\mathbf{M}^\star))^2 + \mathsf{tail}_2(\mathbf{M}^\star; s).$$

(c) *The matrix* $\mathbf{R}$ *and* k *satisfy* $(\hat{\mathbf{A}}\mathbf{R})^\top \hat{\mathbf{A}}\mathbf{R} \succeq 39\epsilon \mathbf{P}^\star_{[k]}$ *and* $\sigma_k(\mathbf{M}^\star) - \sigma_{k+1}(\mathbf{M}^\star) \geq 40\epsilon/s$.

Explanation of Theorem 7. There are a few essential points to the theorem, which we outline below.

- The parameter ϵ^2 upper bounds $\|\mathbf{M}^\star - \hat{\mathbf{M}}\|_{\mathrm{F}}^2$. When instantiated as in Section 4.4, ϵ^2 bounds $\|\mathbf{M}_{1\otimes 1}(f, g) - \mathbf{M}_{1\otimes 1}(f^\star, g^\star)\|_{\mathrm{F}}^2 = \mathcal{R}(f, g; \mathcal{D}_{1\otimes 1})$. Because $\mathcal{D}_{\mathrm{train}}$ covers $\mathcal{D}_{1\otimes 1}$, this ensures that we can choose ϵ sufficiently small for non-vacuous bounds.

- The theorem guarantees the existence of some index $k \in [s-1]$ for which the error terms with respect to the rank-k approximation is small. It may not be the case that $k = s-1$, and indeed the construction of this index k can be subtle. Fortunately, this index k is only important for the analysis, and need not be known by the algorithm. Such an index k leads to a partition of the singular values that enable us to better control the relative spectral gap (see formal definition in Appendix I.1). This is the key to obtaining our improved bounds compared to the literature.

- Part (a) of the theorem bounds Δ_0 and Δ_1. Our bound on Δ_0 is *much smaller* than that on Δ_1, scaling quadratically in ϵ instead of linearly. This emphasizes the importance of weighting by the co-factors $\mathbf{A}^\star_{[k]}$ and $\mathbf{B}^\star_{[k]}$ in Eq. (I.1), or equivalently (via the discussion in Section 4.4), by the rank-reduced embeddings $f_k^\star, g_k^\star$ in Definition 4.3.

- Part (b) stipulates that truncating the spectrum of $\mathbf{M}^\star$ at the index k is not much worse than truncating the spectrum at the stipulated index s. Note that $\mathsf{tail}_2(\mathbf{M}^\star; k)$ corresponds to $\textbf{tail}_2^\star(k)$ (in Proposition 4.1) under the choices in Section 4.4. Hence, this is useful for handling the term $\textbf{tail}_2^\star(k)$ that emerges in the risk decomposition therein.

- Finally, the statement ensures that, even though $\mathbf{M}^\star$ *may not* have a spectral gap at its s-singular value, the stipulated index k does ensure $\sigma_k(\mathbf{M}^\star) - \sigma_{k+1}(\mathbf{M}^\star) \geq 40\epsilon/s$. Moreover, it also ensures that, after the rotation $\mathbf{R}$, the column-space of $\hat{\mathbf{A}}\mathbf{R}$ contains the column space of $\mathbf{A}^\star$; this corresponds to Eq. (4.1), and ensures that the chosen rotation $\mathbf{R}$ makes $(\mathbf{R}^\top f, \mathbf{R}^\top g)$ *aligned proxies*. This latter statement also gives some quantitative wiggle room when applying limiting arguments for continuous distributions (see more details in Appendix M).

I.1. Proof roadmap

Technical challenges. The key challenge throughout the proof of Theorem 7 is that many classical matrix perturbation bounds (e.g. Wedin's theorem) require some form of *separation* (i.e. *gaps*) among the singular values of the matrix to which it is being applied. In sharp contrast, we assume no such condition on gaps in the spectrum of $\mathbf{M}^\star$.

Specifically, we appeal to a lemma due to Tu et al. (2016) (see the restatement in Lemma I.5), which controls (up to a rotation) the Frobenius error of the factors $\mathbf{X}_1 - \mathbf{R}\mathbf{X}_2,\ \mathbf{Y}_1 - \mathbf{R}\mathbf{Y}_2$ in terms of the Frobenius error between their outer products $\mathbf{Z}_1 = \mathbf{X}_1\mathbf{Y}_1^\top$ and $\mathbf{Z}_2 = \mathbf{X}_2\mathbf{Y}_2^\top$. When applied directly to $\hat{\mathbf{M}} = \hat{\mathbf{A}}\hat{\mathbf{B}}^\top$ and $\mathbf{M}^\star = \mathbf{A}^\star(\mathbf{B}^\star)^\top$, Lemma I.5 has numerous limitations: (a) it requires the factorization of $\hat{\mathbf{M}}$ and $\mathbf{M}^\star$ to have the same rank k; (b) it requires a sufficiently large spectral gap on $\sigma_k(\mathbf{M}^\star) - \sigma_{k+1}(\mathbf{M}^\star)$; (c) the error bounds scale with the inverse of this gap, which can be very loose when $\sigma_k(\mathbf{M}^\star)$ becomes small.

Our techniques. Instead of applying Lemma I.5 directly, we construct a certain partition of the spectrum of $\mathbf{M}^\star$, what we call a "well-tempered partition" (Definition I.3), which partitions the indices $[s]$ of the top-s singular values of $\mathbf{M}^\star$ into intervals where (a) all singular values are of similar magnitude, and (b) the separation between the intervals is sufficiently large. Condition (b) is necessary for applying gap-dependent perturbation bounds, but condition (a) allows us to refine these bounds tremendously.

Specifically, we denote the subsets in this partition as $\mathcal{K}_i = \{k_i + 1, k_i + 2, \ldots, k_{i+1}\}$; we call k_i the *pivot*. We show that the partition ensures that the *relative gap*

$$\delta_{k_i} = \frac{\sigma_{k_i}(\mathbf{M}^\star) - \sigma_{k_i+1}(\mathbf{M}^\star)}{\sigma_{k_i}(\mathbf{M}^\star)} \tag{I.5}$$

is at least $\Omega(1/s)$. By contrast, note that with exponentially decaying singular values, the *absolute gap* $\sigma_{k_i}(\mathbf{M}^\star) - \sigma_{k_i+1}(\mathbf{M}^\star)$ can be exponentially small.

With a careful change-of-basis, the above spectral partition induces a decomposition of $\mathbf{A}^\star_{[k]}, \mathbf{B}^\star_{[k]}$ and $\hat{\mathbf{A}}, \hat{\mathbf{B}}$ into blocks according to the indices in the set $\mathcal{K}_i$. We then apply Lemma I.5 *separately along each block*, arguing that the factorization error is small *block-wise*. This significantly sharpens our control over Δ_0 (recall the definition in Eq. (I.1)) because we weight the error in block i by the largest singular value in that block. Working through the algebra, we end up only paying for the *relative gap*, which as noted above is $\Omega(1/s)$.

For both Δ_0 and Δ_1, the above partition also has the advantage (indeed, necessity) that, by restricting to each set $\mathcal{K}_i$ in the partition, we only need to consider the factorizations of the same rank, and lower-bounded relative spectral gap. Recall that Lemma I.5 establishes error bounds on factors in terms of error bounds on their outer-product. By decomposing our matrices into their restriction to the singular values index by $\mathcal{K}_i$, we therefore need some way of controlling the following: *Denote by $\mathbf{M}^\star_{\mathcal{K}_i}, \hat{\mathbf{M}}_{\mathcal{K}_i}$ the SVDs of $\mathbf{M}^\star, \hat{\mathbf{M}}$ containing only singular values indexed by $j \in \mathcal{K}_i$. How large is $\|\mathbf{M}^\star_{\mathcal{K}_i} - \hat{\mathbf{M}}_{\mathcal{K}_i}\|_{\mathrm{F}}$, in terms of $\|\mathbf{M}^\star - \hat{\mathbf{M}}\|_{\mathrm{F}}$?*

Again, our control over relative spectral gaps come to the rescue. Here, we invoke Theorem 1, which shows that the error in the SVDs between these objects grows only with the *relative* spectral gap, which as we have stressed, is well-controlled. This again reduces the dependence on the small singular values of $\mathbf{M}^\star$, which improves our bounds.

The formal proof is quite involved. Hence, we begin with an extensive setup of preliminaries, simplification and useful notation before diving into the main arguments. But to summarize, the key tools are: (a) Lemma I.5 due to Tu et al. (2016), (b) our novel construction of the "well-tempered partition" of the spectrum of $\mathbf{M}^\star$, and (c) our novel relative-error perturbation bound.

I.2. Proof preliminaries

Singular value notation. We introduce the following notation for the singular values of $\mathbf{M}^\star$ and their relative gaps:

$$\sigma_k^\star := \sigma_k(\mathbf{M}^\star), \quad \sigma_0^\star = +\infty, \quad \delta_k^\star := 1 - \frac{\sigma_{k+1}(\mathbf{M}^\star)}{\sigma_k(\mathbf{M}^\star)}, \quad \delta_0^\star = 1.$$

Explicit factorization. We argue that, without loss of generality, we can pick factors $\hat{\mathbf{A}}, \hat{\mathbf{B}}, \mathbf{A}^\star, \mathbf{B}^\star$ of a canonical form. Construct the SVDs of $\hat{\mathbf{M}}$ and $\mathbf{M}^\star$ as

$$\hat{\mathbf{M}} = \hat{\mathbf{U}}\hat{\mathbf{\Sigma}}\hat{\mathbf{V}}^\top, \quad \mathbf{M}^\star = \mathbf{U}^\star\mathbf{\Sigma}^\star(\mathbf{V}^\star)^\top,$$

where $\hat{\mathbf{U}}, \mathbf{U}^\star \in \mathbb{R}^{n\times d}$, $\hat{\mathbf{V}}, \mathbf{V}^\star \in \mathbb{R}^{m\times d}$, and $\hat{\mathbf{\Sigma}}, \mathbf{\Sigma}^\star \in \mathbb{R}^{p\times d}$ are diagonal matrices with non-negative entries arranged in (non-strictly) descending order. Note that $p \le \min\{n, m\}$. We now argue that we may assume the factor matrices take the following form, without loss of generality:

$$\hat{\mathbf{A}} = \hat{\mathbf{U}}\hat{\mathbf{\Sigma}}^{\frac{1}{2}}, \quad \hat{\mathbf{B}} = \hat{\mathbf{V}}\hat{\mathbf{\Sigma}}^{\frac{1}{2}}, \quad \mathbf{A}^\star = \mathbf{U}^\star(\mathbf{\Sigma}^\star)^{\frac{1}{2}}, \quad \mathbf{B}^\star = \mathbf{V}^\star(\mathbf{\Sigma}^\star)^{\frac{1}{2}}. \tag{I.6}$$

One can check that a valid choice of rank-k SVD for $\mathbf{M}^\star$ is given by $\mathbf{A}^\star_{[k]}(\mathbf{B}^\star_{[k]})^\top$, where

$$\mathbf{A}^\star_{[k]} = \mathbf{U}^\star(\mathbf{\Sigma}^\star_{[k]})^{\frac{1}{2}}, \quad \mathbf{B}^\star_{[k]} = \mathbf{V}^\star(\mathbf{\Sigma}^\star_{[k]})^{\frac{1}{2}},$$

and $\mathbf{\Sigma}^\star_{[k]}$ zeroes out all but the first k entries of $\mathbf{\Sigma}^\star$. The assumption that the matrices take the above form is justified by the following lemma, which shows that any bounds on Δ_0, Δ_1 hold for the factorization in Eq. (I.6).

Lemma I.1 *Assume that $\hat{\mathbf{A}}, \hat{\mathbf{B}}, \mathbf{A}^\star, \mathbf{B}^\star$ take the form Eq. (I.6). Let $(\hat{\mathbf{A}}', \hat{\mathbf{B}}')$ and $(\mathbf{A}^{\star\prime}, \mathbf{B}^{\star\prime})$ be any other rank-d balanced factorizations of the matrices $\hat{\mathbf{M}}$ and $\mathbf{M}^\star$, respectively. Then, for any $\mathbf{R}' \in \mathbb{O}(d)$, there exists a $\mathbf{R} \in \mathbb{O}(d)$ such that*

$$\|(\mathbf{A}^\star_{[k]} - \hat{\mathbf{A}}\mathbf{R})(\mathbf{B}^\star_{[k]})^\top\|_{\mathrm{F}}^2 = \|(\mathbf{A}^{\star\prime}_{[k]} - \hat{\mathbf{A}}'\mathbf{R}')(\mathbf{B}^{\star\prime}_{[k]})^\top\|_{\mathrm{F}}^2, \quad \|\mathbf{A}^\star_{[k]}(\mathbf{B}^\star_{[k]} - \hat{\mathbf{B}}\mathbf{R})^\top\|_{\mathrm{F}}^2 = \|\mathbf{A}^{\star\prime}_{[k]}(\mathbf{B}^{\star\prime}_{[k]} - \hat{\mathbf{B}}'\mathbf{R}')^\top\|_{\mathrm{F}}^2$$
$$\|\mathbf{A}^\star_{[k]} - \hat{\mathbf{A}}\mathbf{R}\|_{\mathrm{F}}^2 = \|\mathbf{A}^{\star\prime}_{[k]} - \hat{\mathbf{A}}'\mathbf{R}'\|_{\mathrm{F}}^2, \quad \|\mathbf{B}^\star_{[k]} - \hat{\mathbf{B}}\mathbf{R}\|_{\mathrm{F}}^2 = \|\mathbf{B}^{\star\prime}_{[k]} - \hat{\mathbf{B}}'\mathbf{R}'\|_{\mathrm{F}}^2.$$

Moreover, if

$$\mathrm{rowspace}\left((\hat{\mathbf{A}}_{[k]}\mathbf{R})^\top(\hat{\mathbf{A}}_{[k]}\mathbf{R})\right) \supseteq \mathrm{rowspace}\left((\mathbf{A}^\star_{[k]})^\top(\mathbf{A}^\star_{[k]})\right),$$

then

$$\mathrm{rowspace}\left((\hat{\mathbf{A}}'_{[k]}\mathbf{R}')^\top(\hat{\mathbf{A}}'_{[k]}\mathbf{R}')\right) \supseteq \mathrm{rowspace}\left((\mathbf{A}^{\star\prime}_{[k]})^\top(\mathbf{A}^{\star\prime}_{[k]})\right).$$

The above lemma is proved with the following fact.

Lemma I.2 *Let $(\mathbf{A}, \mathbf{B})$ be a rank-at-most-d balanced factorization of a matrix $\mathbf{M} \in \mathbb{R}^{n\times m}$. Denote a SVD $\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ (with $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$). Then, there exists a rotation matrix $\mathbf{R} \in \mathbb{O}(d)$ such that $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}^{1/2}\mathbf{R}$ and $\mathbf{B} = \mathbf{V}\mathbf{\Sigma}^{1/2}\mathbf{R}$. Moreover, this $\mathbf{R}$ satisfies $\mathbf{A}_{[k]} = \mathbf{U}\mathbf{\Sigma}_{[k]}^{1/2}\mathbf{R}$, where $\mathbf{\Sigma}_{[k]}$ masks the all but the first k entries of $\mathbf{\Sigma}$, where $\mathbf{A}_{[k]}$ is consistent with the definition of SVD as Eq. (I.3). We similarly define $\mathbf{B}_{[k]}$.*

The proofs of Lemmas I.1 and I.2 are given in Appendix K.1.

Masking. It will be convenient to have compact notation for masking entries of matrices. Recall that the index d refers to the "inner dimension", e.g. $\hat{\mathbf{A}} \in \mathbb{R}^{n\times d}$.

To begin, we define masking for *square matrices*. For matrices $\mathbf{X} \in \mathbb{R}^{p\times d}$, and $\mathcal{K} \subseteq [d]$, define the matrix $\mathbf{X}_{\mathcal{K}} \in \mathbb{R}^{p\times d}$ by masking $\mathbf{X}$'s entries in $\mathcal{K}$:

$$(\mathbf{X}_{\mathcal{K}})_{ij} = \mathbb{I}\{i \in \mathcal{K} \text{ and } j \in \mathcal{K}\} \cdot \mathbf{X}_{ij}.$$

We also define the shorthand notation

$$\mathbf{X}_{>k} := \mathbf{X}_{[d]\setminus[k]},$$

with the convention that $\mathbf{X}_{[0]} = 0$.

Next, we define masking for *factors matrices*. Given a matrix of the form $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}^{\frac{1}{2}}$, where $\mathbf{U} \in \mathbb{R}^{n\times d}$ and $\mathbf{\Sigma} \in \mathbb{R}^{p\times d}$ is diagonal, we define

$$\mathbf{A}_{\mathcal{K}} := \mathbf{U}\mathbf{\Sigma}_{\mathcal{K}}^{\frac{1}{2}}, \quad \mathbf{A}_{[k]} = \mathbf{U}\mathbf{\Sigma}_{[k]}^{\frac{1}{2}}, \quad \mathbf{A}_{>k} = \mathbf{U}\mathbf{\Sigma}_{>k}^{\frac{1}{2}}, \quad (\mathbf{A}, \mathbf{U}, \mathbf{\Sigma}) \in \{(\mathbf{A}^\star, \mathbf{U}^\star, \mathbf{\Sigma}^\star), (\hat{\mathbf{A}}, \hat{\mathbf{U}}, \hat{\mathbf{\Sigma}})\}.$$

We define analogous notation for $\hat{\mathbf{B}}, \mathbf{B}^\star$. Finally, we define

$$\mathbf{M}_{\mathcal{K}} = \mathbf{A}_{\mathcal{K}}\mathbf{B}_{\mathcal{K}}^\top, \quad \mathbf{M}_{[k]} = \mathbf{A}_{[k]}\mathbf{B}_{[k]}^\top, \quad \mathbf{M}_{>k} = \mathbf{A}_{>k}\mathbf{B}_{>k}^\top$$
$$(\mathbf{A}, \mathbf{B}, \mathbf{M}) \in \{(\mathbf{A}^\star, \mathbf{B}^\star, \mathbf{M}^\star), (\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{M}})\}.$$

In particular, $\mathbf{M}_{[k]}$ is the rank-k approximation of $\mathbf{M}$, for $\mathbf{M} \in \{\hat{\mathbf{M}}, \mathbf{M}^\star\}$.

Partitions & Compatibility. Importantly, we consider the set $\mathcal{K}$ which partitions the inner dimension d into disjoint intervals. We call these sets monotone partitions.

Definition I.1 (Monotone Partitions) *We say that $(\mathcal{K}_i)_{i=1}^{\ell}$ is a partition of $[d]$ if $\bigcup_{i=1}^{\ell} \mathcal{K}_i = [d]$, and the sets $\{\mathcal{K}_i\}_{i=1}^{\ell}$ are pairwise disjoint. We say that it is a* monotone *partition if there exists integers $0 = k_1 < \cdots < k_\ell < k_{\ell+1} = p$ such that $\mathcal{K}_i = \{k_i + 1, \ldots, k_{i+1}\}$. In particular, this means $\mathcal{K}_\ell = \{k_\ell + 1, \ldots, p\}$. We call the entries k_i the* pivots *of the monotone partition, and call the entry k_ℓ the* final pivot.

Definition I.2 (Compatibility) *We say a matrix $\mathbf{X} \in \mathbb{R}^{d\times d}$ is* compatible *with a partition $(\mathcal{K}_i)_{i=1}^{\ell}$ of $[d]$ if $\mathbf{X} = \sum_{i=1}^{\ell} \mathbf{X}_{\mathcal{K}_i}$.*

In particular, if $(\mathcal{K}_i)_{i=1}^{\ell}$ is a monotone partition, then compatibility means that $\mathbf{X}$ is a block-diagonal matrix whose blocks corresponding to the indices in the sets $\mathcal{K}_i$ for all $i = 1, \cdots, \ell$.

I.3. Key error decomposition results

The following lemma decomposes the error across the partitions:

Lemma I.3 *Let $(\mathcal{K}_i)_{i=1}^{\ell}$ be a monotone partition, and let $\mathbf{R} \in \mathbb{R}^{p\times d}$ be compatible with $(\mathcal{K}_i)_{i=1}^{\ell}$. Then*

$$(\mathbf{A}^\star_{[k_\ell]} - \hat{\mathbf{A}}\mathbf{R})(\mathbf{B}^\star_{[k_\ell]})^\top = \sum_{i=1}^{\ell-1} \left(\mathbf{A}^\star_{\mathcal{K}_i} - \hat{\mathbf{A}}_{\mathcal{K}_i}\mathbf{R}_{\mathcal{K}_i}\right)(\mathbf{B}^\star_{\mathcal{K}_i})^\top, \tag{I.7a}$$

$$(\mathbf{A}^\star_{[k_\ell]} - \hat{\mathbf{A}}\mathbf{R}) = \sum_{i=1}^{\ell-1} \left(\mathbf{A}^\star_{\mathcal{K}_i} - \hat{\mathbf{A}}_{\mathcal{K}_i}\mathbf{R}_{\mathcal{K}_i}\right) - \hat{\mathbf{A}}_{>k_\ell}\mathbf{R}_{>k_\ell}, \tag{I.7b}$$

with the analogous composition being true for $(\mathbf{B}^\star_{[k_\ell]} - \hat{\mathbf{B}}\mathbf{R})$ and $\mathbf{A}^\star_{[k_\ell]}(\mathbf{B}^\star_{[k_\ell]} - \hat{\mathbf{B}}\mathbf{R})^\top$.

In what follows, we choose the matrix $\mathbf{R}$ above to be orthogonal. However, orthogonality of $\mathbf{R}$ is not strictly necessary the decomposition in Lemma I.3. Conveniently, the decomposition of $(\mathbf{A}^\star_{[k_\ell]} - \hat{\mathbf{A}}\mathbf{R})(\mathbf{B}^\star_{[k_\ell]})^\top$ does not incur a dependence on the tail $\hat{\mathbf{A}}_{>k_\ell}\mathbf{R}_{>k_\ell}$ of $\hat{\mathbf{A}}$, thereby avoiding all singular values after the final pivot. This is one of the reasons why the weighted error Δ_0 ends up being smaller than the unweighted Δ_1. The proof of Lemma I.3 is given in Subsubsection K.2.1.

In our analysis, we consider partitions of $[d]$ that enjoy favorable spectral properties: first, every pivot k_i has large relative spectral gap $\delta^\star_{k_i}$ (for $\mathbf{M}^\star$), and second, the largest singular value in each partition is at most a constant times that of the pivot of the next partition.

Definition I.3 (Well-Tempered Partition) *We say a partition $(\mathcal{K}_i)_{i=1}^{\ell}$ is (δ,μ)-*well-tempered *if it is monotone, and for all $i \in [\ell]$, the corresponding pivots $k_i = \min_k\{k-1 : k \in \mathcal{K}_i\}$ satisfy*

(a) $\delta^\star_{k_i} \geq \delta$;

(b) $\max\{\sigma^\star_{k'} : k' \in \mathcal{K}_i\} \leq \mu \cdot \sigma^\star_{k_{i+1}}$.

For such a partition, we define the constants

$$M_{\mathrm{space}} := \sum_{i=1}^{\ell} (\delta^\star_{k_i})^{-2}, \;\; \textit{and} \;\; M_{\mathrm{spec}} := \sum_{i=1}^{\ell} (\sigma^\star_{k_i})^{-1}.$$

At the end of the proof, we show that a well-tempered partition always exists with $\mu = \mathcal{O}(1)$ and $\delta = 1/k$, and where M_{space} and M_{spec} are well-behaved. For now, let us carry the analysis out in terms of the properties of the supposed partition. The key object in the analysis is the normalized error:

Definition I.4 (Normalized Factored Error) *Let $(\mathcal{K}_i)_{i=1}^{\ell}$ be a monotone partition. We define the* normalized error term *for $i \in [\ell]$ as*

$$E_i(\mathbf{R}) := \left\{(\delta^\star_{k_i} \wedge \delta^\star_{k_{i+1}})^2(\sigma^\star_{k_{i+1}})\right\} \cdot \max\left\{\|\mathbf{A}^\star_{\mathcal{K}_i} - \hat{\mathbf{A}}_{\mathcal{K}_i}\mathbf{R}_{\mathcal{K}_i}\|_{\mathrm{F}}^2, \|\mathbf{B}^\star_{\mathcal{K}_i} - \hat{\mathbf{B}}_{\mathcal{K}_i}\mathbf{R}_{\mathcal{K}_i}\|_{\mathrm{F}}^2\right\}.$$

We now bound our given error terms in terms of the E_i-quantities. The proof of the following lemma is given in Subsubsection K.2.2.

Lemma I.4 *Let $\mathbf{R} \in \mathbb{O}(d)$ be compatible with a (δ,μ)-well-tempered partition $(\mathcal{K}_i)_{i=1}^{\ell}$ with pivots $\{k_i\}_{i\in[\ell]}$. Then*

$$\Delta_0(\mathbf{R}, k_\ell) \leq 2\mu \cdot M_{\mathrm{space}} \cdot \max_{i\in[\ell-1]} E_i(\mathbf{R}) \tag{I.8a}$$

$$\Delta_1(\mathbf{R}, k_\ell) \leq \frac{M_{\mathrm{spec}}}{\delta^2} \cdot \max_{i\in[\ell-1]} E_i(\mathbf{R}) + \sum_{i>k_\ell} \sigma_i(\hat{\mathbf{M}}). \tag{I.8b}$$

I.4. Controlling the normalized errors

As shown in Lemma I.4, bounds on both Δ_0 and Δ_1 amount to bounding (suitably rotated) errors between the factorizations of the ground-truth and estimated matrix. To do so, we invoke the following factorization lemma due to Tu et al. (2016).

Lemma I.5 ((Tu et al., 2016), Lemma 5.14) *Let* $(\mathbf{A}, \mathbf{B})$ *and* $(\mathbf{A}', \mathbf{B}')$ *be rank-*r *balanced factorization of matrices* $\mathbf{M}$ *and* $\mathbf{M}'$*, respectively. Suppose that* $\|\mathbf{M} - \mathbf{M}'\|_{\mathrm{op}} \le \frac{1}{2}\sigma_r(\mathbf{M})$*. Then, there exists an orthogonal matrix* $\mathbf{O} \in \mathbb{O}(r)$ *such that*

$$\|\mathbf{A} - \mathbf{A}'\mathbf{O}\|_{\mathrm{F}}^2 + \|\mathbf{B} - \mathbf{B}'\mathbf{O}\|_{\mathrm{F}}^2 \le c_0 \frac{\|\mathbf{M} - \mathbf{M}'\|_{\mathrm{F}}^2}{\sigma_r(\mathbf{M})},$$

where $c_0 = \frac{2}{\sqrt{2}-1}$.

As a consequence, we can deduce the following bound on the normalized error terms.

Lemma I.6 *Let* $(\mathcal{K}_i)_{i=1}^{\ell}$ *be a* (δ, μ)*-well-tempered partition with pivots* $\{k_i\}_{i \in [\ell]}$*. Define*

$$\tilde{\epsilon}_{\mathrm{op}} := \max_{i \in [\ell+1]} \updelta_{k_i}^{\star} \|\hat{\mathbf{M}}_{[k_i]} - \mathbf{M}_{[k_i]}^{\star}\|_{\mathrm{op}}, \quad \tilde{\epsilon}_{\mathrm{fro}} := \max_{i \in [\ell+1]} \updelta_{k_i}^{\star} \|\hat{\mathbf{M}}_{[k_i]} - \mathbf{M}_{[k_i]}^{\star}\|_{\mathrm{F}}. \tag{I.9}$$

Then, if $\tilde{\epsilon}_{\mathrm{op}} \le \frac{\delta \sigma_{k_\ell}^{\star}}{4}$*, there exists a* $\mathbf{R} \in \mathbb{O}(d)$ *which is compatible with* $(\mathcal{K}_i)_{i=1}^{\ell}$ *such that*

$$\max_{i \in [\ell]} E_i(\mathbf{R}) \le 4 c_0 \tilde{\epsilon}_{\mathrm{fro}}^2 \lesssim \tilde{\epsilon}_{\mathrm{fro}}^2,$$

where $c_0 = \frac{2}{\sqrt{2}-1}$.

To prove the above lemma, we invoke Lemma I.5 to bound $\|\mathbf{A}_{\mathcal{K}_i}^{\star} - \hat{\mathbf{A}}_{\mathcal{K}_i}\mathbf{R}_{\mathcal{K}_i}\|_{\mathrm{F}}$ and $\|\mathbf{B}_{\mathcal{K}_i}^{\star} - \hat{\mathbf{B}}_{\mathcal{K}_i}\mathbf{R}_{\mathcal{K}_i}\|_{\mathrm{F}}$ in terms of $\|\mathbf{M}_{\mathcal{K}_i}^{\star} - \hat{\mathbf{M}}_{\mathcal{K}_i}\|_{\mathrm{F}}$. Then we notice that $\hat{\mathbf{M}}_{\mathcal{K}_i} = \hat{\mathbf{M}}_{[k_{i+1}]} - \hat{\mathbf{M}}_{[k_i]}$ and $\mathbf{M}_{\mathcal{K}_i}^{\star} = \mathbf{M}_{[k_{i+1}]}^{\star} - \mathbf{M}_{[k_i]}^{\star}$, so we can derive a bound directly in terms of the differences between rank-k_i SVDs. The proof is given in Subsubsection K.2.3. In applying Lemma I.6, we crudely bound $\tilde{\epsilon}_{\mathrm{op}} \le \tilde{\epsilon}_{\mathrm{fro}}$, but the above lemma is stated so that a more refined analysis may be possible.

We now bound $\tilde{\epsilon}_{\mathrm{fro}}$ using a generic bound for the SVD decomposition, which we state below.

Theorem 1 (Perturbation of SVD Approximation with Relative Gap) *Let* $\mathbf{M}^{\star}, \hat{\mathbf{M}} \in \mathbb{R}^{n \times m}$*. Fix a* $k \le \min\{n, m\}$ *for which* $\sigma_k(\mathbf{M}^{\star}) > 0$ *and the relative spectral gap* $\updelta_k(\mathbf{M}^{\star})$ *(Eq. (1.1)) is positive. Then, if* $\|\mathbf{M}^{\star} - \hat{\mathbf{M}}\|_{\mathrm{op}} \le \eta \sigma_k(\mathbf{M}^{\star}) \updelta_k(\mathbf{M}^{\star})$ *for some* $\eta \in (0, 1)$*, we have that the rank-*k *SVD approximations of* $\mathbf{M}^{\star}$ *and* $\hat{\mathbf{M}}$*, denoted as* $\mathbf{M}_{[k]}^{\star}$ *and* $\hat{\mathbf{M}}_{[k]}$*, are unique, and satisfy*

$$\left\|\hat{\mathbf{M}}_{[k]} - \mathbf{M}_{[k]}^{\star}\right\|_{\mathrm{F}} \le \frac{9\|\hat{\mathbf{M}} - \mathbf{M}^{\star}\|_{\mathrm{F}}}{\updelta_k(\mathbf{M}^{\star})(1 - \eta)}.$$

Taking $\eta = 1/10$, and noting that $\updelta_{k_i}^{\star} \ge \delta$ and $\sigma_{k_i}^{\star} \ge \sigma_{k_\ell}^{\star}$ for all pivots in a (δ, μ)-well-tempered partition, we have the following corollary.

Corollary I.1 *Suppose* $(\mathcal{K}_i)_{i=1}^{\ell}$ *is* (δ, μ)*-well-tempered, and* $\|\hat{\mathbf{M}} - \mathbf{M}^{\star}\|_{\mathrm{op}} \le \sigma_{k_\ell}^{\star} \delta / 10$*. Then,* $\tilde{\epsilon}_{\mathrm{fro}} \le 10 \|\hat{\mathbf{M}} - \mathbf{M}^{\star}\|_{\mathrm{F}}$.

Combining with Lemma I.6, and then Lemma I.4, we obtain the following guarantee.

Proposition I.1 *Suppose* $(\mathcal{K}_i)_{i=1}^{\ell}$ *is* (δ,μ)*-well-tempered, and* $\|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{F}} \le \delta\sigma^\star_{k_\ell}/40$*. Then, there exists an orthogonal matrix* $\mathbf{R} \in \mathbb{O}(d)$ *compatible with* $(\mathcal{K}_i)_{i=1}^{\ell}$ *such that*

$$\max_{i\in[\ell]} E_i(\mathbf{R}) \le 400c_0\|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{F}}^2 \lesssim \|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{F}}^2.$$

Therefore, by Lemma I.4, this orthogonal $\mathbf{R}$ *and index* k_ℓ *satisfy*

$$\Delta_0(\mathbf{R}, k_\ell) \lesssim \mu M_{\mathrm{space}}\|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{F}}^2, \tag{I.10a}$$

$$\Delta_1(\mathbf{R}, k_\ell) \lesssim \frac{M_{\mathrm{spec}}}{\delta^2}\|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{F}}^2 + \sqrt{r}\|\mathbf{M}^\star - \hat{\mathbf{M}}\|_{\mathrm{F}} + \mathsf{tail}_1(\mathbf{M}^\star; k_\ell). \tag{I.10b}$$

Proof If $\|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{F}} \le \delta\sigma^\star_{k_\ell}/40$, then also $\|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{op}} \le \sigma^\star_{k_\ell}\delta/10$, so $\tilde{\epsilon}_{\mathrm{fro}} \le 10\|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{F}}$. Hence, $\tilde{\epsilon}_{\mathrm{op}} \le \tilde{\epsilon}_{\mathrm{fro}} \le 10\|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{F}} \le \delta\sigma^\star_{k_\ell}/4$. Thus, by Lemma I.6 followed by Corollary I.1, there exists an orthogonal matrix $\mathbf{R} \in \mathbb{O}(d)$ compatible with $(\mathcal{K}_i)_{i=1}^{\ell}$ for which

$$\max_{i\in[\ell]} E_i(\mathbf{R}) \le 4c_0\tilde{\epsilon}_{\mathrm{fro}}^2 \le 400c_0\|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{F}}^2 \lesssim \|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{F}}^2.$$

Eq. (I.10a) now follows directly from Lemma I.4. To achieve Eq. (I.10b), we see that directly from Lemma I.4,

$$\Delta_1(\mathbf{R}, k_\ell) \lesssim \frac{M_{\mathrm{spec}}}{\delta^2}\|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{F}}^2 + \sum_{i>k_\ell}\sigma_i(\hat{\mathbf{M}}). \tag{I.11}$$

Using that $\mathrm{rank}(\hat{\mathbf{M}}) = r$, we have

$$\begin{aligned}
\sum_{i>k_\ell}\sigma_i(\hat{\mathbf{M}}) = \sum_{i=k_\ell+1}^{r}\sigma_i(\hat{\mathbf{M}}) &\le \sum_{i=k_\ell+1}^{r}\sigma_i(\mathbf{M}^\star) + \sum_{i=k_\ell+1}^{r}|\sigma_i(\mathbf{M}^\star) - \sigma_i(\hat{\mathbf{M}})| \\
&\le \sum_{i=k_\ell+1}^{r}\sigma_i(\mathbf{M}^\star) + \sqrt{r\sum_{i=k_\ell+1}^{r}|\sigma_i(\mathbf{M}^\star) - \sigma_i(\hat{\mathbf{M}})|^2} \\
&\le \sum_{i=k_\ell+1}^{r}\sigma_i(\mathbf{M}^\star) + \sqrt{r}\|\mathbf{M}^\star - \hat{\mathbf{M}}\|_{\mathrm{F}} \qquad \text{Lemma K.3} \\
&:= \mathsf{tail}_1(\mathbf{M}^\star; k_\ell) + \sqrt{r}\|\mathbf{M}^\star - \hat{\mathbf{M}}\|_{\mathrm{F}}.
\end{aligned}$$

The desired bound follows by combining with Eq. (I.11). ■

I.5. Existence of well-tempered partition

To conclude the proof, it suffices to demonstrate the existence of a well-tempered partition of the singular values of $\mathbf{M}^\star$, for which $M_{\mathrm{space}}, 1/\delta, \mu, M_{\mathrm{spec}}$ are all of reasonable magnitude. To do so, we focus on the pivots. One important subtlety is that, for any given $k \in \mathbb{N}$, $\delta^\star_k$ may be very small, indeed even equal to zero.

Hence, to construct the well-tempered partition, we take in a target rank k_{i+1}, and show that we can use a slightly smaller rank k_i for which $\delta^\star_{k_i} \ge 1/k_{i+1}$. We then argue that we can construct a sequence of pivots $k_1, k_2, \dots$ for which singular values within those pivots are within a constant factor (the μ-parameter for well-temperedness), the $\delta^\star_{k_i}$ parameters are lower bounded (hence lower bounding the δ parameter). In addition, this partition ensures that the singular values at the pivot points grow *at least* geometrically. This is helpful to control M_{space} and M_{spec}. The following technical lemma is proved in Subsubsection K.3.1.

Lemma I.7 (Singular Value Spacing) *Fix any $s \in \mathbb{N}$ and $\sigma \in [\sigma^\star_s, \sigma^\star_1]$. Then, there exists integer $\ell \in \mathbb{N}$, and an increasing sequence $0 = k_1 < k_2 \cdots < k_\ell < k_{\ell+1} = s$ such that the following is true:*

(a) For $i \in [\ell]$, $\delta^\star_{k_i} \geq 1/k_{i+1} \geq 1/s$.

(b) For $i = \ell$, $\sigma^\star_{k_i+1} \leq 2e\sigma$, and for $i \in [\ell-1]$, $\sigma^\star_{k_i+1} \leq 2e^2\sigma^\star_{k_{i+1}}$.

(c) For $i = \ell$, $\sigma^\star_{k_i} \geq \sigma$, and for $i \in [\ell-1]$, $\sigma^\star_{k_i} \geq e\sigma^\star_{k_{i+1}}$.

With this technical lemma in hand, we can demonstrate the existence of a well-tempered partition with a number of desirable properties. The following is proved in Subsubsection K.3.2.

Proposition I.2 (Well-Tempered Partition) *Fix any $s \in \mathbb{N}$ and $\sigma \in [\sigma^\star_s, \sigma^\star_1]$. There exists a partition $(\mathcal{K}_i)_{i=1}^{\ell}$ of $[s]$, which is (δ, μ)-well-tempered with parameters satisfying*

(a) $\delta \geq 1/s$ and $\mu \leq 2e^2$.

(b) $k_\ell < s$, $\sigma^\star_{k_\ell} \geq \sigma$, and $M_{\mathrm{spec}} \leq \frac{(\sigma)^{-1}}{1-e^{-1}}$.

(c) $M_{\mathrm{space}} \leq \ell_{\sigma,s} \cdot s^2$, where $\ell_{\sigma,s} := \min\{1 + \lceil \log \frac{\|\mathbf{M}^\star\|_{\mathrm{op}}}{\sigma} \rceil,\ s\}$.

(d) $\mathsf{tail}_1(\mathbf{M}^\star; k_\ell) \leq 2es\sigma + \mathsf{tail}_1(\mathbf{M}^\star; s)$ and $\mathsf{tail}_2(\mathbf{M}^\star; k_\ell) \leq 4e^2 s\sigma^2 + \mathsf{tail}_2(\mathbf{M}^\star; s)$.

I.6. Proof of Theorem 7

We recall the theorem here for convenience.

Theorem 7 *Let $\mathbf{A}^\star, \hat{\mathbf{A}} \in \mathbb{R}^{n\times d}$, $\mathbf{B}^\star, \hat{\mathbf{B}} \in \mathbb{R}^{m\times d}$, and suppose $(\mathbf{A}^\star, \mathbf{B}^\star)$ and $(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ are balanced factorizations of $\mathbf{M}^\star = \mathbf{A}^\star(\mathbf{B}^\star)^\top$, and $\hat{\mathbf{M}} = \hat{\mathbf{A}}\hat{\mathbf{B}}^\top$. Let $r = \mathrm{rank}(\hat{\mathbf{M}})$. Fix $\epsilon > 0$ and $s \in \mathbb{N}$ such that $s > 1$, $\epsilon \geq \|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{F}}$, and $\epsilon \leq \frac{\|\mathbf{M}^\star\|_{\mathrm{op}}}{40s}$. Also, for $q \geq 1$, let $\mathsf{tail}_q(\mathbf{M}; k) := \sum_{i>k} \sigma_i(\mathbf{M})^q$. Then,*

(a) There exists an index $k \in [\min\{r, s-1\}]$ and an orthogonal matrix $\mathbf{R} \in \mathbb{O}(d)$ such that

$$\text{(weighted error)} \quad \Delta_0(\mathbf{R}, k) \lesssim \epsilon^2 \cdot s^2 \ell_\star(\epsilon, s) \tag{I.4a}$$

$$\text{(unweighted error)} \quad \Delta_1(\mathbf{R}, k) \lesssim (\sqrt{r} + s^2)\epsilon + s\sigma_s(\mathbf{M}^\star) + \mathsf{tail}_1(\mathbf{M}^\star; s), \tag{I.4b}$$

where we define $\ell_\star(\epsilon, s) := \min\left\{1 + \log \frac{\|\mathbf{M}^\star\|_{\mathrm{op}}}{40s\epsilon},\ s\right\}$.

(b) Moreover, the index k satisfies

$$\mathsf{tail}_2(\mathbf{M}^\star; k) \lesssim s^3\epsilon^2 + s(\sigma_s(\mathbf{M}^\star))^2 + \mathsf{tail}_2(\mathbf{M}^\star; s).$$

(c) *The matrix* $\mathbf{R}$ *and* k *satisfy* $(\hat{\mathbf{A}}\mathbf{R})^\top \hat{\mathbf{A}}\mathbf{R} \succeq 39\epsilon \mathbf{P}^\star_{[k]}$ *and* $\sigma_k(\mathbf{M}^\star) - \sigma_{k+1}(\mathbf{M}^\star) \geq 40\epsilon/s$.

Proof [Proof of Theorem 7] Fix any $s \in \mathbb{N}$. To tune the bound, we also fix a parameter $\sigma \in [\sigma_s(\mathbf{M}^\star), \|\mathbf{M}^\star\|_{\mathrm{op}}]$. We shall tune σ at the end of the proof such that the following inequality is satisfied

$$\epsilon \leq \frac{\sigma}{40s}. \tag{I.12}$$

Extracting the balanced partition. Consider the balanced partition that arises from applying Proposition I.2 with parameter s and singular value parameter σ, and let k_ℓ be the resulting last pivot. Note that $\sigma^\star_{k_\ell} \geq \sigma$, $\delta^\star_{k_\ell} \geq 1/s$, and $k_\ell < s$; i.e. $k_\ell \in [s-1]$. We shall ultimately choose the promised k in the main theorem to be k_ℓ, but retain the ℓ-subscript for clarity in the proof below.

As a consequence of Eq. (I.12), we have

$$\epsilon \leq \frac{\sigma^\star_{k_\ell}}{40s}. \tag{I.13}$$

Note that Eq. (I.13) implies $k_\ell \leq r$, because

$$\sigma_{k_\ell}(\hat{\mathbf{M}}) \geq \sigma^\star_{k_\ell} - \|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{op}} \geq \sigma^\star_{k_\ell} - \epsilon > 0.$$

Moreover, we have

$$(\mathcal{K}_i)_{i=1}^{\ell} \text{ is } (\delta,\mu)\text{-well tempered for } \delta = 1/s, \quad \mu \leq 2e^2 = \mathcal{O}(1). \tag{I.14}$$

In particular,

$$\epsilon \leq \frac{\delta \sigma^\star_{k_\ell}}{40}. \tag{I.15}$$

We shall use Eq. (I.15) as the sufficient condition to invoke Proposition I.1. In addition, due to Weyl's inequality and Eq. (I.13),

$$\sigma_{k_\ell}(\hat{\mathbf{M}}) \geq \sigma^\star_{k_\ell} - \|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{op}} \geq \frac{39}{40}\sigma^\star_{k_\ell} \geq 39\epsilon > 0. \tag{I.16}$$

We shall use this lower bound to verify the positive semi-definite domination of $\mathbf{P}^\star_{[k]}$ at the end of the proof. Finally, we can also check that $\sigma^\star_{k_\ell} - \sigma^\star_{k_{\ell+1}} \geq 40\epsilon/s$ via similar manipulations.

Applying the error bounds. In addition, Eq. (I.15) allows us to apply Proposition I.1. This means that there exists an orthogonal matrix $\mathbf{R} \in \mathbb{O}(d)$ which is compatible with $(\mathcal{K}_i)_{i=1}^{\ell}$ such that the following holds

$$\begin{aligned}
\Delta_0(\mathbf{R}, k_\ell) &\lesssim \epsilon^2 \cdot \mu \cdot M_{\mathrm{space}} \lesssim \epsilon^2 s^2 \ell_{\sigma,s} && (\mu \lesssim 1,\ M_{\mathrm{space}} \lesssim s^2 \ell_{\sigma,s}) \\
\Delta_1(\mathbf{R}, k_\ell) &\lesssim \sqrt{r}\epsilon + \delta^2 M_{\mathrm{spec}} \epsilon^2 + \mathsf{tail}_1(\mathbf{M}^\star; k_\ell) \\
&\lesssim \sqrt{r}\epsilon + \frac{\epsilon^2 s^2}{\sigma} + s\sigma + \mathsf{tail}_1(\mathbf{M}^\star; s) \\
& \qquad (\delta \geq 1/s,\ M_{\mathrm{spec}} \lesssim 1/\sigma,\ \mathsf{tail}_1(\mathbf{M}^\star; s) \lesssim s\sigma + \mathsf{tail}_1(\mathbf{M}^\star; s)) \\
&\lesssim \sqrt{r}\epsilon + \frac{\sigma^2 s^2}{\sigma s^2} + s\sigma + \mathsf{tail}_1(\mathbf{M}^\star; s) && (\epsilon \leq \sigma/(40s) \text{ due to Eq. (I.12)}) \\
&\lesssim \sqrt{r}\epsilon + s\sigma + \mathsf{tail}_1(\mathbf{M}^\star; s).
\end{aligned}$$

Above, we used Proposition I.2 which affords $\mu \lesssim 1$, $M_{\mathrm{space}} \lesssim s^2\ell_{\sigma,s}$, $\delta \geq 1/s$, $M_{\mathrm{spec}} \lesssim 1/\sigma$, and $\mathsf{tail}_1(\mathbf{M}^\star; s) \lesssim s\sigma + \mathsf{tail}_1(\mathbf{M}^\star; s)$. To summarize,

$$\Delta_0(\mathbf{R}, k_\ell) \lesssim \epsilon^2 \cdot \mu \cdot M_{\mathrm{space}} \lesssim \epsilon^2 s^2 \ell_{\sigma,s}, \quad \Delta_1(\mathbf{R}, k_\ell) \leq \sqrt{r}\epsilon + s\sigma + \mathsf{tail}_1(\mathbf{M}^\star; s). \tag{I.18}$$

In addition, note from Proposition I.2 that

$$\mathsf{tail}_2(\mathbf{M}^\star; k_\ell) \lesssim s\sigma^2 + \mathsf{tail}_2(\mathbf{M}^\star; s).$$

Tuning parameter σ. We choose

$$\sigma = \max\{\sigma_s^\star, 40s\epsilon\}. \tag{I.19}$$

This ensures that two of our constraints on σ are satisfied: i.e. $\sigma \geq \sigma_s^\star$, and that $\epsilon \leq \frac{\sigma}{40s}$. For our third constraint, $\sigma \leq \sigma_1^\star$, to hold, this requires that $\epsilon \leq \sigma_1^\star/40s$, which is ensured by the condition of the theorem.

Applying our choice of σ to the error bounds. For this choice of σ, we have

$$\begin{aligned}\ell_{\sigma,s} &:= \min\left\{1 + \lceil \log \frac{\|\mathbf{M}^\star\|_{\mathrm{op}}}{\sigma} \rceil, s\right\} \\ &\leq \min\left\{1 + \lceil \log \underbrace{\frac{\|\mathbf{M}^\star\|_{\mathrm{op}}}{40s\epsilon}}_{\geq 1} \rceil, s\right\} \lesssim \min\left\{1 + \log \frac{\|\mathbf{M}^\star\|_{\mathrm{op}}}{40s\epsilon}, s\right\} := \ell_\star(\epsilon, s).\end{aligned}$$

Thus, by Eq. (I.18)

$$\Delta_0(\mathbf{R}, k_\ell) \lesssim \epsilon^2 \cdot s^2 \ell_\star(\epsilon, s).$$

Similarly, Eq. (I.18),

$$\begin{aligned}\Delta_1(\mathbf{R}, k_\ell) &\lesssim \sqrt{r}\epsilon + s\max\{\sigma_s^\star, 40s\epsilon\} + \mathsf{tail}_1(\mathbf{M}^\star; s) \\ &\lesssim (\sqrt{r} + s^2)\epsilon + s\sigma_s^\star + \mathsf{tail}_1(\mathbf{M}^\star; s).\end{aligned}$$

Finally, using $\sigma = \max\{\sigma_s^\star, 40s\epsilon\}$, we bound

$$\mathsf{tail}_2(\mathbf{M}^\star; k_\ell) \lesssim s\sigma^2 + \mathsf{tail}_2(\mathbf{M}^\star; s) \lesssim s^3\epsilon^2 + s(\sigma_s^\star)^2 + \mathsf{tail}_2(\mathbf{M}^\star; s).$$

We conclude by setting $k = k_\ell$.

Checking PSD domination. Lastly, we check the relevant PSD relation. Recall our choice $k = k_\ell$. Let $\mathcal{V}_k$ denote the range of the projection $\mathbf{P}^\star_{[k]}$, which is the span of the first k basis vectors under Eq. (I.6). Let $\mathbf{v} = (\mathbf{v}_1 + \mathbf{v}_2) \in \mathbb{R}^p$ be such that $\mathbf{v}_1 \in \mathcal{V}_k$, and $\mathbf{v}_2$ is supported on the remaining $p - k$ basis vectors. Then, since $\mathbf{R}$ is orthogonal and compatible with $(\mathcal{K}_i)_{i=1}^\ell$ and since $\bigcup_{i=1}^\ell \mathcal{K}_i = [k]$, we have

$$\mathbf{R} \cdot (\mathbf{v}_1 + \mathbf{v}_2) = (\mathbf{w}_1 + \mathbf{w}_2)$$

where $\|\mathbf{w}_1\| = \|\mathbf{v}_1\|$, $\|\mathbf{w}_2\| = \|\mathbf{v}_2\|$, and again $(\mathbf{w}_1, \mathbf{w}_2)$ decomposes into the first k and remaining $p-k$ coordinates. Using Eq. (I.6), moreover, $\hat{\mathbf{A}}^\top \hat{\mathbf{A}} = \hat{\mathbf{\Sigma}}$. Hence,

$$\begin{aligned}\mathbf{v}^\top \mathbf{R}^\top \hat{\mathbf{A}}^\top \hat{\mathbf{A}} \mathbf{R} \mathbf{v} &= (\mathbf{w}_1 + \mathbf{w}_2)^\top \hat{\mathbf{\Sigma}} (\mathbf{w}_1 + \mathbf{w}_2) \\ &\geq \sigma_k(\hat{\mathbf{\Sigma}}) \|\mathbf{w}_1\|^2 = \sigma_{k_\ell}(\hat{\mathbf{\Sigma}}) \|\mathbf{v}_1\|^2 \\ &= \sigma_k(\hat{\mathbf{\Sigma}}) \mathbf{v}_1^\top \mathbf{P}^\star_{[k]} \mathbf{v}_1 = \sigma_k(\hat{\mathbf{\Sigma}}) \mathbf{v}^\top \mathbf{P}^\star_{[k]} \mathbf{v},\end{aligned}$$

as $\mathbf{v}_1$ is the projection of $\mathbf{v}$ onto $\mathbf{P}^\star_{[k]}$. Lastly, as $\hat{\mathbf{A}}$ is balanced, $\sigma_k(\hat{\mathbf{\Sigma}}) = \sigma_k(\hat{\mathbf{M}}) := \sigma_{k_\ell}(\hat{\mathbf{M}}) \geq 39\epsilon$ due to Eq. (I.16). This completes the proof of our Theorem 7. ■

Appendix J. From Matrix Factorization to Bilinear Embeddings

This section gives the limiting arguments that proceed from results about matrices to results about Hilbert-space embeddings under potentially non-discrete distributions. Specifically, we prove the following

Theorem 8 (Error on $\mathcal{D}_{1\otimes 1}$) *Suppose $(\hat{f}, \hat{g})$ are $\mathbb{R}^r$-embeddings. Then, for any $s \in \mathbb{N}$ and error bound $\epsilon > 0$ such that (i) $\epsilon^2 \geq \inf_{s' \geq s-1} \mathcal{R}_{[s']}(\hat{f}, \hat{g}; \mathcal{D}_{1\otimes 1})$ and (ii) $s < \frac{\|\mathbf{\Sigma}^\star_{1\otimes 1}\|_{\mathrm{op}}}{40\epsilon}$, then we have: (a) if $(\hat{f}, \hat{g})$ are* full-rank, *then there exists an index $k \in [\min\{r, s-1\}]$ and functions $f : \mathcal{X} \to \mathcal{H}$ and $g : \mathcal{Y} \to \mathcal{H}$ such that (f, g) are aligned k-proxies and the error terms are bounded by*

$$(\text{weighted error}) \quad \mathbf{\Delta}_0(f, g, k) + \mathsf{tail}^\star_2(k) \lesssim s^3 \epsilon^2 + s(\boldsymbol{\sigma}^\star_s)^2 + \mathsf{tail}^\star_2(s), \tag{J.1a}$$

$$(\text{unweighted error}) \quad \mathbf{\Delta}_1(f, g, k) \lesssim (\sqrt{r} + s^2)\epsilon + s\boldsymbol{\sigma}^\star_s + \mathsf{tail}^\star_1(s); \tag{J.1b}$$

and (b) if $\epsilon^2_{1\otimes 1} \leq (1 - \alpha^{-1})(\boldsymbol{\sigma}^\star_r)^2$ for some $\alpha \geq 1$, then $(\hat{f}, \hat{g})$ are necessarily full-rank, and $\sigma_r(\hat{f}, \hat{g})^2 \geq (\boldsymbol{\sigma}^\star_r)^2/\alpha$, where we recall the definition of $\sigma_r(\hat{f}, \hat{g})$ in Eq. (3.1).

Remark 9 *A few remarks are in order. The condition $\epsilon^2 \geq \inf_{s' \geq s-1} \mathcal{R}_{[s']}(\hat{f}, \hat{g}; \mathcal{D}_{1\otimes 1})$ is for technical convenience; for intuition, one should think of $\epsilon^2 = \mathcal{R}(\hat{f}, \hat{g}; \mathcal{D}_{1\otimes 1})$ as the risk on the "top-block". Next, we observe the differences in scaling: due to the weighting, $\mathbf{\Delta}_0(f, g, k) + \mathsf{tail}^\star_2(k)$ scales with ϵ^2, and with the* squares *of singular values, whereas $\mathbf{\Delta}_1(f, g, k)$ scales with ϵ, and ℓ_1-sums of singular values. This is essential, because it means that the term $\frac{1}{\sigma^2}(\mathsf{tail}^\star_2(k) + \mathbf{\Delta}_0(f, g, k) + \mathbf{\Delta}_{\mathrm{train}})^2$ in Proposition 4.1 can decay to zero. Lastly, our theorem gives us sufficient conditions on which $\sigma_r(\hat{f}, \hat{g})$, which appears in the aforementioned Proposition 4.1, is indeed lower bounded.*

We begin in Appendix J.1 by stating an intermediate guarantee for Theorem 8, Theorem 10, to whose proof the majority of this appendix is devoted, and provide preliminaries and review proof-specific notation in Appendix J.2.

To prove Theorem 10, we adopt the standard technique of approximation by so-called *simple functions*:

Definition J.1 (Simple Functions) *Let $\mathcal{Z}$ be an abstract domain. We say that a function $\psi : \mathcal{Z} \to \mathbb{R}^p$ is simple if its image $\psi(\mathcal{Z})$ is a set of finite cardinality.*

In Appendix J.3, we show that our factorization theorem for matrices (i.e. Theorem 7) directly implies Theorem 10. Subsequently, Appendix J.4 extends the guarantees to arbitrary (possible non-simple) functions, but with the restriction that they have a finite-dimensional range. The idea is to approximate our actual functions $f, f^\star, g, g^\star$ as the limit of simple functions. The steps in this section are mostly routine, but some care must be taken to ensure all the simple functions can be *balanced* under $\mathcal{D}_{1\otimes 1}$ in the sense of Definition 4.1; recall that (f, g) are balanced (under $\mathcal{D}_{1\otimes 1}$) if

$$\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[ff^\top] = \mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}[gg^\top].$$

To facilitate this, we show that when all the embeddings have we can approximate $f, f^\star, g, g^\star$ by simple functions whose range is "smaller" than the limiting function they approximate. Care must also be taken to handle the rotation matrices which align the functions $(f^\star, g^\star)$ with their estimates (f, g).

Finally, Appendix J.5 removes the restriction of a finite dimensional range, thereby concluding the proof of Theorem 10. Subsubsection J.6.1 contains the proof of all supporting claims. Lastly, Theorem 11 provides the generalization of our main SVD perturbation lemma, Theorem 1, to general distributions.

J.1. Factor recovery for one block, Theorem 8

We first prove a variant of Theorem 8, from which that theorem can be readily derived.

Theorem 10 *Suppose that $(\hat{f}, \hat{g})$ embeddings $\mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle \hat{f}, \hat{g}\rangle - \langle f^\star, g^\star\rangle)^2] \le \epsilon^2$, and pick any positive $s \in \mathbb{N}$ and $s > 1$ such that $s < \frac{\|\mathbf{\Sigma}^\star_{1\otimes 1}\|_{\mathrm{op}}}{40\epsilon}$. Then, $(\hat{f}, \hat{g})$ are full-rank, there exists a $k \in [s-1]$ and functions $f : \mathcal{X} \to \mathcal{H}$ and $g : \mathcal{Y} \to \mathcal{H}$ such that*

(a) $\langle f(x), g(y)\rangle = \langle \hat{f}(x), \hat{g}(y)\rangle$ *for all* (x, y).

(b) The functions (f, g) are valid proxies for $\hat{f}, \hat{g}$ in the sense of Definition 4.2.

(c) The following error terms

$$\begin{aligned}\Delta_0(f, g, k) &:= \max\left\{\mathbb{E}_{\mathcal{D}_{1\otimes 1}}\left[\langle f^\star_k, g^\star_k - g\rangle^2\right], \mathbb{E}_{\mathcal{D}_{1\otimes 1}}\left[\langle f^\star_k - f, g^\star_k\rangle^2\right]\right\}\\ \Delta_1(f, g, k) &:= \max\left\{\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}\|f^\star_k - f\|^2, \mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}\|g^\star_k - g\|^2\right\}\end{aligned}$$

are bounded by

$$\begin{aligned}\Delta_0(f, g, k) + \mathsf{tail}_2(\mathbf{\Sigma}^\star_{1\otimes 1}; k) &\lesssim s^3\epsilon^2 + s(\boldsymbol{\sigma}^\star_s)^2 + \mathsf{tail}_2(\mathbf{\Sigma}^\star_{1\otimes 1}; s)\\ \Delta_1(f, g, k) &\lesssim (\sqrt{r} + s^2)\epsilon + s\boldsymbol{\sigma}^\star_s + \mathsf{tail}_1(\mathbf{\Sigma}^\star_{1\otimes 1}; s),\end{aligned}$$

(d) For any j, $\sigma_j(\hat{f}, \hat{g}) \ge \sigma_j(\mathbf{\Sigma}^\star) - \epsilon$.

Moreover, if instead of assuming $(\hat{f}, \hat{g})$ are full-rank, but in addition we assume that $\epsilon < \sigma_r(\mathbf{\Sigma}^\star_{1\otimes 1})$, then $(\hat{f}, \hat{g})$ are guaranteed to be full-rank so that the conclusion of the above theorem holds.

Let us now prove Theorem 8.

Proof [Proof of Theorem 8] Fix any s. Consider any $s' \geq s-1$, and define $\bar{f}^\star := f^\star_{s'}$ and $\bar{g}^\star := g^\star_{s'}$. Define $\bar{\Delta}_0(f,g,k)$, $\bar{\Delta}_1(f,g,k)$ and $\bar{\boldsymbol{\Sigma}}_{1\otimes 1}$analogously, with $f^\star, g^\star$ replaced by $\bar{f}^\star, \bar{g}^\star$. Finally, set $\epsilon^2 := \mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle f, g\rangle - \langle \bar{f}^\star, \bar{g}^\star\rangle)^2]$. Applying Theorem 10 with $f^\star, g^\star \leftarrow \bar{f}^\star, \bar{g}^\star$, we find the existence of f, g satisfying points (a), (b), as well as well as

$$
\begin{aligned}
\bar{\Delta}_1(f,g,k) &\lesssim (\sqrt{r} + s^2)\epsilon + s\sigma_s(\bar{\boldsymbol{\Sigma}}_{1\otimes 1}) + \mathsf{tail}_1(\bar{\boldsymbol{\Sigma}}_{1\otimes 1}; s) \\
\bar{\Delta}_0(f,g,k) + \mathsf{tail}_2(\bar{\boldsymbol{\Sigma}}_{1\otimes 1}; k) &\lesssim s^3\epsilon^2 + s\sigma_s(\bar{\boldsymbol{\Sigma}}_{1\otimes 1})^2 + \mathsf{tail}_2(\bar{\boldsymbol{\Sigma}}_{1\otimes 1}; s),
\end{aligned}
$$

where above we bounded $\ell_\star(\epsilon, s) \leq s$. Observe that $k \in [s-1]$, it holds that $\bar{f}^\star_k = f^\star_k$ and $\bar{g}^\star_k = g^\star_k$. Moreover, since $\bar{\boldsymbol{\Sigma}}_{1\otimes 1} \preceq \boldsymbol{\Sigma}^\star_{1\otimes 1}$ (since the former is an SVD approximation of the latter), and since $\sigma_s(\bar{\boldsymbol{\Sigma}}_{1\otimes 1})^2 \leq \sigma_s(\boldsymbol{\Sigma}^\star_{1\otimes 1})^2$,

$$
\begin{aligned}
\bar{\Delta}_1(f,g,k) &\lesssim (\sqrt{r} + s^2)\epsilon + s\boldsymbol{\sigma}^\star_s + \mathsf{tail}_1(\boldsymbol{\Sigma}^\star_{1\otimes 1}; s) \\
\bar{\Delta}_0(f,g,k) + \mathsf{tail}_2(\bar{\boldsymbol{\Sigma}}_{1\otimes 1}; k) &\lesssim s^3\epsilon^2 + s(\boldsymbol{\sigma}^\star_s)^2 + \mathsf{tail}_2(\boldsymbol{\Sigma}^\star_{1\otimes 1}; s) + \epsilon^2 \cdot s^3.
\end{aligned}
$$

Lastly, notice that

$$
\begin{aligned}
\mathsf{tail}_2(\boldsymbol{\Sigma}^\star_{1\otimes 1}; k) &= \mathsf{tail}_2(\bar{\boldsymbol{\Sigma}}_{1\otimes 1}; k) + \sum_{i > s'} (\boldsymbol{\sigma}^\star_i)^2 \\
&\leq \mathsf{tail}_2(\bar{\boldsymbol{\Sigma}}_{1\otimes 1}; k) + (\boldsymbol{\sigma}^\star_s)^2 + \mathsf{tail}_2(\boldsymbol{\Sigma}^\star_{1\otimes 1}; s) \\
&\lesssim s^3\epsilon^2 + s(\boldsymbol{\sigma}^\star_s)^2 + \mathsf{tail}_2(\boldsymbol{\Sigma}^\star_{1\otimes 1}; s),
\end{aligned}
$$

where above we use $s' \geq s-1$. Hence, these differences get absorbed by the above bound on $\mathsf{tail}_2(\bar{\boldsymbol{\Sigma}}^\star; k)$, yielding

$$
\Delta_0(f,g,k) + \mathsf{tail}_2(\bar{\boldsymbol{\Sigma}}^\star; k) \lesssim s^3\epsilon^2 + s(\boldsymbol{\sigma}^\star_s)^2 + \mathsf{tail}_2(\boldsymbol{\Sigma}^\star_{1\otimes 1}; s).
$$

Since the above was true for any $s' \geq s-1$, we can replace ϵ with any ϵ satisfying

$$
\epsilon^2 \geq \inf_{s' \geq s-1} \mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle f, g\rangle - \langle f^\star_{s'}, g^\star_{s'}\rangle)^2],
$$

as needed. Finally, the last part of Theorem 8 is directly implied by Theorem 10(d). ■

J.2. Proof preliminaries

For the majority of the proof, we assume that $\mathcal{H} = \mathbb{R}^p$; that is, the embeddings are finite dimensional (recall that all finite dimensional Hilbert spaces are isomorphic). This restriction is the simplest to remove, so we save removing it till the end of the argument. We also study balanced functions f, g directly, and remove the balancing requirement at the end.

Setup. Let $\mathcal{D}_\mathcal{X}$ and $\mathcal{D}_\mathcal{Y}$ be distributions over $\mathcal{X}$ and $\mathcal{Y}$ which have finite support, and let $\mathcal{D}_\otimes := \mathcal{D}_\mathcal{X} \otimes \mathcal{D}_\mathcal{Y}$ denote the product measure. We consider functions $f, f^\star : \mathcal{X} \to \mathbb{R}^p$ and $g, g^\star : \mathcal{Y} \to \mathbb{R}^p$ whose inner products have squared error $\epsilon^2_{\mathrm{pred}}$:

$$
\epsilon^2_{\mathrm{pred}} = \mathbb{E}_{\mathcal{D}_\otimes}[(\langle f, g\rangle - \langle f^\star, g^\star\rangle)^2], \quad \mathcal{D}_\otimes := \mathcal{D}_\mathcal{X} \otimes \mathcal{D}_\mathcal{Y}.
$$

Key objects. When reasoning about functions of random variables, we no longer have finite matrices whose singular values we can reason about. Instead, it is more convenient to describe spectral via expected outer-products. The following objects are central to our consideration:

$$\begin{aligned}
\mathbf{\Sigma} &:= \mathbb{E}_{\mathcal{D}_{\mathcal{X}}}[ff^\top] = \mathbb{E}_{\mathcal{D}_{\mathcal{Y}}}[gg^\top] \\
\mathbf{\Sigma}^\star &:= \mathbb{E}_{\mathcal{D}_{\mathcal{X}}}[(f^\star)(f^\star)^\top] = \mathbb{E}_{\mathcal{D}_{\mathcal{Y}}}[(g^\star)(g^\star)^\top] \\
\mathsf{tail}_q(\mathbf{\Sigma}; k) &:= \sum_{i>k} \sigma_i(\mathbf{\Sigma})^q, \quad q \ge 1, \;\; \mathbf{\Sigma} \in \mathbb{R}^{p\times p} \\
f_k^\star &:= \mathbf{P}_k^\star f, \quad g_k^\star := \mathbf{P}_k^\star g \\
f_k &:= \mathbf{P}_k f, \quad g_k := \mathbf{P}_k g,
\end{aligned}$$

where $\mathbf{P}_k^\star$ is the projection onto any top-k eigenspace of $\mathbf{\Sigma}^\star$ (unique when $\sigma_k(\mathbf{\Sigma}^\star) > \sigma_{k+1}(\mathbf{\Sigma}^\star)$), and $\hat{\mathbf{P}}_k$ is the projection onto any top-k eigenspace of $\mathbf{\Sigma}$.

We consider the following error terms:

$$\begin{aligned}
\Delta_0(\mathbf{R}, k) &= \mathbb{E}_{\mathcal{D}_\otimes}[\langle f_k^\star - \mathbf{R}f, g_k^\star\rangle^2] \vee \mathbb{E}_{\mathcal{D}_\otimes}[\langle f_k^\star, \mathbf{R}g - g_k^\star\rangle^2] \\
\Delta_1(\mathbf{R}, k) &= \mathbb{E}_{\mathcal{D}_{\mathcal{X}}}[\|f_k^\star - \mathbf{R}f\|^2] \vee \mathbb{E}_{\mathcal{D}_{\mathcal{Y}}}[\|\mathbf{R}g - g_k^\star\|^2] \\
\epsilon^2 &\ge \mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle f, g\rangle - \langle f^\star, g^\star\rangle)^2] \\
\epsilon_{\mathrm{pred},k}^2 &:= \mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle f_k, g_k\rangle - \langle f_k^\star, g_k^\star\rangle)^2].
\end{aligned} \tag{J.2}$$

Outer product notation. To reduce notational clutter, we introduce a compact notation for vector outer products. Given a vector $v \in \mathbb{R}^p$, or more generally, functions $f : \mathcal{X} \to \mathbb{R}^p$ and $g : \mathcal{Y} \to \mathbb{R}^p$, we let $v^{\otimes 2} := vv^\top$, $f^{\otimes 2} := ff^\top$, $g^{\otimes 2} := gg^\top$. Notice that the typesetting of $\otimes 2$ differs from the standard tensor product $\otimes$ so as to avoid confusion with tensor-products of distributions, as in $\mathcal{D}_{\mathcal{X}} \otimes \mathcal{D}_{\mathcal{Y}}$.

J.3. Guarantee for simple functions

For simple functions, Theorem 10 items (a)-(d) translate to the following guarantees.

Proposition J.1 *Suppose that (f, g) and $(f^\star, g^\star)$ are simple functions, and balanced under $\mathcal{D}_\otimes = \mathcal{D}_{\mathcal{X}} \otimes \mathcal{D}_{\mathcal{Y}}$. Further, suppose ϵ as in Eq. (J.2) and $s \in \mathbb{N}$ satisfies $\epsilon \le \frac{\|\mathbf{\Sigma}^\star\|_{\mathrm{op}}}{40s}$. Then, there exists an index $k \in [s-1]$ and an orthogonal matrix $\mathbf{R} \in \mathbb{O}(p)$ such that*

$$\Delta_0(\mathbf{R}, k) \lesssim \epsilon^2 \cdot s^2 \ell_\star(\epsilon, s) \tag{J.3a}$$
$$\Delta_1(\mathbf{R}, k) \lesssim (\sqrt{r} + s^2)\epsilon + s\sigma_s(\mathbf{\Sigma}^\star) + \mathsf{tail}_1(\mathbf{\Sigma}^\star; s), \tag{J.3b}$$

where we define $\ell_\star(\epsilon, s) := \min\left\{1 + \log \frac{\|\mathbf{\Sigma}^\star\|_{\mathrm{op}}}{40s\epsilon},\, s\right\}$. Second, the index k satisfies

$$\mathsf{tail}_2(\mathbf{\Sigma}^\star; k) \lesssim s^3\epsilon^2 + s(\sigma_s(\mathbf{\Sigma}^\star))^2 + \mathsf{tail}_2(\mathbf{\Sigma}^\star; s).$$

Third, $\mathbf{R}$ and k satisfy

$$\mathbf{R}\mathbf{\Sigma}\mathbf{R}^\top \succeq \epsilon\mathbf{P}_k^\star, \quad \sigma_k(\mathbf{\Sigma}^\star) - \sigma_{k+1}(\mathbf{\Sigma}^\star) \ge 40\epsilon/s,$$

and lastly $\max_j |\sigma_j(\mathbf{\Sigma}^\star) - \sigma_j(\mathbf{\Sigma})| \le \epsilon$.

The key property of simple functions we use is that their expectations can be reduced to those over finitely-supported distributions. The following is proved in Subsubsection J.6.1.

Lemma J.1 *Let $f_1, \dots, f_a : \mathcal{X} \to \mathbb{R}^p$ and $g_1, \dots, g_b : \mathcal{Y} \to \mathbb{R}^p$ be simple functions, and let $\mathcal{D}_{\mathcal{X}}$ and $\mathcal{D}_{\mathcal{Y}}$ be measures over $\mathcal{X}$ and $\mathcal{Y}$, respectively. Then, there exist finitely-supported distributions $\bar{\mathcal{D}}_{\mathcal{X}}$ and $\bar{\mathcal{D}}_{\mathcal{Y}}$ such that, for all functions $\Psi : \mathbb{R}^{p(a+b)} \to \mathcal{V}_\Psi$ mapping to some Euclidean space $\mathcal{V}_\Psi$ (possibly different for each Ψ), we have*

$$\mathbb{E}_{\mathcal{D}_{\mathcal{X}} \otimes \mathcal{D}_{\mathcal{Y}}}[\Psi(f_1(x), \dots, f_a(x), g_1(y), \dots, g_b(y))] = \mathbb{E}_{\bar{\mathcal{D}}_{\mathcal{X}} \otimes \bar{\mathcal{D}}_{\mathcal{Y}}}[\Psi(f_1(x), \dots, f_a(x), g_1(y), \dots, g_b(y))].$$

We now turn to the proof of Proposition J.1.

Proof [Proof of Proposition J.1]

By Lemma J.1, we may assume without loss of generality that $\mathcal{D}_{\mathcal{X}}$ and $\mathcal{D}_{\mathcal{Y}}$ are distributions with finite support; indeed, by appropriate choices of Ψ, the discretization preserves expected outer-products (e.g. $\mathbf{\Sigma}^\star$), balancing, the projection $\mathbf{P}_k^\star$, and Δ_0, Δ_1.

Continuing, assume $n = |\mathrm{supp}(\mathcal{D}_{\mathcal{X}})|$ and $m = |\mathrm{supp}(\mathcal{D}_{\mathcal{Y}})|$. By augumenting the support with probability-zero points, we may assume without loss of generality that $p \le \min\{n, m\}$. Let $x_1, \dots, x_n$ and $y_1, \dots, y_m$ denote the elements of $\mathrm{supp}(\mathcal{D}_{\mathcal{X}})$ and $\mathrm{supp}(\mathcal{D}_{\mathcal{Y}})$. For $i \in [n]$ and $j \in [m]$, define $\mathsf{p}_i := \mathbb{P}_{x \sim \mathcal{D}_{\mathcal{X}}}[x = x_i]$ and $\mathsf{q}_j = \mathbb{P}_{y \sim \mathcal{D}_{\mathcal{Y}}}[y = y_j]$. We define the matrices $\mathbf{M}^\star, \hat{\mathbf{M}} \in \mathbb{R}^{n \times m}$ via

$$\mathbf{M}^\star_{ij} = \sqrt{\mathsf{p}_i \mathsf{q}_j} \cdot \langle f^\star(x_i), g^\star(y_j) \rangle, \quad \hat{\mathbf{M}}_{ij} = \sqrt{\mathsf{p}_i \mathsf{q}_j} \cdot \langle f(x_i), g(y_j) \rangle.$$

Further, define matrices $\mathbf{A}^\star, \hat{\mathbf{A}} \in \mathbb{R}^{n \times p}$ and $\mathbf{B}^\star, \hat{\mathbf{B}} \in \mathbb{R}^{m \times p}$ via their rows:

$$\mathbf{A}^\star_{(i,:)} = \sqrt{\mathsf{p}_i} f^\star(x_i)^\top, \quad \mathbf{B}^\star_{(j,:)} = \sqrt{\mathsf{q}_j} g^\star(y_j)^\top, \quad \hat{\mathbf{A}}_{(i,:)} = \sqrt{\mathsf{p}_i} f(x_i)^\top, \quad \hat{\mathbf{B}}_{(j,:)} = \sqrt{\mathsf{q}_j} g(y_j)^\top.$$

We readily check that

$$\mathbf{M}^\star = \mathbf{A}^\star (\mathbf{B}^\star)^\top \quad \hat{\mathbf{M}} = \hat{\mathbf{A}} \hat{\mathbf{B}}^\top.$$

Proposition J.1 follows directly from Theorem 7, after invoking the substitutions invoked by the following lemma (and taking $\mathbf{R} \leftarrow \mathbf{R}^\top$):

Lemma J.2 *The following identities hold.*

(a) $\sigma_i(\mathbf{M}^\star) = \sigma_i(\mathbf{\Sigma}^\star)$ *and* $\sigma_i(\hat{\mathbf{M}}) = \sigma_i(\mathbf{\Sigma})$.

(b) $\|\mathbf{M}^\star - \hat{\mathbf{M}}\|_{\mathrm{F}}^2 = \mathbb{E}_{\mathcal{D}_{\mathcal{X}} \otimes \mathcal{D}_{\mathcal{Y}}}[(\langle f^\star(x), g^\star(y) \rangle - \langle f(x), g(y) \rangle)^2] := \epsilon_{\mathrm{pred}}^2 \le \epsilon^2$. *Consequently, by Weyl's inequality,* $|\sigma_i(\mathbf{\Sigma}^\star) - \sigma_i(\mathbf{\Sigma})| \le \|\mathbf{M}^\star - \hat{\mathbf{M}}\|_{\mathrm{op}} \le \|\mathbf{M}^\star - \hat{\mathbf{M}}\|_{\mathrm{F}} \le \epsilon$.

(c) $(\mathbf{A}^\star)^\top \mathbf{A}^\star = \mathbb{E}_{\mathcal{D}_{\mathcal{X}}}[(f^\star)(f^\star)^\top]$, $(\mathbf{B}^\star)^\top \mathbf{B}^\star = \mathbb{E}_{\mathcal{D}_{\mathcal{Y}}}[(g^\star)(g^\star)^\top]$, *so that* $(\mathbf{A}^\star)^\top \mathbf{A}^\star = (\mathbf{B}^\star)^\top \mathbf{B}^\star$. *Similarly,* $\hat{\mathbf{A}}^\top \hat{\mathbf{A}} = \mathbb{E}_{\mathcal{D}_{\mathcal{X}}}[f f^\top]$ *and* $\hat{\mathbf{B}}^\top \hat{\mathbf{B}} = \mathbb{E}_{\mathcal{D}_{\mathcal{Y}}}[g g^\top]$, *so that* $\hat{\mathbf{A}}^\top \hat{\mathbf{A}} = \hat{\mathbf{B}}^\top \hat{\mathbf{B}}$.

(d) Using SVD approximations in the sense of Eq. (I.3), we have that $\mathbf{A}_k^\star$*'s i-th row is* $\sqrt{\mathsf{p}_i} \cdot f_k^\star(x_i)^\top$ *and* $\mathbf{B}_k^\star$*'s j-th row is* $\sqrt{\mathsf{q}_j} \cdot g_k^\star(y_j)^\top$ *(notice, the k is in the subscript). Similarly, we have that* $\hat{\mathbf{A}}_k$*'s i-th row is* $\sqrt{\mathsf{p}_i} \cdot f_k(x_i)^\top$ *and* $\hat{\mathbf{B}}_k$*'s j-th row is* $\sqrt{\mathsf{q}_j} \cdot g_k(y_j)^\top$.

(e) $\Delta_0(\mathbf{R}, k) = \|(\mathbf{A}^\star_{[k]} - \hat{\mathbf{A}} \mathbf{R}^\top)(\mathbf{B}^\star_{[k]})^\top\|_{\mathrm{F}}^2 \vee \|\mathbf{A}^\star_{[k]}(\mathbf{B}^\star_{[k]} - \hat{\mathbf{B}} \mathbf{R}^\top)^\top\|_{\mathrm{F}}^2$.

(f) $\Delta_1[\mathbf{R}, k] = \|\mathbf{A}^\star_{[k]} - \hat{\mathbf{A}}\mathbf{R}^\top\|_{\mathrm{F}}^2 \vee \|\mathbf{B}^\star_{[k]} - \hat{\mathbf{B}}\mathbf{R}^\top\|_{\mathrm{F}}^2$.

Proof The proof of point (a) relies on point (c), namely $(\mathbf{A}^\star)^\top\mathbf{A}^\star = (\mathbf{B}^\star)^\top\mathbf{B}^\star$ (the argument is not circular, because the proof of point (c) does not rely on point (a)). Using this, we see $(\mathbf{A}^\star)^\top\mathbf{A}^\star = (\mathbf{B}^\star)^\top\mathbf{B}^\star$. Thus, from Lemma I.2, $\sigma_i(\mathbf{M}^\star) = \sigma_i((\mathbf{A}^\star)^\top\mathbf{A}^\star)$. Invoking point (c) again, we find $\sigma_i(\mathbf{M}^\star) = \sigma_i(\mathbb{E}_{\mathcal{D}_\mathcal{X}}[(f^\star)^{\otimes 2}]) := \sigma_i(\mathbf{\Sigma}^\star)$. A similar argument applies to showing $\sigma_i(\hat{\mathbf{M}}) = \sigma_i(\mathbf{\Sigma})$.

The proof of points (b)-(f) rely on the same sorts of computations. We prove point (b) as an illustration.

$$\begin{aligned}
\|\mathbf{M}^\star - \hat{\mathbf{M}}\|_{\mathrm{F}}^2 &= \sum_{ij}(\sqrt{\mathsf{p}_i\mathsf{q}_j}\cdot\langle f^\star(x_i), g^\star(y_j)\rangle - \sqrt{\mathsf{p}_i\mathsf{q}_j}\langle f(x_i), g(y_j)\rangle)^2 \\
&= \sum_{ij}\mathsf{p}_i\mathsf{q}_j(\langle f^\star(x_i), g^\star(y_j)\rangle - \langle f(x_i), g(y_j)\rangle)^2 \\
&= \mathbb{E}_{\mathcal{D}_\mathcal{X}\otimes\mathcal{D}_\mathcal{Y}}(\langle f^\star(x), g^\star(y)\rangle - \langle f(x), g(y)\rangle)^2.
\end{aligned}$$

The remaining points can be proved analogously. ■

This concludes the proof of Proposition J.1. ■

J.4. Extension beyond simple functions

We now extend the guarantees of the previous section to the case beyond simple functions. The analogue of Proposition J.1 is as follows:

Proposition J.2 *Suppose that $f, g, f^\star, g^\star$ map to $\mathbb{R}^p$, and are balanced under $\mathcal{D}_\otimes = \mathcal{D}_\mathcal{X}\otimes\mathcal{D}_\mathcal{Y}$,* but are not necessarily simple functions. *Further, suppose ϵ as in Eq. (J.2) and $s \in \mathbb{N}$ satisfies*

$$\epsilon < \frac{\|\mathbf{\Sigma}^\star\|_{\mathrm{op}}}{40s} \quad \textit{(strict inequality).} \tag{J.4}$$

Then, there exists an index $k \in [s-1]$ and an orthogonal matrix $\mathbf{R} \in \mathbb{O}(p)$ such that

$$\Delta_0(\mathbf{R}, k) \lesssim \epsilon^2 \cdot s^3 \tag{J.5a}$$

$$\Delta_1(\mathbf{R}, k) \lesssim (\sqrt{r} + s^2)\epsilon + s\sigma_s(\mathbf{\Sigma}^\star) + \mathsf{tail}_1(\mathbf{\Sigma}^\star; s). \tag{J.5b}$$

Second, the index k satisfies

$$\mathsf{tail}_2(\mathbf{\Sigma}^\star; k) \lesssim s^3\epsilon^2 + s(\sigma_s(\mathbf{\Sigma}^\star))^2 + \mathsf{tail}_2(\mathbf{\Sigma}^\star; s).$$

Third, $\mathbf{R}$ and k satisfy

$$\mathbf{R}^\top\mathbf{\Sigma}\mathbf{R} \succeq \epsilon\mathbf{P}^\star_k, \quad \mathbf{\Sigma} := \mathbb{E}_{\mathcal{D}_{1\otimes 1}}[ff^\top],$$

and lastly $\max_{j\in[p]}|\sigma_j(\mathbf{\Sigma}^\star) - \sigma_j(\mathbf{\Sigma})| \le \epsilon$.

Before proving the above two propositions, we review some facts about $\mathcal{L}_2$ convergence, and some basic results for measure-theoretic probability theory which can be found in any standard reference (e.g. Çinlar (2011)).

$\mathcal{L}_2$ **convergence.** We first review the definition of $\mathcal{L}_2$ convergence.

Definition J.2 **($\mathcal{L}_2$ Convergence)** *Let $\mathcal{D}$ be a measure on $\mathcal{Z}$. We say that $\psi : \mathcal{Z} \to \mathbb{R}^p$ is in $\mathcal{L}_2(\mathcal{D})$ if $\mathbb{E}_{\mathcal{D}}[\|\psi\|^2] < \infty$. Let $(\psi_\tau)_{\tau \geq 1}$ be a sequence of functions in $\mathcal{L}_2(\mathcal{D})$, $\psi \in \mathcal{L}_2(\mathcal{D})$, and let $\mathcal{D}$ be a measure on $\mathcal{Z}$. We say that ψ_τ converges to ψ in $\mathcal{L}_2(\mathcal{D})$, denoted*

$$\psi_\tau \overset{\mathcal{L}_2(\mathcal{D})}{\to} \psi,$$

if $\lim_{\tau \to \infty} \mathbb{E}_{\mathcal{D}}\|\psi_\tau - \psi\|^2 = 0$.

The following lemma is standard in probability theory (again, see e.g., (Çinlar, 2011, Section 2)).

Lemma J.3 *Let $\mathcal{D}$ be a measure on $\mathcal{Z}$. Given any $\psi \in \mathcal{L}_2(\mathcal{D})$, there exists a sequence of simple functions $\psi_\tau \in \mathcal{L}_2(\mathcal{D})$ such that $\psi_\tau \overset{\mathcal{L}_2(\mathcal{D})}{\to} \psi$.*

We shall often use the following lemma, which is easy to check.

Lemma J.4 *If $\psi_\tau \overset{\mathcal{L}_2(\mathcal{D})}{\to} \psi$, then $\lim_{\tau \to \infty} \mathbb{E}[\psi_\tau^{\otimes 2}] = \mathbb{E}[\psi^{\otimes 2}]$.*

The following fact is also useful.

Lemma J.5 *If $\psi_\tau \overset{\mathcal{L}_2(\mathcal{D})}{\to} \psi$ and if $\mathrm{range}(\mathbb{E}[\psi_\tau^{\otimes 2}]) \subseteq \mathrm{range}(\mathbb{E}[\psi^{\otimes 2}])$ for all τ, then there exists some τ_0 such that, for all $\tau \geq \tau_0$, $\mathrm{range}(\mathbb{E}[\psi_\tau^{\otimes 2}]) = \mathrm{range}(\mathbb{E}[\psi^{\otimes 2}])$.*

Proofs of Lemmas J.4 and J.5 are given in Subsubsection J.6.2.

Approximation by simple functions. Using the machinery introduced above, we approximate $f, g, f^\star, g^\star$ by a sequence of simple functions. Our approximation preserves an important property regarding the ranges of their covariances.

Lemma J.6 *There exists a sequence of simple functions $f_{(\tau)}, g_{(\tau)}$ $f^\star_{(\tau)}, g^\star_{(\tau)}$ such that*

$$f_{(\tau)} \overset{\mathcal{L}_2(\mathcal{D}_\mathcal{X})}{\to} f, \quad g_{(\tau)} \overset{\mathcal{L}_2(\mathcal{D}_\mathcal{Y})}{\to} g, \quad f^\star_{(\tau)} \overset{\mathcal{L}_2(\mathcal{D}_\mathcal{X})}{\to} f^\star, \quad g^\star_{(\tau)} \overset{\mathcal{L}_2(\mathcal{D}_\mathcal{Y})}{\to} g^\star,$$

and, the covariances

$$\mathbf{\Sigma}_{(\tau),f} := \mathbb{E}[(f_{(\tau)})^{\otimes 2}], \quad \mathbf{\Sigma}_{(\tau),g} := \mathbb{E}[(g_{(\tau)})^{\otimes 2}], \quad \mathbf{\Sigma}^\star_{(\tau),f} := \mathbb{E}[(f^\star_{(\tau)})^{\otimes 2}], \quad \mathbf{\Sigma}^\star_{(\tau),g} := \mathbb{E}[(g^\star_{(\tau)})^{\otimes 2}],$$

satisfy $\mathrm{range}(\mathbf{\Sigma}_{(\tau),f}) \cup \mathrm{range}(\mathbf{\Sigma}_{(\tau),g}) \subseteq \mathrm{range}(\mathbf{\Sigma})$ *and* $\mathrm{range}(\mathbf{\Sigma}^\star_{(\tau),f}) \cup \mathrm{range}(\mathbf{\Sigma}^\star_{(\tau),g}) \subseteq \mathrm{range}(\mathbf{\Sigma}^\star)$.

The above lemma is proved in Subsubsection J.6.3.

Constructing the balanced functions. We cannot invoke Proposition J.1 directly on the simple functions constructed above because they are not *balanced*. Below we show that we can balance them, and that the matrices which achieve this converge to the identity.

Lemma J.7 *There exists a sequence of invertible $p \times p$-matrices $(\mathbf{T}_{[\tau]})_{\tau \geq 1}$ and $(\mathbf{T}^\star_{[\tau]})_{\tau \geq 1}$ such that*

(a) For all τ sufficiently large, $\mathbb{E}_{\mathcal{D}_\mathcal{X}}[(\mathbf{T}_{[\tau]} f_{(\tau)})^{\otimes 2})] = \mathbb{E}_{\mathcal{D}_\mathcal{Y}}[(\mathbf{T}_{[\tau]}^{-\top} g_{(\tau)})^{\otimes 2})]$ and $\mathbb{E}_{\mathcal{D}_\mathcal{X}}[(\mathbf{T}^\star_{[\tau]} f^\star_{(\tau)})^{\otimes 2})] = \mathbb{E}_{\mathcal{D}_\mathcal{Y}}[((\mathbf{T}^\star_{[\tau]})^{-\top} g^\star_{(\tau)})^{\otimes 2})]$.

(b) $\lim_{\tau \to \infty} \mathbf{T}_{[\tau]} = \lim_{\tau \to \infty} \mathbf{T}^\star_{[\tau]} = \mathbf{I}_p$.

The above lemma is proved in Subsubsection J.6.4. With these balancing matrices, we devise a new sequence of balanced functions and associated quantities:

$$f_{[\tau]} = \mathbf{T}_{[\tau]} f_{(\tau)}, \quad g_{[\tau]} = \mathbf{T}_{[\tau]}^{-\top} g_{(\tau)}, \quad \mathbf{\Sigma}_{[\tau]} := \mathbb{E}_{\mathcal{D}_\mathcal{X}}[(f_{[\tau]})^{\otimes 2})]$$
$$f^\star_{[\tau]} = \mathbf{T}^\star_{[\tau]} f^\star_{(\tau)}, \quad g^\star_{[\tau]} = (\mathbf{T}^\star_{[\tau]})^{-\top} g^\star_{(\tau)}, \quad \mathbf{\Sigma}^\star_{[\tau]} := \mathbb{E}_{\mathcal{D}_\mathcal{X}}[(f^\star_{[\tau]})^{\otimes 2})],$$

and, letting $\mathbf{P}^\star_{k,[\tau]}$ project onto the top k singular values of $\mathbf{\Sigma}^\star_{[\tau]}$ and defining $\mathbf{P}_{k,[\tau]}$ analogously, we set

$$f^\star_{k,[\tau]} = \mathbf{P}^\star_{k,[\tau]} f^\star_{[\tau]}, \quad g^\star_{k,[\tau]} = \mathbf{P}^\star_{k,[\tau]} f^\star_{[\tau]}$$
$$f_{k,[\tau]} = \mathbf{P}_{k,[\tau]} f_{[\tau]}, \quad g_{k,[\tau]} = \mathbf{P}_{k,[\tau]} g_{[\tau]}.$$

We also define the errors

$$\epsilon^2_{[\tau]} := \mathbb{E}_{\mathcal{D}_\otimes}[(\langle f_{[\tau]}, g_{[\tau]} \rangle - \langle f^\star_{[\tau]}, g^\star_{[\tau]} \rangle)^2], \quad \epsilon^2_{k,[\tau]} := \mathbb{E}_{\mathcal{D}_\otimes}[(\langle f_{[\tau]}, g_{[\tau]} \rangle - \langle f^\star_{[\tau]}, g^\star_{[\tau]} \rangle)^2].$$

Lastly, we define

$$\Delta_{0,[\tau]}(\mathbf{R}, k) = \mathbb{E}_{\mathcal{D}_\otimes}[\langle f^\star_{k,[\tau]} - \mathbf{R} f_{[\tau]}, g^\star_{k,[\tau]} \rangle^2] \vee \mathbb{E}_{\mathcal{D}_\otimes}[\langle f^\star_{k,[\tau]}, \mathbf{R} g_{[\tau]} - g^\star_{k,[\tau]} \rangle^2]$$
$$\Delta_{1,[\tau]}(\mathbf{R}, k) = \mathbb{E}_{\mathcal{D}_\mathcal{X}}[\| f^\star_{k,[\tau]} - \mathbf{R} f_{[\tau]} \|^2] \vee \mathbb{E}_{\mathcal{D}_\mathcal{Y}}[\| g^\star_{k,[\tau]} - \mathbf{R} g_{[\tau]} \|^2],$$

and recall

$$\Delta_0(\mathbf{R}, k) = \mathbb{E}_{\mathcal{D}_\otimes}[\langle f^\star_k - \mathbf{R} f, g^\star_k \rangle^2] \vee \mathbb{E}_{\mathcal{D}_\otimes}[\langle f^\star_k, \mathbf{R} g - g^\star_k \rangle^2],$$
$$\Delta_1(\mathbf{R}, k) = \mathbb{E}_{\mathcal{D}_\mathcal{X}}[\| f^\star_k - \mathbf{R} f \|^2] \vee \mathbb{E}_{\mathcal{D}_\mathcal{Y}}[\| \mathbf{R} g - g^\star_k \|^2].$$

Analyzing the balanced functions. In order to conclude the proof, we establish numerous useful properties of the balanced function sequence. The following lemma is proved in Subsubsection J.6.5.

Lemma J.8 *The followings are true:*

(a) The sequences of balanced functions converge to their targets in $\mathcal{L}_2$:

$$f_{[\tau]} \overset{\mathcal{L}_2(\mathcal{D}_\mathcal{X})}{\to} f, \quad g_{[\tau]} \overset{\mathcal{L}_2(\mathcal{D}_\mathcal{Y})}{\to} g, \quad f^\star_{[\tau]} \overset{\mathcal{L}_2(\mathcal{D}_\mathcal{X})}{\to} f^\star, \quad g^\star_{[\tau]} \overset{\mathcal{L}_2(\mathcal{D}_\mathcal{Y})}{\to} g^\star.$$

More generally, if $\mathbf{R}_{\tau_n}$ is a convergent subsequence converging to $\mathbf{R}$, then $\mathbf{R}_{\tau_n} f_{[\tau_n]} \overset{\mathcal{L}_2(\mathcal{D}_\mathcal{X})}{\to} \mathbf{R} f$ and $\mathbf{R}_{\tau_n} g_{[\tau_n]} \overset{\mathcal{L}_2(\mathcal{D}_\mathcal{Y})}{\to} \mathbf{R} g$ as $n \to \infty$.

(b) *We have* $\lim_{\tau\to\infty} \mathbf{\Sigma}^\star_{[\tau]} = \mathbf{\Sigma}^\star$. *Hence, by Weyl's inequality,* $\lim_{\tau\to\infty} \mathsf{tail}_q(\mathbf{\Sigma}^\star_{[\tau]}; k) = \mathsf{tail}_q(\mathbf{\Sigma}^\star; k)$ *for any* $q, k \geq 1$ *(note that we have assumed here finite-dimensional embeddings, so the covariance operators are matrices and thus the sense of convergence is unambiguous).*

(c) *Similarly,* $\lim_{\tau\to\infty} \mathbf{\Sigma}_{[\tau]} = \mathbf{\Sigma}$. *More generally, if* $\mathbf{R}_{\tau_n}$ *is a convergent subsequence converging to* $\mathbf{R}$*, then* $\lim_{n\to\infty} \mathbf{R}_{\tau_n}\mathbf{\Sigma}_{[\tau_n]}\mathbf{R}_{\tau_n}^\top = \mathbf{R}\mathbf{\Sigma}\mathbf{R}^\top$.

(d) *For any* k *for which* $\sigma_k(\mathbf{\Sigma}^\star) > \sigma_{k+1}(\mathbf{\Sigma}^\star)$, $\lim_{\tau\to\infty} \mathbf{P}^\star_{k,[\tau]} = \mathbf{P}^\star_k$*, where* $\mathbf{P}^\star_k$ *projects onto the top* k*-eigenspace of* $\mathbf{\Sigma}^\star$. *Similarly, for any* k *for which* $\sigma_k(\mathbf{\Sigma}) > \sigma_{k+1}(\mathbf{\Sigma})$, $\lim_{\tau\to\infty} \mathbf{P}_{k,[\tau]} = \mathbf{P}_k$*, where* $\mathbf{P}_k$ *projects onto the top* k*-eigenspace of* $\mathbf{\Sigma}$.

(e) *For any* k *for which* $\sigma_k(\mathbf{\Sigma}^\star) > \sigma_{k+1}(\mathbf{\Sigma}^\star)$,

$$f^\star_{k,[\tau]} \overset{\mathcal{L}_2(\mathcal{D}_\mathcal{X})}{\to} f^\star_k, \quad g^\star_{k,[\tau]} \overset{\mathcal{L}_2(\mathcal{D}_\mathcal{Y})}{\to} g^\star_k.$$

Similarly, for any k *for which* $\sigma_k(\mathbf{\Sigma}) > \sigma_{k+1}(\mathbf{\Sigma})$,

$$f_{k,[\tau]} \overset{\mathcal{L}_2(\mathcal{D}_\mathcal{X})}{\to} f_k, \quad g_{k,[\tau]} \overset{\mathcal{L}_2(\mathcal{D}_\mathcal{Y})}{\to} g_k.$$

(f) *For any* $\mathbf{R} \in \mathbb{O}(p)$*, and* k *for which* $\sigma_k(\mathbf{\Sigma}^\star) > \sigma_{k+1}(\mathbf{\Sigma}^\star)$, $\lim_{\tau\to\infty} \Delta_{0,[\tau]}(\mathbf{R}, k) = \Delta_0(\mathbf{R}, k)$ *and* $\lim_{\tau\to\infty} \Delta_{1,[\tau]}(\mathbf{R}, k) = \Delta_1(\mathbf{R}, k)$. *More generally, if* $\mathbf{R}_{\tau_n}$ *is a convergent subsequence converging to* $\mathbf{R}$*, then we have* $\lim_{n\to\infty} \Delta_{i,[\tau_n]}(\mathbf{R}_{\tau_n}, k) = \Delta_i(\mathbf{R}, k)$, $i \in \{0, 1\}$.

(g) $\lim_{\tau\to\infty} \epsilon^2_{[\tau]} = \epsilon^2_{\mathrm{pred}} \leq \epsilon^2$ *and, for any* k *satisfying both* $\sigma_k(\mathbf{\Sigma}^\star) > \sigma_{k+1}(\mathbf{\Sigma}^\star)$ *and* $\sigma_k(\mathbf{\Sigma}) > \sigma_{k+1}(\mathbf{\Sigma})$ *(supposing such a* k *exists),* $\lim_{\tau\to\infty} \epsilon^2_{k,[\tau]} = \epsilon^2_{\mathrm{pred},k}$.

(h) *For some* η *sufficiently small, and for* ϵ *chosen to satisfy Eq. (J.4) for some* $s \in \mathbb{N}$ *and* $s > 1$*, there exists some* τ_0 *such that, for all* $\tau \geq \tau_0$ *sufficiently large,*

$$\epsilon^2_{[\tau]} \leq (1+\eta)\epsilon^2 \leq 2\epsilon^2 \vee \frac{\|\mathbf{\Sigma}^\star_{[\tau]}\|^2_{\mathrm{op}}}{40^2 s^2}.$$

Concluding the proof. We are now in a position to complete the proofs of Proposition J.2 and Lemma J.10.

Proof [Proof of Proposition J.2] By applying Proposition J.1 to the functions $f_{[\tau]}, g_{[\tau]}, f^\star_{[\tau]}, g^\star_{[\tau]}$ with $\epsilon^2 \leftarrow (1+\eta)\epsilon^2 \geq \epsilon^2_{[\tau]}$ and invoking Lemma J.8 part (h), the following claim is immediate:

Claim J.1 *For all* $\tau \geq \tau_0$*, there exists a* $\mathbf{R}_\tau$ *and* $k_\tau \in [s-1]$ *such that*

$$\Delta_{0,[\tau]}(\mathbf{R}_\tau, k_\tau) \lesssim \epsilon^2 \cdot s^3. \tag{J.6}$$

$$\Delta_{1,[\tau]}(\mathbf{R}_\tau, k_\tau) \lesssim (\sqrt{r} + s^2)\epsilon + s\sigma_s(\mathbf{\Sigma}^\star_{[\tau]}) + \mathsf{tail}_1(\mathbf{\Sigma}^\star_{[\tau]}; s). \tag{J.7}$$

[7] *Moreover, the index* k_τ *satisfies*

$$\mathsf{tail}_2(\mathbf{\Sigma}^\star_{[\tau]}; k_\tau) \lesssim s^3\epsilon^2 + s(\sigma_s(\mathbf{\Sigma}^\star_{[\tau]}))^2 + \mathsf{tail}_2(\mathbf{\Sigma}^\star_{[\tau]}; s).$$

Above, we note $\lesssim$ *hides universal constants independent of* τ. *Morever,*

$$\mathbf{R}_\tau\mathbf{\Sigma}_{[\tau]}\mathbf{R}_\tau^\top \succeq \epsilon\mathbf{P}^\star_{k,[\tau]}, \quad \sigma_{k_\tau}(\mathbf{\Sigma}^\star_{[\tau]}) - \sigma_{k_\tau+1}(\mathbf{\Sigma}^\star_{[\tau]}) \geq 40\epsilon/s.$$

Lastly, $\max_{j\in[p]} |\sigma_j(\mathbf{\Sigma}^\star_{[\tau]}) - \mathbf{\Sigma}_{[\tau]}| \leq \epsilon_{[\tau]}$.

7. A literal invocation of Proposition J.1 would take $\ell_{\star,[\tau]}(\epsilon, s) := \min\{1 + \log \frac{\|\mathbf{\Sigma}^\star_{[\tau]}\|_{\mathrm{op}}}{40(1+\eta)s\epsilon}, s\}$. Here, we use $(1+\eta) \geq 1$.

We may now conclude the proof of Proposition J.2. Since $[s-1]$ is a finite set, and $\mathbb{O}(p)$ is compact, there exists a subsequence $(\mathbf{R}_{\tau_n}, k_{\tau_n})$ so that $\tau_n \ge \tau_0$ for all n, $k_{\tau_n} = k$ for some fixed $k \in [s-1]$, and $\mathbf{R}_{\tau_n} \to \mathbf{R}$ for some fixed $\mathbf{R} \in \mathbb{O}(p)$. By Lemma J.8 part (b) and Weyl's inequality, it must be the case that this k satisfies $\sigma_k(\mathbf{\Sigma}^\star) - \sigma_{k+1}(\mathbf{\Sigma}^\star) \ge 40\epsilon/s > 0$. Hence,

$$\begin{aligned}
\Delta_0(\mathbf{R}, k) &= \lim_{n\to\infty} \Delta_{0,[\tau_n]}(\mathbf{R}_{\tau_n}, k_{\tau_n}) && \text{(Lemma J.8 part (f))}\\
&\lesssim \epsilon^2 \cdot s^2 \cdot \lim_{n\to\infty} \ell_{\star,[\tau_n]}(\epsilon, s) && \text{(Claim J.1)}\\
&= \epsilon^2 \cdot s^2 \cdot \lim_{n\to\infty} \min\left\{1 + \log \frac{\|\mathbf{\Sigma}^\star_{[\tau_n]}\|_{\mathrm{op}}}{40 s\epsilon},\, s\right\} && \text{(see Claim J.1)}\\
&= \epsilon^2 \cdot s^2 \cdot \underbrace{\min\left\{1 + \log \frac{\|\mathbf{\Sigma}^\star\|_{\mathrm{op}}}{40 s\epsilon},\, s\right\}}_{=\ell_\star(\epsilon,s)}, && \text{(Lemma J.8 part (b))}
\end{aligned}$$

and,

$$\begin{aligned}
\Delta_1(\mathbf{R}, k) &= \lim_{n\to\infty} \Delta_{1,[\tau_n]}(\mathbf{R}_{\tau_n}, k_{\tau_n}) && \text{(Lemma J.8 part (f))}\\
&\lesssim (\sqrt{r} + s^2)\epsilon + \lim_{n\to\infty}\left(s\sigma_s(\mathbf{\Sigma}^\star_{[\tau_n]}) + \mathsf{tail}_1(\mathbf{\Sigma}^\star_{[\tau_n]}; s)\right) && \text{(Claim J.1)}\\
&\le (\sqrt{r} + s^2)\epsilon + s\sigma_s(\mathbf{\Sigma}^\star) + \mathsf{tail}_1(\mathbf{\Sigma}^\star; s). && \text{(Lemma J.8 part (b))}
\end{aligned}$$

Second,

$$\begin{aligned}
\mathsf{tail}_2(\mathbf{\Sigma}^\star; k) &= \lim_{n\to\infty} \mathsf{tail}_2(\mathbf{\Sigma}^\star_{[\tau_n]}; k_{\tau_n}) && \text{(Lemma J.8 part (b))}\\
&\lesssim s^3\epsilon^2 + \lim_{n\to\infty}\left(s(\sigma_s(\mathbf{\Sigma}^\star_{[\tau_n]}))^2 + \mathsf{tail}_2(\mathbf{\Sigma}^\star_{[\tau_n]}; s)\right) && \text{(Claim J.1)}\\
&\lesssim s^3\epsilon^2 + s(\sigma_s(\mathbf{\Sigma}^\star))^2 + \mathsf{tail}_2(\mathbf{\Sigma}^\star; s). && \text{(Lemma J.8 part (b))}
\end{aligned}$$

Third,

$$\begin{aligned}
\mathbf{R}\mathbf{\Sigma}\mathbf{R}^\top &= \lim_{n\to\infty} \mathbf{R}_{\tau_n}\mathbf{\Sigma}_{[\tau_n]}\mathbf{R}_{\tau_n}^\top && \text{(Lemma J.8 part (c))}\\
&\succeq \epsilon \lim_{n\to\infty} \mathbf{P}^\star_{k,[\tau_n]} && \text{(Claim J.1)}\\
&= \epsilon\mathbf{P}^\star_k. && \text{(Lemma J.8 part (d))}
\end{aligned}$$

Finally, by parts (b) and (c), the fact that $\max_{j\in[p]} |\sigma_j(\mathbf{\Sigma}^\star_{[\tau]}) - \mathbf{\Sigma}_{[\tau]}| \le \epsilon_{[\tau]}$ due to Lemma J.8, and Weyl's inequality. This concludes the proof of Proposition J.2. ■

J.5. From finite to infinite dimensional embeddings: Proof of Theorem 10

We give the proof of Theorem 10 from Proposition J.2. Fix any $\hat{f} : \mathcal{X} \to \mathbb{R}^r$ and $\hat{g} : \mathcal{Y} \to \mathbb{R}^r$ that satisfy

$$\mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle \hat{f}, \hat{g}\rangle_{\mathbb{R}^r} - \langle f^\star, g^\star\rangle_{\mathcal{H}})^2] \le \epsilon^2 < \frac{\|\mathbf{\Sigma}^\star_{1\otimes 1}\|^2_{\mathrm{op}}}{(40s)^2}. \tag{J.8}$$

Now, fix a $p \in \mathbb{N}$, which we shall take sufficiently large. Since $\mathbf{\Sigma}^\star_{1\otimes 1}$ is trace class, we may assume without loss of generality that the space spanned by its top p eigenvectors is unique.[8] Let $\mathcal{V}^\star_p$ denote the eigenspace spanned by these eigenvectors, and note that $\mathcal{V}^\star_p$ is isomorphic to $\mathbb{R}^p$. Finally, let let $\iota : \mathbb{R}^r \to \mathcal{V}^\star_p \subset \mathcal{H}$ by an isometric inclusion in the sense of Definition 4.2.

Suppose first that $(\hat{f}, \hat{g})$ are full-rank, and let $\mathbf{T}$ be the balancing operator guaranteed by Lemma 4.3. As ι is an isometry, Eq. (J.8) implies

$$\mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle \iota(\mathbf{T}\hat{f}), \iota(\mathbf{T}^{-1}\hat{g})\rangle_{\mathcal{H}} - \langle f^\star, g^\star\rangle_{\mathcal{H}})^2] \le \epsilon^2.$$

as well as the following equality, which can be checked by evaluating the induced quadratic forms.

$$\mathbf{\Sigma}_\iota := \mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[\iota(\hat{f})\iota(\hat{f})^\top] = \mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}[\iota(\hat{g})\iota(\hat{g})^\top], \tag{J.9}$$

so that $\iota(\hat{f})$ and $\iota(\hat{g})$ are balanced and take values in $\mathcal{V}^\star_p$. A standard expansion and Cauchy-Schwartz inequality imply that

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle \iota(\hat{f}), \iota(\hat{g})\rangle_{\mathcal{H}} - \langle f^\star_p, g^\star_p\rangle_{\mathcal{H}})^2] \\
&= \mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle \hat{f}, \hat{g}\rangle_{\mathbb{R}^r} - \langle f^\star_p, g^\star_p\rangle_{\mathcal{H}})^2] \\
&\le \mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle \hat{f}, \hat{g}\rangle_{\mathbb{R}^r} - \langle f^\star, g^\star\rangle_{\mathcal{H}})^2] + \mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle f^\star, g^\star\rangle_{\mathcal{H}} - \langle f^\star_p, g^\star_p\rangle_{\mathcal{H}})^2] \\
&\quad + 2\sqrt{\mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle \hat{f}, \hat{g}\rangle_{\mathbb{R}^r} - \langle f^\star, g^\star\rangle_{\mathcal{H}})^2]\mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle f^\star, g^\star\rangle_{\mathcal{H}} - \langle f^\star_p, g^\star_p\rangle_{\mathcal{H}})^2]} \\
&= \epsilon^2 + 2\mathbf{tail}^\star_2(p)^{1/2}\epsilon + \mathbf{tail}^\star_2(p) && \text{(Lemma M.4)} \\
&= \epsilon^2 + 2\mathbf{tail}^\star_2(p)^{1/2}\epsilon + \mathbf{tail}^\star_2(p) := \epsilon^2_{[p]}. && \text{(J.10)}
\end{aligned}$$

Recall that $\epsilon^2 < \frac{\|\mathbf{\Sigma}^\star_{1\otimes 1}\|^2_{\mathrm{op}}}{(40s)^2}$. Hence, by choosing some $p \ge s$ sufficiently large, we can ensure $\mathbf{tail}^\star_2(p)$ is small enough that

$$\epsilon_{[p]} < \frac{\|\mathbf{\Sigma}^\star_{1\otimes 1}\|_{\mathrm{op}}}{40\epsilon}, \quad \epsilon^2_{[p]} \le 2\epsilon^2. \tag{J.11}$$

As $\|\mathbf{\Sigma}^\star_{1\otimes 1}\|_{\mathrm{op}} = \|\mathbf{\Sigma}^\star_p\|_{\mathrm{op}}$ due to Eq. (J.13), we then have

$$\epsilon_{[p]} < \frac{\|\mathbf{\Sigma}^\star_p\|_{\mathrm{op}}}{40\epsilon}. \tag{J.12}$$

Continuing the proof, let $\mathbf{\Sigma}^\star_{[p]}$ denote $\mathbb{E}_{\mathcal{D}_{1\otimes 1}}[f^\star_p(f^\star_p)^\top]$, viewed as an operator on $\mathcal{V}^\star_p$, and notice that $\mathbf{\Sigma}^\star = \mathbb{E}_{\mathcal{D}_{1\otimes 1}}[g^\star_p(g^\star_p)^\top]$ by Assumption 2.4, and that

$$\sigma_i(\mathbf{\Sigma}^\star_{[p]}) = \begin{cases} \sigma_i(\mathbf{\Sigma}^\star_{1\otimes 1}) & i \in [p] \\ 0 & i > p \end{cases}. \tag{J.13}$$

Viewing $\mathcal{V}^\star_p$ as isomorphic to $\mathbb{R}^p$, define the error terms consider by Proposition J.2:

$$\begin{aligned}
\Delta_0(\mathbf{R}, k) &= \mathbb{E}_{\mathcal{D}_\otimes}[\langle f^\star_k - \mathbf{R}\iota(\mathbf{T}\hat{f}), g^\star_k\rangle^2_{\mathcal{V}^\star_p}] \vee \mathbb{E}_{\mathcal{D}_\otimes}[\langle f^\star_k, \mathbf{R}\iota(\mathbf{T}^{-1}\hat{g}) - g^\star_k\rangle^2_{\mathcal{V}^\star_p}] \\
\Delta_1(\mathbf{R}, k) &= \mathbb{E}_{\mathcal{D}_\mathcal{X}}[\|f^\star_k - \mathbf{R}\iota(\mathbf{T}\hat{f})\|^2_{\mathcal{V}^\star_p}] \vee \mathbb{E}_{\mathcal{D}_\mathcal{Y}}[\|\mathbf{R}\iota(\mathbf{T}^{-1}\hat{g}) - g^\star_k\|^2_{\mathcal{V}^\star_p}],
\end{aligned}$$

8. If $\mathrm{rank}(\mathbf{\Sigma}^\star_{1\otimes 1})$ is finite, let p equal the rank. Otherwise, the eigenvectors must have decay so for any p, there exists some $p' \ge p$ which has eigengap.

where above the norms and inner products are the standard Euclidean inner product on $\mathcal{V}_p^\star$, and where $\mathbf{R}$ is an orthogonal transformation of $\mathcal{V}_p^\star$. From Eqs. (J.11) and (J.12), we can apply Proposition J.2 to find that there exists an orthogonal operator $\mathbf{R}$ for which

$$\Delta_0(\mathbf{R}, k) \lesssim \epsilon_{[p]}^2 \cdot s^3$$
$$\Delta_1(\mathbf{R}, k) \lesssim (\sqrt{r} + s^2)\epsilon_{[p]} + s\sigma_s(\mathbf{\Sigma}_{[p]}^\star) + \mathsf{tail}_1(\mathbf{\Sigma}_{[p]}^\star; s).$$

Moreover, the index k satisfies

$$\mathsf{tail}_2(\mathbf{\Sigma}_{[p]}^\star; k) \lesssim s^3\epsilon_{[p]}^2 + s(\sigma_s(\mathbf{\Sigma}_{[p]}^\star))^2 + \mathsf{tail}_2(\mathbf{\Sigma}_{[p]}^\star; s).$$

Lastly, $\mathbf{R}$ and k satisfy

$$\mathbf{R}\mathbf{\Sigma}_\iota\mathbf{R}^\top \succeq \epsilon_{[p]}\mathbf{P}_{[k]}^\star, \tag{J.14}$$

where $\mathbf{P}_{[k]}^\star$ is the projection onto the top-k singular space of $\mathbf{\Sigma}_{[p]}^\star$, namely $\mathcal{V}_p^\star$, and were $\mathbf{\Sigma}_\iota$ is as in Eq. (J.9). We now argue that for this transformation $\mathbf{R}$, the embeddings (f, g) defined by[9]

$$f := \mathbf{R}\iota(\mathbf{T}\hat{f}), \quad g := \mathbf{R}\iota(\mathbf{T}^{-1}\hat{g}),$$

satisfy the conclusion of Theorem 10.

Proof of part (a). This follows from direct computation, as ι and multiplication by $\mathbf{R}$ are isometries.

Proof of part (b). This follows from the definition of f and g, and from Eq. (J.14).

Proof of part (c). By Eq. (J.13), we have

$$\mathsf{tail}_j(\mathbf{\Sigma}_{[p]}^\star; k) \le \mathbf{tail}_j^\star(k), \quad i \in \{1, 2\}.$$

Combining this and the facts that $p \ge s$ and that $\epsilon_{[p]} \lesssim \epsilon$ (see Eq. (J.10)),

$$\Delta_1(\mathbf{R}, k) \lesssim (\sqrt{r} + s^2)\epsilon + s\sigma_s(\mathbf{\Sigma}_{1\otimes1}^\star) + \mathbf{tail}_1^\star(s), \tag{J.15}$$

as well as $\Delta_0(\mathbf{R}, k) \lesssim \epsilon^2 \cdot s^3$. Similarly, the index k satisfies

$$\mathsf{tail}_2(\mathbf{\Sigma}_{[p]}^\star; k) \lesssim s^3\epsilon^2 + s(\sigma_s(\mathbf{\Sigma}_{1\otimes1}^\star))^2 + \mathbf{tail}_2^\star(s).$$

Moreover, from Eq. (J.13), we see $\mathbf{tail}_2^\star(k) = \mathsf{tail}_2(\mathbf{\Sigma}_{[p]}^\star; k) + \mathbf{tail}_2^\star(p)$. Since $p \ge s$,

$$\begin{aligned}\mathbf{tail}_2^\star(k) &\lesssim s^3\epsilon^2 + s(\sigma_s(\mathbf{\Sigma}_{1\otimes1}^\star))^2 + \mathbf{tail}_2^\star(s) + \mathbf{tail}_2^\star(p)\\ &\lesssim s^3\epsilon^2 + s(\sigma_s(\mathbf{\Sigma}_{1\otimes1}^\star))^2 + \mathbf{tail}_2^\star(s).\end{aligned} \tag{J.16}$$

Thus,

$$\mathbf{tail}_2^\star(k) + \Delta_0(\mathbf{R}, k) \lesssim s^3\epsilon^2 + s(\sigma_s(\mathbf{\Sigma}_{1\otimes1}^\star))^2 + \mathbf{tail}_2^\star(s). \tag{J.17}$$

To conclude

$$\Delta_0(\mathbf{R}, k) = \mathbb{E}_{\mathcal{D}_\otimes}[\langle f_k^\star - f, g_k^\star\rangle_{\mathcal{V}_p^\star}^2] \vee \mathbb{E}_{\mathcal{D}_\otimes}[\langle f_k^\star, g - g_k^\star\rangle_{\mathcal{V}_p^\star}^2], \tag{J.18}$$

$$\Delta_1(\mathbf{R}, k) = \mathbb{E}_{\mathcal{D}_\mathcal{X}}[\|f_k^\star - f\|_{\mathcal{V}_p^\star}^2] \vee \mathbb{E}_{\mathcal{D}_\mathcal{Y}}[\|g_k^\star - g\|_{\mathcal{V}_p^\star}^2]. \tag{J.19}$$

Hence, part (c) of Theorem 10 follows from the above identification, and Eqs. (J.15) and (J.17).

Proof of part (d). Applying the last part of Proposition J.2 and using $p \ge r$ implies that

$$\sigma_j(\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[\iota(\tilde{f})\iota(\tilde{f})^\top]) > \sigma_j(\mathbf{\Sigma}_{[p]}^\star) - \epsilon = \sigma_j(\mathbf{\Sigma}^\star) - \epsilon.$$

9. under the natural inclusion of $\mathbf{R}\iota(\hat{f})$ from $\mathcal{V}_p^\star$ to $\mathcal{H}$

Removing the full-rank assumption $\hat{f}, \hat{g}$. To replace the assumption that $(\hat{f}, \hat{g})$ is full-rank with the assumption that $\epsilon < \sigma_r(\boldsymbol{\Sigma}^\star)$, apply Lemma L.8 to show that there exist $\tilde{f}, \tilde{g}$ such that (a) $(\tilde{f}, \tilde{g})$ is full-rank if and only if $(\hat{f}, \hat{g})$ is, (b) $\langle \hat{f}, \hat{g} \rangle = \langle \tilde{f}, \tilde{g} \rangle$ almost surely on $\mathcal{D}_{\mathcal{X},1}$, and (c)

$$\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[\tilde{f}\tilde{f}^\top] = \mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}[\tilde{g}\tilde{g}^\top], \quad \sigma_r(\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[\tilde{f}\tilde{f}^\top]) = \hat{\sigma}.$$

It suffices to show that $\tilde{f}, \tilde{g}$ is full-rank. To this end, let ι be an isometric embedding of $\mathbb{R}^r \to \mathcal{V}_p^\star$ for $p \geq r$, and note we have that $\mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle \hat{f}, \hat{g} \rangle_{\mathbb{R}^r} - \langle f^\star, g^\star \rangle_{\mathcal{H}})^2] \leq \epsilon^2$, and since ι is an isometry,

$$\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[\iota(\tilde{f})\iota(\tilde{f})^\top] = \mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}[\iota(\tilde{g})\iota(\tilde{g})^\top].$$

Applying the last part of Proposition J.2 and using $p \geq r$ implies that

$$\sigma_r(\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[\iota(\tilde{f})\iota(\tilde{f})^\top]) > \sigma_r(\boldsymbol{\Sigma}^\star_{[p]}) - \epsilon = \sigma_r(\boldsymbol{\Sigma}^\star) - \epsilon,$$

which is striclty positive for $\epsilon < \sigma_r(\boldsymbol{\Sigma}^\star)$. Since ι is an isometry, we have shown that $(\tilde{f}, \tilde{g})$ are full-rank, which implies that $(\hat{f}, \hat{g})$ are also full-rank.

J.6. Proof of supporting claims

J.6.1. PROOF OF LEMMA J.1

Proof Define subsets of $\mathbb{R}^p$ by $\mathcal{U}_f = \bigcup_{i=1}^a \{f_i(\mathcal{X})\}$ and $\mathcal{W}_g = \bigcup_{j=1}^b \{g_j(\mathcal{Y})\}$. Since f_i and g_j are simple, $\mathcal{U}_f$ and $\mathcal{W}_f$ are finite. Define the sets

$$\mathcal{X}_{\mathbf{u}} := \bigcap_{i=1}^a f_i^{-1}(\mathbf{u}^{(i)}), \quad \mathbf{u} = (\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(a)}) \in (\mathcal{U}_f)^a$$
$$\mathcal{Y}_{\mathbf{w}} := \bigcap_{j=1}^b g_j^{-1}(\mathbf{w}^{(j)}), \quad \mathbf{w} = (\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(b)}) \in (\mathcal{W}_g)^b.$$

Let $\boldsymbol{\mathcal{U}}_f := \{\mathbf{u} \in (\mathcal{U}_f)^a : \mathcal{X}_{\mathbf{u}} \neq \emptyset\}$ and $\boldsymbol{\mathcal{W}}_g := \{\mathbf{w} \in (\mathcal{W}_g)^b : \mathcal{Y}_{\mathbf{w}} \neq \emptyset\}$. Note that $\boldsymbol{\mathcal{U}}_f$ and $\boldsymbol{\mathcal{W}}_g$ are finite sets (since $\mathcal{U}_f$ and $\mathcal{W}_f$ are). By construction, for each $\mathbf{u} \in \boldsymbol{\mathcal{U}}_f$ (resp. $\mathbf{w} \in \boldsymbol{\mathcal{W}}_g$), there exists an $x_{\mathbf{u}} \in \mathcal{X}$ (resp. $y_{\mathbf{w}} \in \mathcal{Y}$) such that for all $i \in [a]$ and $j \in [b]$,

$$f_i(x_{\mathbf{u}}) = \mathbf{u}^{(i)}, \quad g_j(y_{\mathbf{w}}) = \mathbf{w}^{(j)}.$$

By construction, we also see that the sets $\mathcal{X}_{\mathbf{u}}$ and $\mathcal{Y}_{\mathbf{w}}$ indexed by $\mathbf{u} \in \boldsymbol{\mathcal{U}}_f$ and $\mathbf{w} \in \boldsymbol{\mathcal{W}}_g$ form a partition of $\mathcal{X}$ and $\mathcal{Y}$, so we may define functions $\phi_{\mathcal{X}}$ and $\phi_{\mathcal{Y}}$ by

$$\phi_{\mathcal{X}}(x) := x_{\mathbf{u}},\ x \in \mathcal{X}_{\mathbf{u}}, \quad \mathbf{u} \in \boldsymbol{\mathcal{U}}_f$$
$$\phi_{\mathcal{Y}}(y) := y_{\mathbf{w}},\ y \in \mathcal{Y}_{\mathbf{w}}, \quad \mathbf{w} \in \boldsymbol{\mathcal{W}}_g.$$

Note that since $x_{\mathbf{u}} \in \mathcal{X}_{\mathbf{u}}$, $\phi_{\mathcal{X}}$ is idempotent: $\phi_{\mathcal{X}} = \phi_{\mathcal{X}} \circ \phi_{\mathcal{X}}$; similarly, $\phi_{\mathcal{Y}} = \phi_{\mathcal{Y}} \circ \phi_{\mathcal{Y}}$.

By definition of $\mathcal{X}_{\mathbf{u}}$ and $\mathcal{Y}_{\mathbf{w}}$, it holds for all $x \in \mathcal{X}_{\mathbf{u}}$ and $y \in \mathcal{Y}_{\mathbf{w}}$ that

$$f_i(x) = f_i \circ \phi_{\mathcal{X}}(x), \quad g_j(y) = g_j \circ \phi_{\mathcal{Y}}(y).$$

Hence, for any Ψ, $\Psi(f_1(x), \dots, f_a(x), g_1(y), \dots, g_b(y))$ can be written as some function $\tilde{\Psi}(\phi_{\mathcal{X}}(x), \phi_{\mathcal{Y}}(y))$. To conclude, let $\bar{\mathcal{D}}_{\mathcal{X}}$ denote the distribution of $\phi_{\mathcal{X}}(x)$ under $\mathcal{D}_{\mathcal{X}}$ and $\bar{\mathcal{D}}_{\mathcal{Y}}$ denote the distribution of $\phi_{\mathcal{Y}}(y)$ under $\mathcal{D}_{\mathcal{Y}}$. Then,

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_{\mathcal{X}} \otimes \mathcal{D}_{\mathcal{Y}}}[\Psi(f_1(x), \dots, f_a(x), g_1(y), \dots, g_b(y))] \\
&\quad = \mathbb{E}_{\mathcal{D}_{\mathcal{X}} \otimes \mathcal{D}_{\mathcal{Y}}}[\tilde{\Psi}(\phi_{\mathcal{X}}(x), \phi_{\mathcal{Y}}(y))] \\
&\quad = \mathbb{E}_{\mathcal{D}_{\mathcal{X}} \otimes \mathcal{D}_{\mathcal{Y}}}[\tilde{\Psi}(\phi_{\mathcal{X}} \circ \phi_{\mathcal{X}}(x), \phi_{\mathcal{Y}} \circ \phi_{\mathcal{Y}}(y))] && \text{(Idempotence of } \phi_{\mathcal{X}}, \phi_{\mathcal{Y}}) \\
&\quad = \mathbb{E}_{\bar{\mathcal{D}}_{\mathcal{X}} \otimes \bar{\mathcal{D}}_{\mathcal{Y}}}[\tilde{\Psi}(\phi_{\mathcal{X}}(x), \phi_{\mathcal{Y}}(y))] \\
&\quad = \mathbb{E}_{\bar{\mathcal{D}}_{\mathcal{X}} \otimes \bar{\mathcal{D}}_{\mathcal{Y}}}[\Psi(f_1(x), \dots, f_a(x), g_1(y), \dots, g_b(y))].
\end{aligned}$$

This completes the proof. ■

J.6.2. PROOF OF LEMMAS J.4 AND J.5

Proof [Proof of Lemma J.4] We bound

$$\begin{aligned}
&\|\mathbb{E}[\psi_\tau^{\otimes 2}] - \mathbb{E}[\psi^{\otimes 2}]\|_{\mathrm{F}}^2 \\
&= \left\|\mathbb{E}[\psi(\psi_\tau - \psi)^\top] + \mathbb{E}[(\psi_\tau - \psi)\psi^\top] + \mathbb{E}[(\psi_\tau - \psi)(\psi_\tau - \psi)^\top]\right\|_{\mathrm{F}}^2 \\
&\le 3\|\mathbb{E}[\psi(\psi_\tau - \psi)^\top]\|_{\mathrm{F}}^2 + 3\|\mathbb{E}[(\psi_\tau - \psi)\psi^\top]\|_{\mathrm{F}}^2 + 3\|\mathbb{E}[(\psi_\tau - \psi)(\psi_\tau - \psi)^\top]\|_{\mathrm{F}}^2 \\
&\le 3\mathbb{E}[\|\psi(\psi_\tau - \psi)^\top\|_{\mathrm{F}}^2] + 3\mathbb{E}[\|(\psi_\tau - \psi)\psi^\top\|_{\mathrm{F}}^2] + 3\mathbb{E}[\|(\psi_\tau - \psi)(\psi_\tau - \psi)\|_{\mathrm{F}}^2]. \\
&&& \text{(Jensen's Inequality)} \\
&\le 6\mathbb{E}[\|\psi\|^2\|\psi_\tau - \psi\|^2] + 3\mathbb{E}[\|\psi_\tau - \psi\|^4] \\
&\le 6\mathbb{E}[\|\psi\|^2]\mathbb{E}[\|\psi_\tau - \psi\|^2] + 3\mathbb{E}[\|\psi_\tau - \psi\|^4]. && \text{(Cauchy Schwartz)}
\end{aligned}$$

This last term goes to 0 as $\tau \to \infty$ by definition of $\mathcal{L}_2$ convergence. ■

Proof [Proof of Lemma J.5] Let $r := \mathrm{rank}(\mathbb{E}[\psi^{\otimes 2}])$, so that $\sigma_r(\mathbb{E}[\psi^{\otimes 2}]) > 0$. By Lemma J.4, there exists some τ_0 so that for all $\tau \ge \tau_0$ sufficiently large, $\mathbb{E}[\psi_\tau^{\otimes 2}] \succeq \mathbb{E}[\psi^{\otimes 2}] - \frac{\sigma_r}{2}\mathbf{I}_p$. For all such $\tau \ge \tau_0$, and any $\mathbf{v} \in \mathrm{range}(\mathbb{E}[\psi^{\otimes 2}]) \setminus \{0\}$, we have

$$\mathbf{v}^\top \mathbb{E}[\psi_\tau^{\otimes 2}]\mathbf{v} \ge \mathbf{v}^\top \mathbb{E}[\psi^{\otimes 2}]\mathbf{v} - \sigma_r\|\mathbf{v}\|^2/2 = \sigma_r\|\mathbf{v}\|^2/2 > 0.$$

Thus, $\dim(\mathrm{nullspace}(\mathbb{E}[\psi_\tau^{\otimes 2}])) \le \dim(\mathrm{nullspace}(\mathbb{E}[\psi^{\otimes 2}]))$. Hence,

$$\dim(\mathrm{range}(\mathbb{E}[\psi_\tau^{\otimes 2}])) \ge \dim(\mathrm{range}(\mathbb{E}[\psi^{\otimes 2}])).$$

On the other hand, by assumption, $\mathrm{range}(\mathbb{E}[\psi_\tau^{\otimes 2}]) \subseteq \mathrm{range}(\mathbb{E}[\psi^{\otimes 2}])$. Hence, $\mathrm{range}(\mathbb{E}[\psi_\tau^{\otimes 2}]) = \mathrm{range}(\mathbb{E}[\psi^{\otimes 2}])$. ■

J.6.3. PROOF OF LEMMA J.6

Proof We give the construction of the sequence $f_{(\tau)}$, the others are similar. By Lemma J.3, there exists a sequence of functions f_τ such that $f_\tau \overset{\mathcal{L}_2(\mathcal{D}_\mathcal{X})}{\to} f$, i.e. $\lim_{\tau\to\infty} \mathbb{E}_{\mathcal{D}_\mathcal{X}}\|f_\tau - f\|^2 = 0$. Let $\mathbf{P}$ denote the orthogonal projection onto $\mathrm{range}(\mathbf{\Sigma})$. Then, $\mathbf{P}f = f$ $\mathcal{D}_\mathcal{X}$-almost surely by Lemma L.7. Hence,

$$\begin{aligned}
\lim_{\tau\to\infty} \mathbb{E}_{\mathcal{D}_\mathcal{X}}\|\mathbf{P}f_\tau - f\|^2 &= \lim_{\tau\to\infty} \mathbb{E}_{\mathcal{D}_\mathcal{X}}\|\mathbf{P}(f_\tau - f)\|^2 \\
&\le \lim_{\tau\to\infty} \mathbb{E}_{\mathcal{D}_\mathcal{X}}\|f_\tau - f\|^2 && (\mathbf{P} \text{ is an orthogonal projection}) \\
&= 0. && (f_\tau \overset{\mathcal{L}_2(\mathcal{D}_\mathcal{X})}{\to} f)
\end{aligned}$$

Hence, let $f_{(\tau)} := \mathbf{P}f_\tau \overset{\mathcal{L}_2(\mathcal{D}_\mathcal{X})}{\to} f$. Moreover, $\mathbf{\Sigma}_{(\tau),f} := \mathbb{E}[(f_{(\tau)})^{\otimes 2}] = \mathbf{P}\mathbb{E}[(f_\tau)^{\otimes 2}]\mathbf{P}^\top$, which ensures the inclusion of the rowspaces. ■

J.6.4. PROOF OF LEMMA J.7

We demonstrate the existence of $\mathbf{T}_{[\tau]}$, and note that the existence of $\mathbf{T}^\star_{[\tau]}$ is similar. Recall from Lemma J.6 that $\mathbf{\Sigma}_{(\tau),f} := \mathbb{E}[(f_{(\tau)})^{\otimes 2}]$, and $\mathbf{\Sigma}_{(\tau),g} := \mathbb{E}[(g_{(\tau)})^{\otimes 2}]$, and $\mathrm{range}(\mathbf{\Sigma}_{(\tau),f}) \cup \mathrm{range}(\mathbf{\Sigma}_{(\tau),g}) \subseteq \mathrm{range}(\mathbf{\Sigma})$. By Lemma J.6 and Lemma J.4, it follows that

$$\lim_{\tau\to\infty} \mathbf{\Sigma}_{(\tau),f} = \lim_{\tau\to\infty} \mathbf{\Sigma}_{(\tau),g} = \mathbf{\Sigma}$$

and for some τ_0, $\mathrm{range}(\mathbf{\Sigma}_{(\tau),f}) = \mathrm{range}(\mathbf{\Sigma}_{(\tau),g}) = \mathrm{range}(\mathbf{\Sigma})$ (taking τ_0 to be the maximum of the two $\tau_{0,f}$ and $\tau_{0,g}$ required for $\mathbf{\Sigma}_{(\tau),f}$ and $\mathbf{\Sigma}_{(\tau),g}$ individually). Let $r = \dim(\mathrm{range}(\mathbf{\Sigma}))$. By inflating τ_0 if necessary, we can ensures $\sigma_r(\mathbf{\Sigma}_{(\tau),f}) = \sigma_r(\mathbf{\Sigma}_{(\tau),g}) > \sigma_r(\mathbf{\Sigma})/2$ for all $\tau \ge \tau_0$. Hence, by Lemma L.2, for all $\tau \ge \tau_0$, there exist some $\mathbf{T}_{[\tau]} \in \mathbb{S}^p_>$ such that

$$\mathbf{T}_{[\tau]}\mathbf{\Sigma}_{(\tau),f}\mathbf{T}_{[\tau]} = \mathbf{T}^{-1}_{[\tau]}\mathbf{\Sigma}_{(\tau),g}\mathbf{T}^{-1}_{[\tau]},$$

satisfying

$$\max\{\|\mathbf{T}_{[\tau]}\|_{\mathrm{op}}, \|\mathbf{T}^{-1}_{[\tau]}\|_{\mathrm{op}}\} \le (1+\Delta)^{1/4}, \quad \text{where } \Delta := \frac{\|\mathbf{\Sigma}_{(\tau),f} - \mathbf{\Sigma}_{(\tau),g}\|_{\mathrm{op}}}{2\sigma_r(\mathbf{\Sigma})} \overset{\tau\to\infty}{\to} 0 .$$

Notice that $\mathbf{T}_{[\tau]} = \mathbf{T}^\top_{[\tau]}$ ($\mathbb{S}^p_>$ contains only symmetric matrices), we also have

$$\mathbb{E}[(\mathbf{T}_{[\tau]}f_{(\tau)})^{\otimes 2}] = \mathbf{T}_{[\tau]}\mathbf{\Sigma}_{(\tau),f}\mathbf{T}^\top_{[\tau]} = \mathbf{T}^{-\top}_{[\tau]}\mathbf{\Sigma}_{(\tau),g}\mathbf{T}^{-1}_{[\tau]}\mathbb{E}[(\mathbf{T}^{-\top}_{[\tau]}g_{(\tau)})^{\otimes 2}],$$

and that since $\lim_{\tau\to\infty}\max\{\|\mathbf{T}_{[\tau]}\|_{\mathrm{op}}, \|\mathbf{T}^{-1}_{[\tau]}\|_{\mathrm{op}}\} = 1$, $\lim_{\tau\to\infty}\mathbf{T}_{[\tau]} = \mathbf{I}_p$.

J.6.5. PROOF OF LEMMA J.8

Part (a). We have

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_\mathcal{X}}\|f_{[\tau]} - f\|^2 &= \mathbb{E}_{\mathcal{D}_\mathcal{X}}\|\mathbf{T}_{[\tau]}f_{(\tau)} - f\|^2 \\
&\le 2\mathbb{E}_{\mathcal{D}_\mathcal{X}}\|(\mathbf{T}_{[\tau]} - \mathbf{I}_p)f_{(\tau)}\|^2 + 2\mathbb{E}_{\mathcal{D}_\mathcal{X}}\|f_{(\tau)} - f\|^2 \\
&\le 2\|\mathbf{T}_{[\tau]} - \mathbf{I}_p\|^2_{\mathrm{op}} \cdot \mathbb{E}_{\mathcal{D}_\mathcal{X}}\|f_{(\tau)}\|^2 + 2\mathbb{E}_{\mathcal{D}_\mathcal{X}}\|f_{(\tau)} - f\|^2.
\end{aligned}$$

Since $f_{(\tau)} \overset{\mathcal{L}_2(\mathcal{D}_{\mathcal{X}})}{\to} f$, $\sup_\tau \mathbb{E}_{\mathcal{D}_{\mathcal{X}}}\|f_{(\tau)}\|^2 \le M$ for some $M < \infty$. Hence, by Lemmas J.6 and J.7,

$$\lim_{\tau\to\infty} \mathbb{E}_{\mathcal{D}_{\mathcal{X}}}\|f_{[\tau]} - f\|^2 \le \lim_{\tau\to\infty} 2\|\mathbf{T}_{[\tau]} - \mathbf{I}_p\|_{\mathrm{op}}^2 M + \lim_{\tau\to\infty} 2\mathbb{E}_{\mathcal{D}_{\mathcal{X}}}\|f_{(\tau)} - f\|^2 = 0.$$

The proofs of the other guarantees are similar.

Parts (b) and (c). These follow from part (a) and Lemma J.4.

Part (d). Set $\zeta := \sigma_k(\mathbf{\Sigma}^\star) - \sigma_{k+1}(\mathbf{\Sigma}^\star)$, and assume $\zeta > 0$. By part (b), $\lim_{\tau\to\infty} \mathbf{\Sigma}^\star_{[\tau]} = \mathbf{\Sigma}^\star$. Hence, by Weyl's inequality, there exists some τ_0 such that for all $\tau \ge \tau_0$, $\sigma_{k+1}(\mathbf{\Sigma}^\star_{[\tau]}) < \sigma_k(\mathbf{\Sigma}^\star) - \zeta/2$. The convergence then follows by Wedin's Theorem (see e.g. Lemma D.2). The convergence of $\mathbf{P}_{k,[\tau]}$ to $\mathbf{P}_k$ is analogous.

Part (e). This follows from parts (a) and (d).

Part (f). This can be checked by using parts (a) and (e), together with standard applications of Cauchy Schwartz and/or Jensen's inequality.

Part (g). The first statement can be checked by using part (a), together with standard applications of Cauchy Schwartz and/or Jensen's inequality. The second uses part (e) instead of part (a).

Part (h). Since $\lim_{\tau\to\infty} \frac{\|\mathbf{\Sigma}^\star_{[\tau]}\|_{\mathrm{op}}^2}{40^2 s^2} = \frac{\|\mathbf{\Sigma}^\star\|_{\mathrm{op}}^2}{40^2 s^2}$ by part (b) and Weyl's inequality, part (h) follows from part (g) and Eq. (J.4). This completes the proof. ∎

J.7. SVD perturbation for distribution embeddings

In this section, we reiterate the limiting analysis to establish an embedding analogue of our main perturbation result for the singular-value decomposition (Theorem 1). Specifically, the main result of this section is:

Theorem 11 *Let (f, g) be balanced embeddings, with $\mathbf{\Sigma} = \mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[ff^\top] = \mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}[gg^\top]$, and let $\mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle \hat{f}, \hat{g}\rangle - \langle f^\star, g^\star\rangle)^2] \le \epsilon^2$. Then,*

(e) Let $\mathbf{\Sigma} = \mathbb{E}[ff^\top]$. Then, $\sum_{i\ge 1} |\sigma_i(\mathbf{\Sigma}) - \boldsymbol{\sigma}_i^\star|^2 \le \epsilon^2$

(f) Fix $k \in \mathbb{N}$, and let $\mathbf{P}_k$ denote the projection onto the top k eigenvectors of $\mathbf{\Sigma}$.[10] Set $\boldsymbol{\delta}_k^\star := 1 - \frac{\boldsymbol{\sigma}^\star_{k+1}}{\boldsymbol{\sigma}^\star_k}$, and suppose that $\boldsymbol{\sigma}_k^\star > 0$. Then, if $\epsilon \le \eta \boldsymbol{\sigma}_k^\star \boldsymbol{\delta}_k^\star$ for a given $\eta \in [0, 1)$. Then

$$\mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle \mathbf{P}_k f, \mathbf{P}_k g\rangle - \langle f_k^\star, g_k^\star\rangle^2)] \le \frac{81\epsilon^2}{(\boldsymbol{\delta}_k^\star(1-\eta))^2}.$$

Our proof follows by approximation to simple functions.

Lemma J.9 *Ssuppose that (f, g) and $(f^\star, g^\star)$ are simple functions embedding into $\mathbb{R}^p$, and balanced under $\mathcal{D}_\otimes = \mathcal{D}_{\mathcal{X}} \otimes \mathcal{D}_{\mathcal{Y}}$, and that ϵ is as in Eq. (J.2). Then*

(a) It holds that $\sum_{i\ge 1} |\sigma_i(\mathbf{\Sigma}) - \sigma_i(\mathbf{\Sigma}^\star)|^2 \le \epsilon^2$.

10. Under the conditions of this statement, it holds that $\mathbf{P}_k$ is unique.

(b) *Suppose* $\sigma_k(\mathbf{\Sigma}^\star) > 0$, *and set* $\delta_k^\star := 1 - \frac{\sigma_{k+1}(\mathbf{\Sigma}^\star)}{\sigma_k(\mathbf{\Sigma}^\star)}$ *and suppose that* $\epsilon \le \eta\sigma_k(\mathbf{\Sigma}^\star)\delta_k^\star$, *where* $\eta \in [0, 1)$. *Then*

$$\epsilon_{\mathrm{pred},k}^2 \le \frac{81\epsilon^2}{(\delta_k^\star(1-\eta))^2}.$$

Proof [Proof of Lemma J.9] For the first point, we have

$$\begin{aligned}
\sum_{i\ge1} |\sigma_i(\mathbf{\Sigma}) - \sigma_i(\mathbf{\Sigma}^\star)|^2 &= \sum_{i\ge1} |\sigma_i(\hat{\mathbf{M}}) - \sigma_i(\mathbf{M}^\star)|^2 && \text{(Lemma J.2, part (a))}\\
&\le \|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{F}}^2 && \text{(Lemma K.3)}\\
&\le \epsilon^2. && \text{(Lemma J.2, part (b))}
\end{aligned}$$

For the second point, we can verify from Lemma J.2 part (c) and the same computation as in the proof of part (b) that

$$\mathbb{E}_{\mathcal{D}_\mathcal{X}\otimes\mathcal{D}_\mathcal{Y}}[(\langle f_k^\star(x), g_k^\star(y)\rangle - \langle f_k(x), g_k(y)\rangle)^2] = \|\mathbf{M}_{[k]}^\star - \hat{\mathbf{M}}_{[k]}\|_{\mathrm{F}}^2, \tag{J.20}$$

where $\mathbf{M}_{[k]}^\star$ and $\hat{\mathbf{M}}_{[k]}$ denote the rank-k SVD approximation of $\mathbf{M}^\star$ and $\hat{\mathbf{M}}$, respectively. We now invoke our main SVD perturbation bound, Theorem 1. This states that if $\sigma_k(\mathbf{M}^\star) > 0$ and $\delta = 1 - \sigma_{k+1}(\mathbf{M}^\star)/\sigma_k(\mathbf{M}^\star) > 0$, and if $\|\mathbf{M}^\star - \hat{\mathbf{M}}\|_{\mathrm{op}} \le \eta\sigma_k(\mathbf{M}^\star)\delta$ for some $\eta \in (0, 1)$, then the Frobenius norm error between the rank-k SVD's of $\mathbf{M}^\star$ and $\hat{\mathbf{M}}$ is bounded by

$$\|\hat{\mathbf{M}}_{[k]} - \mathbf{M}_{[k]}^\star\|_{\mathrm{F}}^2 \le \frac{81\|\hat{\mathbf{M}} - \mathbf{M}^\star\|_{\mathrm{F}}^2}{\delta^2(1-\eta)^2}.$$

Using the correspondences in Lemma J.2, we can take $\delta = \delta_k^\star := 1 - \frac{\sigma_{k+1}(\mathbf{\Sigma}^\star)}{\sigma_k^\star(\mathbf{\Sigma}^\star)}$, and that it is sufficient that $\epsilon \le \eta\sigma_k(\mathbf{\Sigma}^\star)\delta_k^\star$ (since $\sigma_k(\mathbf{M}^\star) = \sigma_k(\mathbf{\Sigma}^\star)$ and $\epsilon = \|\mathbf{M}^\star - \hat{\mathbf{M}}\|_{\mathrm{F}} \ge \|\mathbf{M}^\star - \hat{\mathbf{M}}\|_{\mathrm{op}}$). Using Eq. (J.20) and Lemma J.2 part (b), we conclude that for $\epsilon \le \eta\sigma_k^\star\delta_k^\star$,

$$\mathbb{E}_{\mathcal{D}_\mathcal{X}\otimes\mathcal{D}_\mathcal{Y}}[(\langle f_k^\star(x), g_k^\star(y)\rangle - \langle f_k(x), g_k(y)\rangle)^2] = \|\mathbf{M}_{[k]}^\star - \hat{\mathbf{M}}_{[k]}\|_{\mathrm{F}}^2 \le \frac{81\epsilon^2}{(\delta_k^\star(1-\eta))^2}.$$

This completes the proof. ∎

Next, we remove the requirement of simple functions.

Lemma J.10 *Suppose that* (f, g) *and* $(f^\star, g^\star)$ *are pairs of embeddings into* $\mathbb{R}^p$, *and are balanced under* $\mathcal{D}_\otimes = \mathcal{D}_\mathcal{X} \otimes \mathcal{D}_\mathcal{Y}$, but are not necessarily simple functions, *and that* ϵ *is as in Eq. (J.2). Then*

(a) *It holds that* $\sum_{i\ge1} |\sigma_i(\mathbf{\Sigma}) - \sigma_i(\mathbf{\Sigma}^\star)|^2 \le \epsilon^2$.

(b) *Suppose* $\sigma_k(\mathbf{\Sigma}^\star) > 0$, *and set* $\delta_k^\star := 1 - \frac{\sigma_{k+1}(\mathbf{\Sigma}^\star)}{\sigma_k(\mathbf{\Sigma}^\star)}$ *and suppose that* $\epsilon \le \eta\sigma_k(\mathbf{\Sigma}^\star)\delta_k^\star$, *where* $\eta \in [0, 1)$. *Then*

$$\epsilon_{\mathrm{pred},k}^2 \le \frac{81\epsilon^2}{(\delta_k^\star(1-\eta))^2}.$$

Proof [Proof of Lemma J.10] Let's start with part (a). By invoking Lemma J.9 part (a) for each τ,

$$\sum_{i \geq 1} |\sigma_i(\mathbf{\Sigma}^\star_{[\tau]}) - \sigma_i(\mathbf{\Sigma}_{[\tau]})|^2 \leq \epsilon^2_{[\tau]}.$$

Taking $\tau \to \infty$, Lemma J.8 parts (b) and (c) ensure $\mathbf{\Sigma}^\star_{[\tau]} \to \mathbf{\Sigma}^\star$, $\mathbf{\Sigma}_{[\tau]} \to \mathbf{\Sigma}$. Thus, Weyl's inequality implies that $\lim_{\tau \to \infty} \sum_{i \geq 1} |\sigma_i(\mathbf{\Sigma}^\star_{[\tau]}) - \sigma_i(\mathbf{\Sigma}_{[\tau]})|^2 = \sum_{i \geq 1} |\sigma_i(\mathbf{\Sigma}^\star) - \sigma_i(\mathbf{\Sigma})|^2$. Lemma J.8 part (g) gives $\epsilon^2_{[\tau]} \to \epsilon^2_{\mathrm{pred}} \leq \epsilon^2$, completing the proof of the statement.

Next, let's turn to part (b). Fix a k for which $\sigma_k(\mathbf{\Sigma}^\star) > 0$. From part (a) and the condition that $\epsilon < \sigma_k(\mathbf{\Sigma}^\star)$, it also follows that $\sigma_k(\mathbf{\Sigma}) > 0$. Define $\delta^\star_{k,[\tau]} := 1 - \frac{\sigma_{k+1}(\mathbf{\Sigma}^\star_{[\tau]})}{\sigma_k(\mathbf{\Sigma}^\star_{[\tau]})}$. Lemma J.8 part (b) ensures $\mathbf{\Sigma}^\star_{[\tau]} \to \mathbf{\Sigma}^\star$, so that (again using Weyl's inequality), $\delta^\star_{k,[\tau]}$ is well defined for all τ sufficiently large, and converges to $\delta^\star_k := 1 - \frac{\sigma_{k+1}(\mathbf{\Sigma}^\star)}{\sigma_k(\mathbf{\Sigma}^\star)}$. Using Lemma J.8 again, the assumption that $\epsilon \leq \eta \sigma_k(\mathbf{\Sigma}^\star)\delta^\star_k$ implies that there is a sequence of $\eta_{k,[\tau]} \downarrow \eta$ such that $\epsilon_{k,[\tau]} \leq \eta_{k,[\tau]}\sigma_k(\mathbf{\Sigma}^\star_{[\tau]})\delta^\star_{k,[\tau]}$ for all τ sufficiently large. Invoking Lemma J.9 part (b) for these τ,

$$\epsilon^2_{k,[\tau]} \leq \frac{81\epsilon^2}{(\delta^\star_{k,[\tau]}(1 - \eta_{k,[\tau]}))^2}.$$

Taking limits $\tau \to \infty$ and again calling Lemma J.8 concludes the proof. ■

The proof of Theorem 11 follows by extending Lemma J.10 to infinite dimensional embeddings along the lines of Appendix J.5. Details are similar (though considerably simpler) and are omitted for brevity. ■

Appendix K. Supporting Linear Algebraic Proofs

K.1. Balancing without loss of generality

Proof [Proof of Lemma I.2] Let $\mathbf{A} = \mathbf{U}_A \mathbf{\Sigma}_A \mathbf{V}_A^\top$ and $\mathbf{B} = \mathbf{U}_B \mathbf{\Sigma}_B \mathbf{V}_B^\top$ where $\mathbf{\Sigma}_A, \mathbf{\Sigma}_B$ are diagonal matrices with elements ranked in descending order, and $\mathbf{V}_A, \mathbf{V}_B \in \mathbb{R}^{d \times d}$. Since the $\mathbf{\Sigma}_A, \mathbf{\Sigma}_B$ and $\mathbf{V}_A, \mathbf{V}_B$ can be constructed from any eigen-decomposition of $\mathbf{A}^\top\mathbf{A}$ and $\mathbf{B}^\top\mathbf{B}$, and since both are equal, we have $\mathbf{\Sigma}_A = \mathbf{\Sigma}_B = \mathbf{\Sigma}$ and we can choose the basis $\mathbf{V}_B$ such that $\mathbf{V}_A = \mathbf{V}_B$. Moreover, $\mathbf{V}_A = \mathbf{V}_B = \mathbf{R}$ for some $\mathbf{R} \in \mathbb{O}(d)$, since they are $d \times d$ matrices with orthonormal columns. Hence, $\mathbf{A} = \mathbf{U}_A \mathbf{\Sigma}\mathbf{R}$ and $\mathbf{B} = \mathbf{V}_B \mathbf{\Sigma}\mathbf{R}$ for some $\mathbf{R} \in \mathbb{O}(d)$.

To justify $\mathbf{A}_{[k]} = \mathbf{U}\mathbf{\Sigma}^{\frac{1}{2}}_{[k]}\mathbf{R}$, set $\bar{\mathbf{A}} := \mathbf{A}\mathbf{R}^\top = \mathbf{U}\mathbf{\Sigma}$; defined $\mathbf{B}_{[k]}$ and $\bar{\mathbf{B}}_{[k]}$ similarly. Then, it can be checked that $\bar{\mathbf{P}}_{[k]}$ projects onto the top k eigenspace of $\bar{\mathbf{A}}^\top\bar{\mathbf{A}}$, so that $\mathbf{P}_{[k]} = \mathbf{R}^\top\bar{\mathbf{P}}_{[k]}\mathbf{R}$ projects onto the top k eigenspace of $\mathbf{A}^\top\mathbf{A}$. Hence,

$$\mathbf{A}_{[k]} = \mathbf{A}\mathbf{P}_{[k]} = \bar{\mathbf{A}}\mathbf{R} \cdot \mathbf{R}^\top\bar{\mathbf{P}}_{[k]}\mathbf{R} = \mathbf{U}\mathbf{\Sigma}^{\frac{1}{2}}_{[k]}\mathbf{R},$$

as needed. Similar argument holds for $\mathbf{B}_{[k]}$. ■

Proof [Proof of Lemma I.1] From Lemma I.2, there exists a $\hat{\mathbf{R}} \in \mathbb{O}(d)$ for which $\hat{\mathbf{A}}' = \hat{\mathbf{A}}\hat{\mathbf{R}}, \hat{\mathbf{B}}' = \hat{\mathbf{B}}\hat{\mathbf{R}}$, and a $\mathbf{R}^\star \in \mathbb{O}(d)$ for which $\mathbf{A}^{\star\prime} = \mathbf{A}^\star\mathbf{R}^\star, \mathbf{B}^{\star\prime} = \mathbf{B}^\star\mathbf{R}^\star$, and for which we can take $\mathbf{A}^{\star\prime}_{[k]} =$

$\mathbf{A}^{\star}_{[k]}\mathbf{R}^{\star}, \mathbf{B}^{\star\prime}_{[k]} = \mathbf{B}^{\star}_{[k]}\mathbf{R}^{\star}$. Given $\mathbf{R}' \in \mathbb{O}(d)$, we choose our orthogonal matrix to apply to the non-primed terms as $\mathbf{R} = \hat{\mathbf{R}}\mathbf{R}'(\mathbf{R}^{\star})^{-1} \in \mathbb{O}(d)$. We compute

$$\begin{aligned}
\|(\mathbf{A}^{\star}_{[k]} - \hat{\mathbf{A}}\mathbf{R})(\mathbf{B}^{\star}_{[k]})^{\top}\|_{\mathrm{F}}^2 &= \|(\mathbf{A}^{\star}_{[k]} - \hat{\mathbf{A}}\hat{\mathbf{R}}\mathbf{R}'(\mathbf{R}^{\star})^{-1})(\mathbf{B}^{\star}_{[k]})^{\top}\|_{\mathrm{F}}^2 \\
&= \|(\mathbf{A}^{\star}_{[k]} - \hat{\mathbf{A}}'\mathbf{R}'(\mathbf{R}^{\star})^{-1})(\mathbf{B}^{\star}_{[k]})^{\top}\|_{\mathrm{F}}^2 \\
&= \|(\mathbf{A}^{\star\prime}_{[k]}(\mathbf{R}^{\star})^{-1} - \hat{\mathbf{A}}'\mathbf{R}'(\mathbf{R}^{\star})^{-1})(\mathbf{B}^{\star\prime}_{[k]}(\mathbf{R}^{\star})^{-1})^{\top}\|_{\mathrm{F}}^2 \\
&= \|(\mathbf{A}^{\star\prime}_{[k]} - \hat{\mathbf{A}}'\mathbf{R}')(\mathbf{R}^{\star})^{-1}(\mathbf{R}^{\star})^{-\top}(\mathbf{B}^{\star\prime}_{[k]})^{\top}\|_{\mathrm{F}}^2 \\
&= \|(\mathbf{A}^{\star\prime}_{[k]} - \hat{\mathbf{A}}'\mathbf{R}')(\mathbf{B}^{\star\prime}_{[k]})^{\top}\|_{\mathrm{F}}^2,
\end{aligned}$$

where the second-last line uses that $(\mathbf{R}^{\star})^{-1}(\mathbf{R}^{\star})^{-\top} = \mathbf{I}_p$ for $\mathbf{R}^{\star} \in \mathbb{O}(d)$. The equality $\|\mathbf{A}^{\star}_{[k]}(\mathbf{B}^{\star}_{[k]} - \hat{\mathbf{B}}\mathbf{R})^{\top}\|_{\mathrm{F}}^2 = \|\mathbf{A}^{\star\prime}_{[k]}(\mathbf{B}^{\star\prime}_{[k]} - \hat{\mathbf{B}}'\mathbf{R}')^{\top}\|_{\mathrm{F}}^2$ can be verified similarly.

Moreover,

$$\begin{aligned}
\|\mathbf{A}^{\star}_{[k]} - \hat{\mathbf{A}}\mathbf{R}\|_{\mathrm{F}}^2 &= \|\mathbf{A}^{\star\prime}_{[k]}(\mathbf{R}^{\star})^{-1} - \hat{\mathbf{A}}'\hat{\mathbf{R}}^{-1} \cdot \hat{\mathbf{R}}\mathbf{R}'(\mathbf{R}^{\star})^{-1}\|_{\mathrm{F}}^2 \\
&= \|(\mathbf{A}^{\star\prime}_{[k]} - \hat{\mathbf{A}}'\mathbf{R}')(\mathbf{R}^{\star})^{-1}\|_{\mathrm{F}}^2 = \|\mathbf{A}^{\star\prime}_{[k]} - \hat{\mathbf{A}}'\mathbf{R}'\|_{\mathrm{F}}^2,
\end{aligned}$$

where the last line uses the unitary invariant property of the Frobenius norm. The equality $\|\mathbf{B}^{\star}_{[k]} - \hat{\mathbf{B}}\mathbf{R}\|_{\mathrm{F}}^2 = \|\mathbf{B}^{\star\prime}_{[k]} - \hat{\mathbf{B}}'\mathbf{R}'\|_{\mathrm{F}}^2$ follows similarly.

Lastly, since $\mathbf{R} = \hat{\mathbf{R}}\mathbf{R}'(\mathbf{R}^{\star})^{-1}$, we have $\mathbf{R}' = \hat{\mathbf{R}}^{-1}\mathbf{R}\mathbf{R}^{\star}$. Then

$$\begin{aligned}
\mathrm{rowspace}\left((\hat{\mathbf{A}}'_{[k]}\mathbf{R}')^{\top}(\hat{\mathbf{A}}'_{[k]}\mathbf{R}')\right) &= \mathrm{rowspace}\left((\hat{\mathbf{A}}'_{[k]}\hat{\mathbf{R}}^{-1}\mathbf{R}\mathbf{R}^{\star})^{\top}(\hat{\mathbf{A}}'_{[k]}\hat{\mathbf{R}}^{-1}\mathbf{R}\mathbf{R}^{\star})\right) \\
&= \mathrm{rowspace}\left((\hat{\mathbf{A}}_{[k]}\mathbf{R}\mathbf{R}^{\star})^{\top}(\hat{\mathbf{A}}_{[k]}\mathbf{R}\mathbf{R}^{\star})\right) \\
&= \mathrm{rowspace}\left((\mathbf{R}^{\star})^{\top}(\hat{\mathbf{A}}_{[k]}\mathbf{R})^{\top}(\hat{\mathbf{A}}_{[k]}\mathbf{R})\mathbf{R}^{\star}\right) \\
&\overset{(i)}{\supseteq} \mathrm{rowspace}\left((\mathbf{R}^{\star})^{\top}(\mathbf{A}^{\star}_{[k]})^{\top}(\mathbf{A}^{\star}_{[k]})\mathbf{R}^{\star}\right) \\
&= \mathrm{rowspace}\left((\mathbf{A}^{\star}_{[k]}\mathbf{R}^{\star})^{\top}(\mathbf{A}^{\star}_{[k]}\mathbf{R}^{\star})\right) \\
&= \mathrm{rowspace}\left((\mathbf{A}^{\star\prime}_{[k]})^{\top}(\mathbf{A}^{\star\prime}_{[k]})\right),
\end{aligned}$$

where in (i), we use that $\mathrm{rowspace}\left((\hat{\mathbf{A}}_{[k]}\mathbf{R})^{\top}(\hat{\mathbf{A}}_{[k]}\mathbf{R})\right) \supseteq \mathrm{rowspace}\left((\mathbf{A}^{\star}_{[k]})^{\top}(\mathbf{A}^{\star}_{[k]})\right)$, and that $\mathbf{R}$ is a rotation matrix. ■

K.2. Supporting proofs for error decomposition

K.2.1. PROOF OF LEMMA I.3

We first state the following facts.

Fact K.1 *Let $\mathcal{K} := (\mathcal{K}_i)_{i=1}^{\ell}$ be a monotone partition of $[d]$ for some $d \leq \min\{n, m\}$, and consider $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}$ and $\mathbf{B}' = \mathbf{V}'\mathbf{\Sigma}'$. Then,*

(a) *Any diagonal matrix* $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$ *is compatible with* $\mathcal{K}$. *Hence,* $\mathbf{A} = \sum_{i=1}^{\ell} \mathbf{A}_{\mathcal{K}_i}$, *and similar for* $\mathbf{B}'$.

(b) *If* $\mathbf{R} \in \mathbb{R}^{d\times d}$ *is compatibile with* $\mathcal{K}$, *then* $\mathbf{A}\mathbf{R} = \sum_{i=1}^{\ell} \mathbf{A}_{\mathcal{K}_i}\mathbf{R}_{\mathcal{K}_i}$.

(c) *For any* $1 \le i \ne j \le \ell$, $\mathbf{\Sigma}_{\mathcal{K}_i}\mathbf{\Sigma}'_{\mathcal{K}_j} = 0$. *Hence,* $\mathbf{A}_{\mathcal{K}_i}(\mathbf{B}'_{\mathcal{K}_j})^\top = 0$, *and also* $(\mathbf{A}_{\mathcal{K}_i}\mathbf{R}_{\mathcal{K}_i})(\mathbf{B}'_{\mathcal{K}_j})^\top = 0$.

(d) $\mathbf{A}_{\mathcal{K}_i}(\mathbf{B}')^\top = (\mathbf{A}_{\mathcal{K}_i})(\mathbf{B}'_{\mathcal{K}_i})^\top$ *for any* $1 \le i \le \ell$.

If in addition $(\mathcal{K}_i)_{i=1}^{\ell}$ *is a monotone partition with pivots* $0 = k_1 < k_2 \cdots < k_\ell < k_{\ell+1} = p$, *then* $\mathbf{A}_{[k_\ell]} = \sum_{i=1}^{\ell-1} \mathbf{A}_{\mathcal{K}_i}$, *and similarly for* $\mathbf{B}'$.

Proof We use the facts above to prove Lemma I.3:

$$\begin{aligned}
(\mathbf{A}^\star_{[k_\ell]} - \hat{\mathbf{A}}\mathbf{R})(\mathbf{B}^\star_{[k_\ell]})^\top &= \left((\sum_{i=1}^{\ell-1} \mathbf{A}^\star_{\mathcal{K}_i}) - (\sum_{i=1}^{\ell} \hat{\mathbf{A}}_{\mathcal{K}_i}\mathbf{R}_{\mathcal{K}_i}) \right) (\mathbf{B}^\star_{[k_\ell]})^\top \\
&= \sum_{i=1}^{\ell-1} \left(\mathbf{A}^\star_{\mathcal{K}_i} - \hat{\mathbf{A}}_{\mathcal{K}_i}\mathbf{R}_{\mathcal{K}_i} \right) (\mathbf{B}^\star_{[k_\ell]})^\top - (\hat{\mathbf{A}}_{\mathcal{K}_\ell}\mathbf{R}_{\mathcal{K}_\ell})(\mathbf{B}^\star_{[k_\ell]})^\top \\
&= \sum_{j=1}^{\ell-1} \sum_{i=1}^{\ell-1} \left(\mathbf{A}^\star_{\mathcal{K}_i} - \hat{\mathbf{A}}_{\mathcal{K}_i}\mathbf{R}_{\mathcal{K}_i} \right) (\mathbf{B}^\star_{\mathcal{K}_j})^\top - \underbrace{\sum_{j=1}^{\ell-1} (\hat{\mathbf{A}}_{\mathcal{K}_\ell}\mathbf{R}_{\mathcal{K}_\ell})(\mathbf{B}^\star_{\mathcal{K}_j})^\top}_{=0} \\
&= \sum_{i=1}^{\ell-1} \left(\mathbf{A}^\star_{\mathcal{K}_i} - \hat{\mathbf{A}}_{\mathcal{K}_i}\mathbf{R}_{\mathcal{K}_i} \right) (\mathbf{B}^\star_{\mathcal{K}_i})^\top.
\end{aligned}$$

Proofs for the compositions of $(\mathbf{A}^\star_{[k_\ell]} - \hat{\mathbf{A}}\mathbf{R})$, $(\mathbf{B}^\star_{[k_\ell]} - \hat{\mathbf{B}}\mathbf{R})$ and $\mathbf{A}^\star_{[k_\ell]}(\mathbf{B}^\star_{[k_\ell]} - \hat{\mathbf{B}}\mathbf{R})^\top$ follow analogously. ■

K.2.2. PROOF OF LEMMA I.4

Proof [Proof of Lemma I.4] For simplicity, we abbreviate $E_i = E_i(\mathbf{R})$. Let us prove the bound on $\Delta_0(\mathbf{R}, k_\ell)$ first. Recall $\Delta_0(\mathbf{R}, k_\ell) = \|(\mathbf{A}^\star_{[k_\ell]} - \hat{\mathbf{A}}\mathbf{R})(\mathbf{B}^\star_{[k_\ell]})^\top\|_{\mathrm{F}}^2 \vee \|\mathbf{A}^\star_{[k_\ell]}(\mathbf{B}^\star_{[k_\ell]} - \hat{\mathbf{B}}\mathbf{R})^\top\|_{\mathrm{F}}^2$. We explicitly bound the first term $\|(\mathbf{A}^\star_{[k_\ell]} - \hat{\mathbf{A}}\mathbf{R})(\mathbf{B}^\star_{[k_\ell]})^\top\|_{\mathrm{F}}^2$, and note that a similar argument bounds the second term.

Invoking Eq. (I.7a)

$$\begin{aligned}
\|(\mathbf{A}^\star_{[k_\ell]} - \hat{\mathbf{A}}\mathbf{R})(\mathbf{B}^\star_{[k_\ell]})^\top\|_{\mathrm{F}}^2 &= \left\| \sum_{i=1}^{\ell-1} \left(\mathbf{A}^\star_{\mathcal{K}_i} - \hat{\mathbf{A}}_{\mathcal{K}_i}\mathbf{R}_{\mathcal{K}_i} \right) (\mathbf{B}^\star_{\mathcal{K}_i})^\top \right\|_{\mathrm{F}}^2 \\
&= \sum_{i=1}^{\ell-1} \left\| \left(\mathbf{A}^\star_{\mathcal{K}_i} - \hat{\mathbf{A}}_{\mathcal{K}_i}\mathbf{R}_{\mathcal{K}_i} \right) (\mathbf{B}^\star_{\mathcal{K}_i})^\top \right\|_{\mathrm{F}}^2 \\
&\quad + \sum_{i,j=1, i\ne j}^{\ell-1} \underbrace{\left\langle \left(\mathbf{A}^\star_{\mathcal{K}_i} - \hat{\mathbf{A}}_{\mathcal{K}_i}\mathbf{R}_{\mathcal{K}_i} \right) (\mathbf{B}^\star_{\mathcal{K}_i})^\top, \left(\mathbf{A}^\star_{\mathcal{K}_j} - \hat{\mathbf{A}}_{\mathcal{K}_j}\mathbf{R}_{\mathcal{K}_j} \right) (\mathbf{B}^\star_{\mathcal{K}_j})^\top \right\rangle}_{=0}.
\end{aligned}$$

Here, the second term vanishes because $(\mathbf{B}^\star_{\mathcal{K}_i})^\top \mathbf{B}^\star_{\mathcal{K}_j} = 0 = (\mathbf{\Sigma}^\star_{\mathcal{K}_i})^{\frac{1}{2}}(\mathbf{V}^\star)^\top \mathbf{V}^\star (\mathbf{\Sigma}^\star_{\mathcal{K}_j})^{\frac{1}{2}} = (\mathbf{\Sigma}^\star_{\mathcal{K}_i})^{\frac{1}{2}}(\mathbf{\Sigma}^\star_{\mathcal{K}_j})^{\frac{1}{2}} = 0$, since $\mathcal{K}_i \cap \mathcal{K}_j = \emptyset$ for $i \neq j$. Continuing,

$$
\begin{aligned}
\sum_{i=1}^{\ell-1} \left\| \left(\mathbf{A}^\star_{\mathcal{K}_i} - \hat{\mathbf{A}}_{\mathcal{K}_i}\mathbf{R}_{\mathcal{K}_i}\right)(\mathbf{B}^\star_{\mathcal{K}_i})^\top \right\|_{\mathrm{F}}^2 &\le \sum_{i=1}^{\ell-1} \|\mathbf{B}^\star_{\mathcal{K}_i}\|_{\mathrm{op}}^2 \|\mathbf{A}^\star_{\mathcal{K}_i} - \hat{\mathbf{A}}_{\mathcal{K}_i}\mathbf{R}_{\mathcal{K}_i}\|^2 \\
&= \sum_{i=1}^{\ell-1} \max\{\sigma^\star_{k'} : k' \in \mathcal{K}_i\} \cdot \|\mathbf{A}^\star_{\mathcal{K}_i} - \hat{\mathbf{A}}_{\mathcal{K}_i}\mathbf{R}_{\mathcal{K}_i}\|^2 \\
&\le \sum_{i=1}^{\ell-1} (\mu\sigma^\star_{k_{i+1}}) \cdot \|\mathbf{A}^\star_{\mathcal{K}_i} - \hat{\mathbf{A}}_{\mathcal{K}_i}\mathbf{R}_{\mathcal{K}_i}\|^2 \quad ((\delta,\mu)\text{-well-tempered}) \\
&= \mu \sum_{i=1}^{\ell-1} (\delta^\star_{k_i} \wedge \delta^\star_{k_{i+1}})^{-2} \cdot \underbrace{(\delta^\star_{k_i} \wedge \delta^\star_{k_{i+1}})^2 (\sigma^\star_{k_{i+1}}) \|\mathbf{A}^\star_{\mathcal{K}_i} - \hat{\mathbf{A}}_{\mathcal{K}_i}\mathbf{R}_{\mathcal{K}_i}\|^2}_{\le E_i} \\
&\le \mu \Big(\sum_{i=1}^{\ell-1} (\delta^\star_{k_i} \wedge \delta^\star_{k_{i+1}})^{-2}\Big) \cdot \max_{i \in [\ell-1]} E_i \\
&\le \mu \Big(\sum_{i=1}^{\ell-1} (\delta^\star_{k_i})^{-2} + (\delta^\star_{k_{i+1}})^{-2}\Big) \cdot \max_{i \in [\ell-1]} E_i \\
&= \mu \Big(\sum_{i=1}^{\ell-1} (\delta^\star_{k_i})^{-2} + \sum_{i=2}^{\ell} (\delta^\star_{k_i})^{-2}\Big) \cdot \max_{i \in [\ell-1]} E_i.
\end{aligned}
$$

By convention, $\delta^\star_0 = 1 \ge \delta^\star_{k_i}$ (all relative gaps are at most 1). Thus, $\sum_{i=2}^{\ell}(\delta^\star_{k_i})^{-2} \le \sum_{i=1}^{\ell}(\delta^\star_{k_i})^{-2}$, and the above is at most $2\mu(\sum_{i=1}^{\ell}(\delta^\star_{k_i})^{-2}) \max_{i\in[\ell]} E_i = 2\mu M_{\mathrm{space}} \max_{i\in[\ell]} E_i$. This completes the proof of the first argument.

Let's now turn to $\Delta_1(\mathbf{R}, k_\ell) = \|\mathbf{A}^\star_{[k_\ell]} - \hat{\mathbf{A}}\mathbf{R}\|_{\mathrm{F}}^2 \vee \|\mathbf{B}^\star_{[k_\ell]} - \hat{\mathbf{B}}\mathbf{R}\|_{\mathrm{F}}^2$. Again, we focus on $\|\mathbf{A}^\star_{[k_\ell]} - \hat{\mathbf{A}}\mathbf{R}\|_{\mathrm{F}}^2$. From Eq. (I.7b)

$$
\begin{aligned}
\|\mathbf{A}^\star_{[k_\ell]} - \hat{\mathbf{A}}\mathbf{R}\|_{\mathrm{F}}^2 &= \left\| \sum_{i=1}^{\ell-1} \left(\mathbf{A}^\star_{\mathcal{K}_i} - \hat{\mathbf{A}}_{\mathcal{K}_i}\mathbf{R}_{\mathcal{K}_i}\right) - \hat{\mathbf{A}}_{>k_\ell}\mathbf{R}_{>k_\ell} \right\|_{\mathrm{F}}^2 \\
&= \sum_{i=1}^{\ell-1} \left\| \mathbf{A}^\star_{\mathcal{K}_i} - \hat{\mathbf{A}}_{\mathcal{K}_i}\mathbf{R}_{\mathcal{K}_i} \right\|_{\mathrm{F}}^2 + \|\hat{\mathbf{A}}_{>k_\ell}\mathbf{R}_{>k_\ell}\|_{\mathrm{F}}^2 \\
&\quad + \sum_{i\neq j}^{\ell-1} \underbrace{\left\langle \mathbf{A}^\star_{\mathcal{K}_i} - \hat{\mathbf{A}}_{\mathcal{K}_i}\mathbf{R}_{\mathcal{K}_i}, \mathbf{A}^\star_{\mathcal{K}_j} - \hat{\mathbf{A}}_{\mathcal{K}_j}\mathbf{R}_{\mathcal{K}_j} \right\rangle}_{=0} - \sum_{i=1}^{\ell-1} \underbrace{\left\langle \mathbf{A}^\star_{\mathcal{K}_i} - \hat{\mathbf{A}}_{\mathcal{K}_i}\mathbf{R}_{\mathcal{K}_i}, \hat{\mathbf{A}}_{>k_\ell}\mathbf{R}_{>k_\ell} \right\rangle}_{=0},
\end{aligned}
$$

where the terms on the second line vanish because they involve inner products of matrices whose columns have disjoint support. We bound

$$\sum_{i=1}^{\ell-1} \left\|\mathbf{A}^\star_{\mathcal{K}_i} - \hat{\mathbf{A}}_{\mathcal{K}_i}\mathbf{R}_{\mathcal{K}_i}\right\|_{\mathrm{F}}^2 = \sum_{i=1}^{\ell-1} \frac{1}{(\delta^\star_{k_i} \wedge \delta^\star_{k_{i+1}})^2(\sigma^\star_{k_{i+1}})} \cdot \underbrace{(\sigma^\star_{k_{i+1}})(\delta^\star_{k_i} \wedge \delta^\star_{k_{i+1}})^2 \left\|\mathbf{A}^\star_{\mathcal{K}_i} - \hat{\mathbf{A}}_{\mathcal{K}_i}\mathbf{R}_{\mathcal{K}_i}\right\|_{\mathrm{F}}^2}_{\le E_i}$$

$$\le \delta^{-2}\sum_{i=1}^{\ell-1} \frac{1}{\sigma^\star_{k_{i+1}}} E_i \qquad \text{(partition is } (\delta,\mu)\text{-well-tempered)}$$

$$\le \delta^{-2} \max_{i\in[\ell-1]} E_i \cdot \sum_{i=1}^{\ell-1}\frac{1}{\sigma^\star_{k_{i+1}}} = \frac{M_{\mathrm{spec}}}{\delta^2}\max_{i\in[\ell-1]} E_i.$$

Finally, using $\mathbf{R} \in \mathbb{O}(d)$ and $\hat{\mathbf{A}}_{>k_\ell} = \hat{\mathbf{U}}\hat{\mathbf{\Sigma}}^{\frac{1}{2}}_{>k_\ell}$ for $\|\hat{\mathbf{U}}\|_{\mathrm{op}} = 1$, we have that

$$\|\hat{\mathbf{A}}_{>k_\ell}\mathbf{R}_{>k_\ell}\|_{\mathrm{F}}^2 \le \|\hat{\mathbf{A}}_{>k_\ell}\|_{\mathrm{F}}^2 \le \|\hat{\mathbf{\Sigma}}^{\frac{1}{2}}_{>k_\ell}\|_{\mathrm{F}}^2 = \sum_{i>k_\ell}\|\hat{\mathbf{\Sigma}}_i^{\frac{1}{2}}\|_{\mathrm{F}}^2 = \sum_{i>k_\ell}\sigma_i(\hat{\mathbf{M}}).$$

In sum,

$$\begin{aligned}\|\mathbf{A}^\star_{[k_\ell]} - \hat{\mathbf{A}}\mathbf{R}\|_{\mathrm{F}}^2 &= \left\|\sum_{i=1}^{\ell-1}\left(\mathbf{A}^\star_{\mathcal{K}_i} - \hat{\mathbf{A}}_{\mathcal{K}_i}\mathbf{R}_{\mathcal{K}_i}\right) - \hat{\mathbf{A}}_{>k_\ell}\mathbf{R}_{>k_\ell}\right\|_{\mathrm{F}}^2 \\ &= \sum_{i=1}^{\ell-1}\left\|\mathbf{A}^\star_{\mathcal{K}_i} - \hat{\mathbf{A}}_{\mathcal{K}_i}\mathbf{R}_{\mathcal{K}_i}\right\|_{\mathrm{F}}^2 + \|\hat{\mathbf{A}}_{>k_\ell}\mathbf{R}_{>k_\ell}\|_{\mathrm{F}}^2 \\ &\le \frac{M_{\mathrm{spec}}}{\delta^2}\max_{i\in[\ell-1]} E_i + \sum_{i>k_\ell}\sigma_i(\hat{\mathbf{M}}),\end{aligned}$$

as needed. ■

K.2.3. PROOF OF LEMMA I.6

Proof [Proof of Lemma I.6] We observe that for any monotone (in particular, well-tempered) partition,

$$\hat{\mathbf{M}}_{\mathcal{K}_i} = \hat{\mathbf{M}}_{[k_{i+1}]} - \hat{\mathbf{M}}_{[k_i]}, \qquad \mathbf{M}^\star_{\mathcal{K}_i} = \mathbf{M}^\star_{[k_{i+1}]} - \mathbf{M}^\star_{[k_i]},$$

where we let $\mathbf{M}^\star_{[k_1]} = \mathbf{M}^\star_{[0]} = 0$. Thus, $\circ = \{\mathrm{op}, \mathrm{F}\}$,

$$\begin{aligned}\max_{i\in[\ell]}(\delta^\star_{k_i} \wedge \delta^\star_{k_{i+1}})\|\hat{\mathbf{M}}_{\mathcal{K}_i} - \mathbf{M}^\star_{\mathcal{K}_i}\|_\circ &\le \max_{i\in[\ell]} \delta^\star_{k_i}\|\hat{\mathbf{M}}_{[k_i]} - \mathbf{M}^\star_{[k_i]}\|_\circ + \delta^\star_{k_{i+1}}\|\hat{\mathbf{M}}_{[k_{i+1}]} - \mathbf{M}^\star_{[k_{i+1}]}\|_\circ \\ &\le 2\max_{i\in[\ell+1]}\delta^\star_{k_i}\|\hat{\mathbf{M}}_{[k_i]} - \mathbf{M}^\star_{[k_i]}\|_\circ =: 2\tilde{\epsilon}_\circ.\end{aligned}$$

Next, for a matrix of the form $\mathbf{A}_{\mathcal{K}_i} \in \mathbb{R}^{n\times d}$, let $\mathbf{A}_{\langle\mathcal{K}_i\rangle} \in \mathbb{R}^{n\times|\mathcal{K}_i|}$ denote its canonical compact representation. We observe then that

$$\hat{\mathbf{A}}_{\langle\mathcal{K}_i\rangle}\hat{\mathbf{B}}^\top_{\langle\mathcal{K}_i\rangle} = \hat{\mathbf{M}}_{\mathcal{K}_i}, \qquad \mathbf{A}^\star_{\langle\mathcal{K}_i\rangle}(\mathbf{B}^\star_{\langle\mathcal{K}_i\rangle})^\top = \mathbf{M}^\star_{\mathcal{K}_i}.$$

Further, observe that $\sigma_{|\mathcal{K}_i|}(\mathbf{M}^\star_{\mathcal{K}_i}) = \min\{\sigma^\star_{k'} : k \in \mathcal{K}_i\} = \sigma^\star_{k_{i+1}}$. Hence, Lemma I.5 implies the following: for a given $i \in [\ell]$, if $2\tilde{\epsilon}_{\mathrm{op}} \le \delta\frac{\sigma^\star_{k_\ell}}{2} \le \frac{(\delta^\star_{k_i} \wedge \delta^\star_{k_{i+1}})\sigma^\star_{k_{i+1}}}{2}$, then there exists an orthogonal matrix $\mathbf{O}_i \in \mathbb{O}(|\mathcal{K}_i|)$ such that

$$\begin{aligned}\|\hat{\mathbf{A}}_{\langle\mathcal{K}_i\rangle}\mathbf{O}_i - \mathbf{A}^\star_{\langle\mathcal{K}_i\rangle}\|_{\mathrm{F}}^2 + \|\hat{\mathbf{B}}_{\langle\mathcal{K}_i\rangle}\mathbf{O}_i - \mathbf{B}^\star_{\langle\mathcal{K}_i\rangle}\|_{\mathrm{F}}^2 &\le \frac{c_0}{\sigma_{|\mathcal{K}_i|}(\mathbf{M}^\star_{\mathcal{K}_i})}\|\hat{\mathbf{M}}_{\mathcal{K}_i} - \mathbf{M}^\star_{\mathcal{K}_i}\|_{\mathrm{F}}^2 \\ &= \frac{c_0}{\sigma^\star_{k_{i+1}}}\|\hat{\mathbf{M}}_{\mathcal{K}_i} - \mathbf{M}^\star_{\mathcal{K}_i}\|_{\mathrm{F}}^2.\end{aligned}$$

Multiplying both sides of the above inequality by $(\delta^\star_{k_i} \wedge \delta^\star_{k_{i+1}})^2\sigma^\star_{k_{i+1}}$, there exists a $\mathbf{O}_i \in \mathbb{O}(|\mathcal{K}_i|)$ such that

$$\begin{aligned}(\delta^\star_{k_i} \wedge \delta^\star_{k_{i+1}})^2\sigma^\star_{k_{i+1}}\left(\|\hat{\mathbf{A}}_{\langle\mathcal{K}_i\rangle}\mathbf{O}_i - \mathbf{A}^\star_{\langle\mathcal{K}_i\rangle}\|_{\mathrm{F}}^2 \vee \|\hat{\mathbf{B}}_{\langle\mathcal{K}_i\rangle}\mathbf{O}_i - \mathbf{B}^\star_{\langle\mathcal{K}_i\rangle}\|_{\mathrm{F}}^2\right) &\le c_0(\delta^\star_{k_i} \wedge \delta^\star_{k_{i+1}})^2\|\hat{\mathbf{M}}_{\mathcal{K}_i} - \mathbf{M}^\star_{\mathcal{K}_i}\|_{\mathrm{F}}^2 \\ &\le 4c_0\tilde{\epsilon}_{\mathrm{fro}}^2.\end{aligned}$$

Now, let $\mathbf{R}$ be the block matrix compatible with $(\mathcal{K}_i)_{i=1}^\ell$, such that $\mathbf{R}_{\langle\mathcal{K}_i\rangle} = \mathbf{O}_i$ (that is, the block of $\mathbf{R}$ corresponding to the set $\mathcal{K}_i$ is the matrix $\mathbf{O}_i$). Since $\mathbf{R}$ is a block-orthogonal matrix, it is orthogonal. Moreover, it is straightforward that

$$\|\hat{\mathbf{A}}_{\langle\mathcal{K}_i\rangle}\mathbf{O}_i - \mathbf{A}^\star_{\langle\mathcal{K}_i\rangle}\|_{\mathrm{F}} = \|\hat{\mathbf{A}}_{\mathcal{K}_i}\mathbf{R}_{\mathcal{K}_i} - \mathbf{A}^\star_{\mathcal{K}_i}\|_{\mathrm{F}},$$

and analogously for the "$\mathbf{B}$"-factors. Hence, for all $i \in [\ell]$

$$E_i = (\delta^\star_{k_i} \wedge \delta^\star_{k_{i+1}})^2\sigma^\star_{k_{i+1}}\left(\|\hat{\mathbf{A}}_{\mathcal{K}_i}\mathbf{R}_{\mathcal{K}_i} - \mathbf{A}^\star_{\mathcal{K}_i}\|_{\mathrm{F}}^2 \vee \|\hat{\mathbf{B}}_{\mathcal{K}_i}\mathbf{R}_{\mathcal{K}_i} - \mathbf{B}^\star_{\mathcal{K}_i}\|_{\mathrm{F}}^2\right) \le 4c_0\tilde{\epsilon}_{\mathrm{fro}}^2.$$

This completes the proof. ■

K.3. Existence of well-tempered partition

K.3.1. PROOF OF LEMMA I.7

We first restate the lemma as below.

Lemma I.7 (Singular Value Spacing) *Fix any $s \in \mathbb{N}$ and $\sigma \in [\sigma^\star_s, \sigma^\star_1]$. Then, there exists integer $\ell \in \mathbb{N}$, and an increasing sequence $0 = k_1 < k_2 \cdots < k_\ell < k_{\ell+1} = s$ such that the following is true:*

(a) For $i \in [\ell]$, $\delta^\star_{k_i} \ge 1/k_{i+1} \ge 1/s$.

(b) For $i = \ell$, $\sigma^\star_{k_i+1} \le 2e\sigma$, and for $i \in [\ell - 1]$, $\sigma^\star_{k_i+1} \le 2e^2\sigma^\star_{k_{i+1}}$.

(c) For $i = \ell$, $\sigma^\star_{k_i} \ge \sigma$, and for $i \in [\ell - 1]$, $\sigma^\star_{k_i} \ge e\sigma^\star_{k_{i+1}}$.

Define

$$\tilde{k}_1 = \begin{cases}\max\left\{k' < s : \delta^\star_{k'} \ge \frac{1}{s}, \sigma^\star_{k'} \ge \sigma\right\} & \text{if such a } k' \ge 0 \text{ exists} \\ 0 & \text{otherwise.}\end{cases}$$

And for $i \geq 1$, define

$$\tilde{k}_{i+1} = \begin{cases} \max\left\{k' < \tilde{k}_i : \delta^\star_{k'} \geq \frac{1}{\tilde{k}_i}, \sigma^\star_{k'} \geq e\sigma^\star_{\tilde{k}_i}\right\} & \text{if such a } k' \geq 0 \text{ exists} \\ 0 & \text{otherwise.} \end{cases}$$

We terminate this recursive definition of $\tilde{k}_i$ the first time there is some $\tilde{k}_i = 0$. Thus, let $\ell := \{i \geq 1 : \tilde{k}_i = 0\}$ (which is a unique index). Finally, for $i \in [\ell]$, choose $k_i = \tilde{k}_{\ell+1-i}$.

We now verify this sequence satisfies the desired properties. Items **(a)** and **(c)** clear from the definition.

Item (b). We mainly prove the argument that for $i \in [\ell - 1]$, $\sigma^\star_{k_i+1} \leq 2e^2\sigma^\star_{k_{i+1}}$.

We may assume $k_{i+1} \neq 1$, since otherwise $k_i = 0 = k_1$, i.e. $i = 1$, and $k_i + 1 = 0 + 1 = 1 = k_{i+1}$, and the bound is vacuous. Continuing, fix an index i, and let $\bar{k}_i := \max\{k' < k_{i+1} : \sigma^\star_{k'} \geq e\sigma^\star_{k_{i+1}}\}$. Notice that in particular $\sigma^\star_{\bar{k}_i+1} < e\sigma^\star_{k_{i+1}}$. Then, we can equivalently express

$$k_i = \begin{cases} \max\left\{k' < \bar{k}_i + 1 : \delta^\star_{k'} \geq \frac{1}{k_{i+1}}\right\} & \text{if such a } k' \geq 0 \text{ exists} \\ 0 & \text{otherwise.} \end{cases}$$

In particular, if $k_i + 1 > \bar{k}_i$, then $k_i = \bar{k}_i$, and thus $\sigma^\star_{k_i+1} = \sigma^\star_{\bar{k}_i+1} < e\sigma^\star_{k_{i+1}} \leq 2e^2\sigma^\star_{k_{i+1}}$, we are finished.

Otherwise, if $k_i + 1 \leq \bar{k}_i$, we know that for all

$$\forall j \in [k_i + 1, \bar{k}_i], \quad \delta^\star_j \leq \frac{1}{k_{i+1}}. \tag{K.1}$$

Hence,

$$\begin{aligned}
\sigma^\star_{k_i+1} &= \sigma^\star_{\bar{k}_i+1}\left(\prod_{j=k_i+1}^{\bar{k}_i} \frac{\sigma^\star_j}{\sigma^\star_{j+1}}\right) \\
&= \sigma^\star_{\bar{k}_i+1}\left(\prod_{j=k_i+1}^{\bar{k}_i} \frac{1}{1-\delta^\star_j}\right) && \left(\tfrac{\sigma^\star_{j+1}}{\sigma^\star_j} = 1 - \delta^\star_j\right) \\
&\leq \sigma^\star_{\bar{k}_i+1}\left(\prod_{j=k_i+1}^{\bar{k}_i} \frac{1}{1-1/k_{i+1}}\right) && \text{(Eq. (K.1))} \\
&\leq \sigma^\star_{\bar{k}_i+1} \cdot (1 - 1/k_{i+1})^{-\left(\bar{k}_i - (k_i+1)\right)} \\
&\leq \sigma^\star_{\bar{k}_i+1} \cdot (1 - 1/k_{i+1})^{-k_{i+1}} \leq e\sigma^\star_{k_{i+1}} \cdot (1 - 1/k_{i+1})^{-k_{i+1}}. && (\sigma^\star_{\bar{k}_i+1} \leq e\sigma^\star_{k_{i+1}})
\end{aligned}$$

Using the elementary inequality $(1 - \frac{1}{n})^n \geq e^{-1}(1 - \frac{1}{n})$ for $n \geq 1$, and the fact that $k_{i+1} \geq 2$, we obtain that $(1 - 1/k_{i+1})^{-k_{i+1}} \leq 2e$. Hence, $\sigma^\star_{k_i+1} \leq 2e^2\sigma^\star_{k_{i+1}}$.

Proof for the argument that $\sigma^\star_{k_\ell+1} \leq 2e\sigma$ is nearly identical, by introducing notation $\bar{k}_\ell$, defined as $\bar{k}_\ell := \max\{k' < s : \sigma^\star_{k'} \geq \sigma\}$ and noticing that k_ℓ can be equivalently expressed as

$$k_\ell = \begin{cases} \max\left\{k' < \bar{k}_\ell + 1 : \delta^\star_{k'} \geq \frac{1}{s}\right\} & \text{if such a } k' \geq 0 \text{ exists} \\ 0 & \text{otherwise,} \end{cases}$$

and the rest of the proof follows the same argument. ■

K.3.2. PROOF OF PROPOSITION I.2

We let $(\mathcal{K}_i)_{i=1}^{\ell}$ denote the partition whose pivots are given by the points in Lemma I.7.

Item (a). From item's (a) and (b) of Lemma I.7, the partition is (δ, μ)-well-tempered for $\delta \geq 1/s$ and $\mu \leq 2e^2$.

Item (b). $\sigma^\star_{k_\ell} \geq \sigma$ follows from Lemma I.7, part (c). From that same lemma, we also see that for $i \in [\ell]$, $\sigma^\star_{k_i} \geq e^{\ell-i}\sigma^\star_{k_\ell} \geq e^{\ell-i}\sigma$. Hence

$$M_{\mathrm{spec}} = \sum_{i=1}^{\ell} (\sigma^\star_{k_i})^{-1} \leq \sum_{i=1}^{\ell} e^{-(\ell-i)}\sigma^{-1} \leq \sigma^{-1}\sum_{i\geq 0} e^{-i} = \frac{\sigma^{-1}}{1-e^{-1}}.$$

Item (c). Finally, we develop bounds on M_{space}. We bound

$$M_{\mathrm{space}} = \sum_{i=1}^{\ell} (\delta^\star_{k_i})^{-2} \leq \ell \max_{i\in[\ell]} (\delta^\star_{k_i})^{-2} \leq \ell s^2.$$

Clearly $\ell \leq s$. Moreover, from Lemma I.7, part (c), since σ_{k_i} grow geometrically by factors of e, we must have that $\ell \leq 1 + \lceil \log \frac{\|\mathbf{M}^\star\|_{\mathrm{op}}}{\sigma} \rceil$. Hence, $\ell \leq \ell_{\sigma,s} := \min\{1 + \lceil \log \frac{\|\mathbf{M}^\star\|_{\mathrm{op}}}{\sigma} \rceil,\ s\}$.

Item (d). We bound

$$\begin{aligned}
\mathsf{tail}_2(\mathbf{M}^\star; k_\ell) &= \sum_{j>k_\ell} (\sigma^\star_j)^2 = \sum_{j=k_\ell+1}^{s} (\sigma^\star_j)^2 + \sum_{j>s} (\sigma^\star_j) = \sum_{j=k_\ell+1}^{s} (\sigma^\star_j)^2 + \mathsf{tail}_2(\mathbf{M}^\star; s) \\
&\leq s(\sigma^\star_{k_\ell+1})^2 + \mathsf{tail}_2(\mathbf{M}^\star; s) \\
&\leq 4e^2 s\sigma^2 + \mathsf{tail}_2(\mathbf{M}^\star; s),
\end{aligned}$$

where in the last line, we used Lemma I.7, part (b). The bound on $\mathsf{tail}_1(\mathbf{M}^\star; k_\ell)$ is analogous. ■
■

K.4. Useful linear algebra facts

We conclude the section by several useful facts about the linear algebra.

Lemma K.1 (Eq. (1), Li and Strang (2020)) *Let* $\mathbf{M}, \mathbf{M}' \in \mathbb{R}^{n\times m}$ *where* $\mathrm{rank}(\mathbf{M}') = r$. *Then,*

$$\forall i \in \{1, \dots, \min\{n,m\} - r\}, \quad \sigma_i(\mathbf{M} - \mathbf{M}') \geq \sigma_{i+r}(\mathbf{M}).$$

Lemma K.2 (Theorem A.14, Bai and Silverstein (2010)) *Let* $\mathbf{M} = \mathbf{A}\mathbf{B}^\top$ *have rank (at most)* r. *Then,* $\sum_{i=1}^r \sigma_i(\mathbf{M}) \leq \sum_{i=1}^r \sigma_i(\mathbf{A})\sigma_i(\mathbf{B})$.

Lemma K.3 (Theorem A.37 (ii), Bai and Silverstein (2010)) *For any* $\mathbf{M}, \mathbf{M}' \in \mathbb{R}^{n\times m}$,

$$\sum_{i=1}^{\nu} (\sigma_i(\mathbf{M}) - \sigma_i(\mathbf{M}'))^2 \leq \|\mathbf{M} - \mathbf{M}'\|_{\mathrm{F}}^2,$$

where the above holds for $\nu = \min(n,m)$, *and thus, also holds for any* $1 \leq \nu \leq \min(n,m)$.

Appendix L. The Balancing Operator

L.1. Properties of the balancing operator

Definition F.4 (Balancing Operator) *Let $\mathbf{X}, \mathbf{Y} \in \mathbb{S}^p_>$. We define the balancing operator*

$$\Psi_{\mathrm{bal}}(\mathbf{Y};\mathbf{X}) := \mathbf{X}^{\frac{1}{2}}(\mathbf{X}^{\frac{1}{2}}\mathbf{Y}\mathbf{X}^{\frac{1}{2}})^{-\frac{1}{2}}\mathbf{X}^{\frac{1}{2}} \in \mathbb{S}^p_>.$$

The uniqueness of $\mathbf{W} = \Psi_{\mathrm{bal}}(\mathbf{Y};\mathbf{X})$ (and hence the well-definedness of the map Ψ_{bal}) is a consequence of the following lemma.

Lemma L.1 *Let $\mathbf{X}, \mathbf{Y} \in \mathbb{S}^p_>$. The balancing operator has the following properties:*

(i) ***Uniqueness:*** *There is a unique $\mathbf{W} = \Psi_{\mathrm{bal}}(\mathbf{Y};\mathbf{X})$ is the unique positive definite matrix satisfying $\mathbf{X} = \mathbf{W}\mathbf{Y}\mathbf{W}$, so that Ψ_{bal} is well-defined.*

(ii) ***Positive scaling:*** $\Psi_{\mathrm{bal}}(\alpha\mathbf{Y};\mathbf{X}) = \alpha^{-\frac{1}{2}}\Psi_{\mathrm{bal}}(\mathbf{Y};\mathbf{X})$.

(iii) ***Anti-monotonicity:*** *If $\mathbf{Y} \succeq \mathbf{Y}'$, then $\Psi_{\mathrm{bal}}(\mathbf{Y};\mathbf{X}) \preceq \Psi_{\mathrm{bal}}(\mathbf{Y}';\mathbf{X})$.*

(iv) ***Comparison with $\mathbf{X}$:*** *If $\mathbf{Y} \succeq \tau\mathbf{X}$, then $\Psi_{\mathrm{bal}}(\mathbf{Y};\mathbf{X}) \preceq \tau^{-\frac{1}{2}}\mathbf{I}_p$, similarly, if $\mathbf{Y} \preceq \tau\mathbf{X}$, $\Psi_{\mathrm{bal}}(\mathbf{Y};\mathbf{X}) \succeq \tau^{-\frac{1}{2}}\mathbf{I}_p$.*

(v) ***Comparison with identity:*** *If $\mathbf{Y} \succeq \tau\mathbf{I}_p$, then $\Psi_{\mathrm{bal}}(\mathbf{Y};\mathbf{X}) \preceq \tau^{-\frac{1}{2}}\mathbf{X}^{\frac{1}{2}}$; similarly, if $\mathbf{Y} \preceq \tau\mathbf{I}_p$, $\Psi_{\mathrm{bal}}(\mathbf{Y};\mathbf{X}) \succeq \tau^{-\frac{1}{2}}\mathbf{X}^{\frac{1}{2}}$.*

(vi) ***Inverse symmetry:*** $\Psi_{\mathrm{bal}}(\mathbf{Y};\mathbf{X}) = \Psi_{\mathrm{bal}}(\mathbf{X};\mathbf{Y})^{-1}$.

(vii) Let $\mathbf{Z} = \mathbf{W}^{\frac{1}{2}}\mathbf{Y}\mathbf{W}^{\frac{1}{2}} = \mathbf{W}^{-\frac{1}{2}}\mathbf{X}\mathbf{W}^{-\frac{1}{2}}$. Then, there exist orthogonal matrices $\mathbf{O}_1, \mathbf{O}_2 \in \mathbb{O}(p)$ such that $\mathbf{Z} \preceq \frac{1}{2}(\mathbf{O}_1\mathbf{X}\mathbf{O}_1^\top + \mathbf{O}_2\mathbf{Y}\mathbf{O}_2^\top)$. Moreover, $\lambda_i(\mathbf{Z}) = \sigma_i(\mathbf{X}^{\frac{1}{2}}\mathbf{Y}^{\frac{1}{2}})$.

Proof Item (i). One can directly check that $\mathbf{W} = \Psi_{\mathrm{bal}}(\mathbf{Y};\mathbf{X})$ satisfies $\mathbf{X} = \mathbf{W}\mathbf{Y}\mathbf{W}$. For uniqueness, $\mathbf{W}$ satisfying $\mathbf{X} = \mathbf{W}\mathbf{Y}\mathbf{W}$ satisfies $\mathbf{I}_p = \mathbf{W}'(\mathbf{X}^{\frac{1}{2}}\mathbf{Y}\mathbf{X}^{\frac{1}{2}})\mathbf{W}'$, where $\mathbf{W}' := \mathbf{X}^{-\frac{1}{2}}\mathbf{W}\mathbf{X}^{-\frac{1}{2}}$. Thus $(\mathbf{W}')^{-2} = \mathbf{X}^{\frac{1}{2}}\mathbf{Y}\mathbf{X}^{\frac{1}{2}}$, so that $(\mathbf{W}')^2 = (\mathbf{X}^{\frac{1}{2}}\mathbf{Y}\mathbf{X}^{\frac{1}{2}})^{-1}$. Note that $\mathbf{X}^{\frac{1}{2}}\mathbf{Y}\mathbf{X}^{\frac{1}{2}} \succ 0$, and since we stipulate $\mathbf{W} \succ 0$, $\mathbf{W}' \succ 0$. Thus, by (Horn and Johnson, 2012, Theorem 7.2.6), it follows that $\mathbf{W}' = (\mathbf{X}^{\frac{1}{2}}\mathbf{Y}\mathbf{X}^{\frac{1}{2}})^{-\frac{1}{2}}$ is the unique positive definite square root of $(\mathbf{X}^{\frac{1}{2}}\mathbf{Y}\mathbf{X}^{\frac{1}{2}})^{-1}$. Solving for $\mathbf{W} = \mathbf{X}^{\frac{1}{2}}\mathbf{W}'\mathbf{X}^{\frac{1}{2}}$, we see $\mathbf{W} = \mathbf{X}^{\frac{1}{2}}(\mathbf{X}^{\frac{1}{2}}\mathbf{Y}\mathbf{X}^{\frac{1}{2}})^{-\frac{1}{2}}\mathbf{X}^{\frac{1}{2}}$.

Item (ii). This is a straightforward computation.

Item (iii). Let $\mathbf{Y} \succeq \mathbf{Y}'$. Then, $\mathbf{X}^{\frac{1}{2}}\mathbf{Y}\mathbf{X}^{\frac{1}{2}} \succeq \mathbf{X}^{\frac{1}{2}}\mathbf{Y}'\mathbf{X}^{\frac{1}{2}}$. The mapping $\mathbf{Z} \mapsto \mathbf{Z}^{-\frac{1}{2}}$ is operator anti-monotone on $\mathbb{S}^p_\geq$ ((Horn and Johnson, 2012, Corollary 7.7.4)). Thus,

$$(\mathbf{X}^{\frac{1}{2}}\mathbf{Y}\mathbf{X}^{\frac{1}{2}})^{-\frac{1}{2}} \preceq (\mathbf{X}^{\frac{1}{2}}\mathbf{Y}'\mathbf{X}^{\frac{1}{2}})^{-\frac{1}{2}}.$$

Therefore,

$$\Psi_{\mathrm{bal}}(\mathbf{Y};\mathbf{X}) = \mathbf{X}^{\frac{1}{2}}(\mathbf{X}^{\frac{1}{2}}\mathbf{Y}\mathbf{X}^{\frac{1}{2}})^{-\frac{1}{2}}\mathbf{X}^{\frac{1}{2}} \preceq \mathbf{X}^{\frac{1}{2}}(\mathbf{X}^{\frac{1}{2}}\mathbf{Y}'\mathbf{X}^{\frac{1}{2}})^{-\frac{1}{2}}\mathbf{X}^{\frac{1}{2}} = \Psi_{\mathrm{bal}}(\mathbf{Y}';\mathbf{X}).$$

Items (iv) and (v) Fix a $r \in \mathbb{R}$. Then

$$\Psi_{\mathrm{bal}}(\tau \mathbf{X}^r; \mathbf{X}) = \mathbf{X}^{\frac{1}{2}}(\mathbf{X}^{\frac{1}{2}} \cdot (\tau \mathbf{X}^r)\mathbf{X}^{\frac{1}{2}})^{-\frac{1}{2}}\mathbf{X}^{\frac{1}{2}} = \mathbf{X}^{\frac{1}{2}}(\tau \mathbf{X}^{r+1})^{-\frac{1}{2}}\mathbf{X}^{\frac{1}{2}} = \tau^{-\frac{1}{2}}\mathbf{X}^{\frac{1-r}{2}}$$

In particular, if $r = 1$, $\Psi_{\mathrm{bal}}(\tau \mathbf{X}^r; \mathbf{X}) = \tau^{-\frac{1}{2}}\mathbf{I}_p$, whereas if $r = 0$, $\Psi_{\mathrm{bal}}(\tau \mathbf{I}_p; \mathbf{X}) = \tau^{-\frac{1}{2}}\mathbf{X}^{\frac{1}{2}}$. The conclusion follows from monotonicity.

Item (vi) If $\mathbf{W}$ satisfies $\mathbf{X} = \mathbf{W}\mathbf{Y}\mathbf{W}$, then $\mathbf{W}' = \mathbf{W}^{-1}$ satisfies $\mathbf{Y} = \mathbf{W}'\mathbf{X}\mathbf{W}'$. The result follows from the uniqueness of Ψ_{bal}.

Item (vii) We start with the following claim:

Claim L.1 *Consider a PSD matrix $\mathbf{\Lambda} = \mathbf{L}\mathbf{L}^\top \in \mathbb{S}^p_>$ with $\mathbf{L} \in \mathbb{R}^{p \times p}$, we have $\mathbf{\Lambda}^{\frac{1}{2}} = \mathbf{O}^\top \mathbf{L}^\top = \mathbf{L}\mathbf{O}$ for some $\mathbf{O} \in \mathbb{O}(p)$.*

Proof Let $\mathbf{L} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ be an SVD of $\mathbf{L}$. Then, $\mathbf{\Lambda} = \mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^\top$, $\mathbf{\Lambda}^{\frac{1}{2}} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\mathbf{V}\mathbf{U}^\top = \mathbf{L}(\mathbf{V}\mathbf{U}^\top)$. Similarly, $\mathbf{\Lambda}^{\frac{1}{2}} = \mathbf{U}\mathbf{V}^\top\mathbf{L}^\top = \mathbf{O}^\top\mathbf{L}^\top$. ■

Now, set $\mathbf{Z} = \mathbf{W}^{-\frac{1}{2}}\mathbf{X}\mathbf{W}^{-\frac{1}{2}} = \mathbf{W}^{\frac{1}{2}}\mathbf{Y}\mathbf{W}^{\frac{1}{2}}$. Then, by the above claim there exist orthogonal matrices $\mathbf{O}_1, \mathbf{O}_2$ such that $\mathbf{O}_1\mathbf{X}^{\frac{1}{2}}\mathbf{W}^{-\frac{1}{2}} = (\mathbf{W}^{-\frac{1}{2}}\mathbf{X}\mathbf{W}^{-\frac{1}{2}})^{\frac{1}{2}} = \mathbf{Z}^{\frac{1}{2}}$ and $\mathbf{W}^{\frac{1}{2}}\mathbf{Y}^{\frac{1}{2}}\mathbf{O}_2 = (\mathbf{W}^{\frac{1}{2}}\mathbf{Y}\mathbf{W}^{\frac{1}{2}})^{\frac{1}{2}} = \mathbf{Z}^{\frac{1}{2}}$. Hence,

$$\mathbf{Z} = \mathbf{O}_1\mathbf{X}^{\frac{1}{2}}\mathbf{W}^{-\frac{1}{2}}\mathbf{W}^{\frac{1}{2}}\mathbf{Y}^{\frac{1}{2}}\mathbf{O}_2 = \mathbf{O}_1\mathbf{X}^{\frac{1}{2}}\mathbf{Y}^{\frac{1}{2}}\mathbf{O}_2.$$

Thus, for any $\mathbf{v} \in \mathbb{R}^p$,

$$\begin{aligned}
\mathbf{v}^\top\mathbf{Z}\mathbf{v} &= \mathbf{v}^\top\mathbf{O}_1\mathbf{X}^{\frac{1}{2}}\mathbf{Y}^{\frac{1}{2}}\mathbf{O}_2\mathbf{v} \\
&\le \|\mathbf{v}^\top\mathbf{O}_1\mathbf{X}^{\frac{1}{2}}\| \cdot \|\mathbf{Y}^{\frac{1}{2}}\mathbf{O}_2\mathbf{v}\| \\
&\le \frac{1}{2}\left(\|\mathbf{v}^\top\mathbf{O}_1\mathbf{X}^{\frac{1}{2}}\|^2 + \|\mathbf{Y}^{\frac{1}{2}}\mathbf{O}_2\mathbf{v}\|^2\right) \\
&= \frac{1}{2}\mathbf{v}^\top\left(\mathbf{O}_1\mathbf{X}\mathbf{O}_1^\top + \mathbf{O}_2\mathbf{Y}\mathbf{O}_2^\top\right)\mathbf{v}.
\end{aligned}$$

Moreover, since $\mathbf{Z} \in \mathbb{S}^p_>$, we have $\lambda_i(\mathbf{Z}) = \sigma_i(\mathbf{Z}) = \sigma_i(\mathbf{O}_1\mathbf{X}^{\frac{1}{2}}\mathbf{Y}^{\frac{1}{2}}\mathbf{O}_2) = \sigma_i(\mathbf{X}^{\frac{1}{2}}\mathbf{Y}^{\frac{1}{2}})$. ■

L.2. Balancing "close" covariances

Lemma L.2 *Let $\mathbf{\Sigma}, \mathbf{\Sigma}' \in \mathbb{S}^p_\ge$ be two matrices with $\mathrm{range}(\mathbf{\Sigma}) = \mathrm{range}(\mathbf{\Sigma}')$. Then, there exists a transformation $\mathbf{T} \in \mathbb{S}^p_>$ such that*

$$\mathbf{T}\mathbf{\Sigma}\mathbf{T} = \mathbf{T}^{-1}\mathbf{\Sigma}'\mathbf{T}^{-1}, \quad \textit{and, since } \mathbf{T} = \mathbf{T}^\top, \ \mathbf{T}\mathbf{\Sigma}\mathbf{T}^\top = \mathbf{T}^{-1}\mathbf{\Sigma}'\mathbf{T}^{-\top}.$$

Moreover, this transformation satisfies, for $r = \mathrm{rank}(\mathbf{\Sigma})$,

$$\max\{\|\mathbf{T}\|_{\mathrm{op}}, \|\mathbf{T}^{-1}\|_{\mathrm{op}}\} \le (1+\Delta)^{1/4}, \quad \textit{where } \Delta := \frac{\|\mathbf{\Sigma} - \mathbf{\Sigma}'\|_{\mathrm{op}}}{\lambda_r(\mathbf{\Sigma}) \wedge \lambda_r(\mathbf{\Sigma}')},$$

$$\sigma_i(\mathbf{T}\mathbf{\Sigma}\mathbf{T}) = \sigma_i(\mathbf{T}\mathbf{\Sigma}\mathbf{T}^\top) = \sigma_i(\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{\Sigma}'^{\frac{1}{2}}) \tag{L.1}$$

Lastly, if $\mathrm{rank}(\mathbf{\Sigma}) = \mathrm{rank}(\mathbf{\Sigma}') = p$, then $\mathbf{T}$ is unique and given by

$$\mathbf{T} = \Psi_{\mathrm{bal}}(\mathbf{\Sigma}'; \mathbf{\Sigma})^{\frac{1}{2}} = \left(\mathbf{\Sigma}^{\frac{1}{2}}(\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{\Sigma}'\mathbf{\Sigma}^{\frac{1}{2}})^{-\frac{1}{2}}\mathbf{\Sigma}^{\frac{1}{2}}\right)^{\frac{1}{2}}$$

Proof The last part of the theorem, when $\mathrm{rank}(\mathbf{\Sigma}) = \mathrm{rank}(\mathbf{\Sigma}') = p$, is a direct consequence of Lemma L.1. We now handle the case when $\mathrm{rank}(\mathbf{\Sigma}) = \mathrm{rank}(\mathbf{\Sigma}') < p$. Let $\mathbf{U} \in \mathbb{R}^{p\times r}$ consist of columns which form an orthonormal basis for $\mathrm{range}(\mathbf{\Sigma}) = \mathrm{range}(\mathbf{\Sigma}')$. Set $\mathbf{X} = \mathbf{U}^\top\mathbf{\Sigma}\mathbf{U}$ and $\mathbf{Y} = \mathbf{U}^\top\mathbf{\Sigma}'\mathbf{U}$. Then,

$$\|\mathbf{X}^{-\frac{1}{2}}\mathbf{Y}\mathbf{X}^{-\frac{1}{2}}\|_{\mathrm{op}} \le 1 + \frac{\|\mathbf{\Sigma}-\mathbf{\Sigma}'\|_{\mathrm{op}}}{\lambda_r(\mathbf{\Sigma})}, \quad \|\mathbf{Y}^{-\frac{1}{2}}\mathbf{X}\mathbf{Y}^{-\frac{1}{2}}\|_{\mathrm{op}} \le 1 + \frac{\|\mathbf{\Sigma}-\mathbf{\Sigma}'\|_{\mathrm{op}}}{\lambda_r(\mathbf{\Sigma}')}.$$

Thus, setting $\Delta := \frac{\|\mathbf{\Sigma}-\mathbf{\Sigma}'\|_{\mathrm{op}}}{\lambda_r(\mathbf{\Sigma})\wedge\lambda_r(\mathbf{\Sigma}')}$, we have

$$\mathbf{Y} \preceq (1+\Delta)\mathbf{X}, \quad \mathbf{X} \preceq (1+\Delta)\mathbf{Y}.$$

Let $\mathbf{W} := \Psi_{\mathrm{bal}}(\mathbf{Y};\mathbf{X})$. Then, from Lemma L.1.(iv),

$$\max\{\|\mathbf{W}\|_{\mathrm{op}}, \|\mathbf{W}^{-1}\|_{\mathrm{op}}\} \le \sqrt{1+\Delta}. \tag{L.2}$$

Moreover, from Lemma L.1.(vii),

$$\begin{aligned}\sigma_i(\mathbf{W}^{\frac{1}{2}}\mathbf{Y}\mathbf{W}^{\frac{1}{2}}) = \sigma_i(\mathbf{W}^{-\frac{1}{2}}\mathbf{X}\mathbf{W}^{-\frac{1}{2}}) &= \sigma_i(\mathbf{X}^{\frac{1}{2}}\mathbf{Y}^{\frac{1}{2}})\\ &= \sigma_i((\mathbf{U}^\top\mathbf{\Sigma}\mathbf{U})^{\frac{1}{2}}(\mathbf{U}^\top\mathbf{\Sigma}'\mathbf{U})^{\frac{1}{2}})\\ &= \sigma_i(\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{\Sigma}'^{\frac{1}{2}}), \quad i \in [r]\end{aligned} \tag{L.3}$$

where the last equality can be verified by a diagonalization argument, and using the fact that $\mathbf{U}$ is a basis for the row space of $\mathbf{\Sigma}$, $\mathbf{\Sigma}'$.

To construct the transformation $\mathbf{T}$, set

$$\mathbf{T} = \mathbf{U}\mathbf{W}^{-\frac{1}{2}}\mathbf{U}^\top + (\mathbf{I}_p - \mathbf{U}\mathbf{U}^\top), \text{ so that } \mathbf{T}^{-1} = \mathbf{U}\mathbf{W}^{\frac{1}{2}}\mathbf{U}^\top + (\mathbf{I}_p - \mathbf{U}\mathbf{U}^\top).$$

Note that $\mathbf{T} \in \mathbb{S}^p_>$, since $\mathbf{W} \in \mathbb{S}^r_>$ and $\mathbf{U}$ is orthonormal. Since $(\mathbf{I}_p - \mathbf{U}\mathbf{U}^\top)\mathbf{\Sigma} = \mathbf{\Sigma}(\mathbf{I}_p - \mathbf{U}\mathbf{U}^\top) = 0$ (and similarly with $\mathbf{\Sigma}'$),

$$\begin{aligned}\mathbf{T}\mathbf{\Sigma}\mathbf{T} &= \mathbf{U}\mathbf{W}^{-\frac{1}{2}}\mathbf{U}^\top\mathbf{\Sigma}\mathbf{U}\mathbf{W}^{-\frac{1}{2}}\mathbf{U}^\top\\ &= \mathbf{U}\mathbf{W}^{-\frac{1}{2}}\mathbf{X}\mathbf{W}^{-\frac{1}{2}}\mathbf{U}^\top\\ &= \mathbf{U}\mathbf{W}^{\frac{1}{2}}\mathbf{Y}\mathbf{W}^{\frac{1}{2}}\mathbf{U}^\top\\ &= \mathbf{U}\mathbf{W}^{\frac{1}{2}}\mathbf{U}^\top\mathbf{\Sigma}'\mathbf{U}\mathbf{W}^{\frac{1}{2}}\mathbf{U}^\top\\ &= \mathbf{T}^{-1}\mathbf{\Sigma}'\mathbf{T}^{-1}.\end{aligned} \tag{L.4}$$

Moreover, by Eq. (L.2),

$$\max\{\|\mathbf{T}\|_{\mathrm{op}}, \|\mathbf{T}^{-1}\|_{\mathrm{op}}\} = \max\{1, \|\mathbf{U}\mathbf{W}^{-\frac{1}{2}}\mathbf{U}^\top\|_{\mathrm{op}}, \|\mathbf{U}\mathbf{W}^{\frac{1}{2}}\mathbf{U}^\top\|_{\mathrm{op}}\} \le (1+\Delta)^{1/4}.$$

Finally, by Eq. (L.4) followed by Eq. (L.3),

$$\sigma_i(\mathbf{T}\mathbf{\Sigma}\mathbf{T}) = \sigma_i(\mathbf{U}\mathbf{W}^{-\frac{1}{2}}\mathbf{X}\mathbf{W}^{-\frac{1}{2}}\mathbf{U}^\top) = \sigma_i(\mathbf{W}^{-\frac{1}{2}}\mathbf{X}\mathbf{W}^{-\frac{1}{2}}) = \sigma_i(\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{\Sigma}'^{\frac{1}{2}}), i \in [r],$$

whereas, for $i > r$, we verify that $\mathbf{W}^{-\frac{1}{2}}\mathbf{X}\mathbf{W}^{-\frac{1}{2}} = 0$. Since $\mathbf{\Sigma}, \mathbf{\Sigma}'$ share the same range and have rank r, we have $\sigma_i(\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{\Sigma}'^{\frac{1}{2}}) = 0$ for $i > r$. ∎

L.2.1. PERTURBATION OF THE BALANCING OPERATOR

Lemma L.3 (Perturbations of Ψ_{bal}, Relative Error) *Fix $\epsilon \in (0,1)$. Then,*

(a) Let $\mathbf{X}, \mathbf{Y}, \mathbf{Y}' \in \mathbb{S}^p_{>}$, with $(1-\epsilon)\mathbf{Y} \preceq \mathbf{Y}' \preceq (1+\epsilon)\mathbf{Y}$. Then,

$$(1+\epsilon)^{-\frac{1}{2}}\Psi_{\mathrm{bal}}(\mathbf{Y};\mathbf{X}) \preceq \Psi_{\mathrm{bal}}(\mathbf{Y}';\mathbf{X}) \preceq (1-\epsilon)^{-\frac{1}{2}}\Psi_{\mathrm{bal}}(\mathbf{Y};\mathbf{X}).$$

(b) Similarly, if $\mathbf{X}, \mathbf{X}', \mathbf{Y} \in \mathbb{S}^p_{>}$, with $(1-\epsilon)\mathbf{X} \preceq \mathbf{X}' \preceq (1+\epsilon)\mathbf{X}$. Then,

$$(1-\epsilon)^{\frac{1}{2}}\Psi_{\mathrm{bal}}(\mathbf{Y};\mathbf{X}) \preceq \Psi_{\mathrm{bal}}(\mathbf{Y};\mathbf{X}') \preceq (1+\epsilon)^{\frac{1}{2}}\Psi_{\mathrm{bal}}(\mathbf{Y};\mathbf{X}).$$

(c) Finally, $\mathbf{X}, \mathbf{X}', \mathbf{Y}, \mathbf{Y}' \in \mathbb{S}^p_{>}$, with $(1-\epsilon)\mathbf{X} \preceq \mathbf{X}' \preceq (1+\epsilon)\mathbf{X}$ and $(1-\epsilon)\mathbf{Y} \preceq \mathbf{Y}' \preceq (1+\epsilon)\mathbf{Y}$, then

$$(1-2\epsilon)\,\Psi_{\mathrm{bal}}(\mathbf{Y};\mathbf{X}) \preceq \Psi_{\mathrm{bal}}(\mathbf{Y}';\mathbf{X}') \preceq \left(1+\frac{2\epsilon}{1-\epsilon}\right)\Psi_{\mathrm{bal}}(\mathbf{Y};\mathbf{X}).$$

Proof [Proof of Lemma L.3] By anti-monotonicity of $\Psi_{\mathrm{bal}}(\cdot;\mathbf{X})$ and the explicit formula for Ψ_{bal},

$$\Psi_{\mathrm{bal}}(\mathbf{Y}';\mathbf{X}) \succeq \Psi_{\mathrm{bal}}((1+\epsilon)\mathbf{Y};\mathbf{X}) = (1+\epsilon)^{-\frac{1}{2}}\Psi_{\mathrm{bal}}(\mathbf{Y};\mathbf{X}).$$

By the same token,

$$(1+\epsilon)^{-\frac{1}{2}}\Psi_{\mathrm{bal}}(\mathbf{Y};\mathbf{X}) \preceq \Psi_{\mathrm{bal}}(\mathbf{Y}';\mathbf{X}) \preceq (1-\epsilon)^{-\frac{1}{2}}\Psi_{\mathrm{bal}}(\mathbf{Y};\mathbf{X}).$$

Hence, the result follows from the inverse symmetry of Ψ_{bal} (Lemma L.1.(vi)).

Finally, combining the first two parts of the lemma, we have

$$\Psi_{\mathrm{bal}}(\mathbf{Y}';\mathbf{X}') \succeq (1+\epsilon)^{-\frac{1}{2}}\Psi_{\mathrm{bal}}(\mathbf{Y};\mathbf{X}') \succeq (1-\epsilon)^{\frac{1}{2}}(1+\epsilon)^{-\frac{1}{2}}\Psi_{\mathrm{bal}}(\mathbf{Y};\mathbf{X}),$$

and

$$\Psi_{\mathrm{bal}}(\mathbf{Y}';\mathbf{X}') \preceq (1-\epsilon)^{-\frac{1}{2}}\Psi_{\mathrm{bal}}(\mathbf{Y};\mathbf{X}') \preceq (1-\epsilon)^{-\frac{1}{2}}(1+\epsilon)^{\frac{1}{2}}\Psi_{\mathrm{bal}}(\mathbf{Y};\mathbf{X}).$$

To conclude, we bound

$$(1-\epsilon)^{\frac{1}{2}}(1+\epsilon)^{-\frac{1}{2}} = \sqrt{\frac{1-\epsilon}{1+\epsilon}} = \sqrt{1-\frac{2\epsilon}{1+\epsilon}} \ge \sqrt{1-2\epsilon} \ge 1-2\epsilon.$$

and

$$(1+\epsilon)^{\frac{1}{2}}(1-\epsilon)^{-\frac{1}{2}} = \sqrt{\frac{1+\epsilon}{1-\epsilon}} = \sqrt{1+\frac{2\epsilon}{1-\epsilon}} \le 1+\frac{2\epsilon}{1-\epsilon}.$$

concluding the proof. ■

Lemma L.4 (Additive Perturbation of Balancing Operator) *Let $\mathbf{X}, \mathbf{X}', \mathbf{Y}, \mathbf{Y}' \in \mathbb{S}^p_>$ be matrices such that $\mu \mathbf{I}_p \preceq \mathbf{X}, \mathbf{X}', \mathbf{Y}, \mathbf{Y}' \preceq M\mathbf{I}_p$, and $\|\mathbf{X} - \mathbf{X}'\|_{\mathrm{op}}, \|\mathbf{Y} - \mathbf{Y}'\|_{\mathrm{op}} \le \Delta \le \mu/3$ for some $\Delta > 0$. Then,*

$$\|\Psi_{\mathrm{bal}}(\mathbf{Y}'; \mathbf{X}') - \Psi_{\mathrm{bal}}(\mathbf{Y}; \mathbf{X})\|_{\mathrm{op}} \le 3\Delta \cdot \frac{\sqrt{M/\mu}}{\mu}.$$

Moreover, $\|\Psi_{\mathrm{bal}}(\mathbf{Y}'; \mathbf{X}')\|_{\mathrm{op}}, \|\Psi_{\mathrm{bal}}(\mathbf{Y}; \mathbf{X})\|_{\mathrm{op}} \le \sqrt{M/\mu}$ and $\Psi_{\mathrm{bal}}(\mathbf{Y}'; \mathbf{X}'), \Psi_{\mathrm{bal}}(\mathbf{Y}; \mathbf{X}) \succeq \sqrt{\mu/M}\mathbf{I}_p$.

Proof [Proof of Lemma L.4] Under the above conditions, it holds that $(1 - \mu^{-1}\Delta)\mathbf{X} \preceq \mathbf{X}' \preceq (1 + \mu^{-1}\Delta)\mathbf{X}$, and similarly for $\mathbf{Y}$ and $\mathbf{Y}'$. Applying Lemma L.3 with $\epsilon = \Delta/\mu \le 1/3$, we have

$$(1 - 2\epsilon)\Psi_{\mathrm{bal}}(\mathbf{Y}; \mathbf{X}) \preceq \Psi_{\mathrm{bal}}(\mathbf{Y}'; \mathbf{X}') \preceq (1 + 3\epsilon)\Psi_{\mathrm{bal}}(\mathbf{Y}; \mathbf{X}).$$

This gives

$$\|\Psi_{\mathrm{bal}}(\mathbf{Y}'; \mathbf{X}') - \Psi_{\mathrm{bal}}(\mathbf{Y}; \mathbf{X})\|_{\mathrm{op}} \le 3\frac{\Delta}{\mu}\|\Psi_{\mathrm{bal}}(\mathbf{Y}; \mathbf{X})\|_{\mathrm{op}}.$$

Lastly, since $\mathbf{Y} \succeq \mu/M\mathbf{X}$ (as $\mathbf{Y} \succeq \mu\mathbf{I}_p$ and $\mathbf{X} \preceq M\mathbf{I}_p$), it holds $\|\Psi_{\mathrm{bal}}(\mathbf{Y}; \mathbf{X})\|_{\mathrm{op}} \le \sqrt{M/\mu}$ by Lemma L.1.(iv). Thus,

$$\|\Psi_{\mathrm{bal}}(\mathbf{Y}'; \mathbf{X}') - \Psi_{\mathrm{bal}}(\mathbf{Y}; \mathbf{X})\|_{\mathrm{op}} \le 3\Delta \cdot \frac{\sqrt{M/\mu}}{\mu}.$$

A similar computation also shows $\|\Psi_{\mathrm{bal}}(\mathbf{Y}'; \mathbf{X}')\|_{\mathrm{op}} \le \sqrt{M/\mu}$, and $\Psi_{\mathrm{bal}}(\mathbf{Y}'; \mathbf{X}'), \Psi_{\mathrm{bal}}(\mathbf{Y}; \mathbf{X}) \succeq \sqrt{\mu/M}\mathbf{I}_p$. ■

We recall the definition of the balanced covariance.

Definition F.5 (Balanced Covariance) *Given $\mathbf{X}, \mathbf{Y} \in \mathbb{S}^p_>$, we define*

$$\mathsf{CovBal}(\mathbf{X}, \mathbf{Y}) = \Psi_{\mathrm{bal}}(\mathbf{Y}; \mathbf{X})^{\frac{1}{2}} \cdot \mathbf{Y} \cdot \Psi_{\mathrm{bal}}(\mathbf{Y}; \mathbf{X})^{\frac{1}{2}}.$$

Remark 12 (Symmetry of CovBal) *Note that, from definition of Ψ_{bal}, we also have $\mathsf{CovBal}(\mathbf{X}, \mathbf{Y}) = \Psi_{\mathrm{bal}}(\mathbf{Y}; \mathbf{X})^{\frac{1}{2}}\mathbf{Y}\Psi_{\mathrm{bal}}(\mathbf{Y}; \mathbf{X})^{\frac{1}{2}} = \Psi_{\mathrm{bal}}(\mathbf{X}; \mathbf{Y})^{-\frac{1}{2}}\mathbf{Y}\Psi_{\mathrm{bal}}(\mathbf{X}; \mathbf{Y})^{-\frac{1}{2}} = \mathsf{CovBal}(\mathbf{Y}, \mathbf{X})$.*

Lemma L.5 (Perturbation of Balanced Covariance) *Let $\mathbf{X}, \mathbf{X}', \mathbf{Y}, \mathbf{Y}' \in \mathbb{S}^p_>$ be the matrices such that $\mu \mathbf{I}_p \preceq \mathbf{X}, \mathbf{X}', \mathbf{Y}, \mathbf{Y}' \preceq M\mathbf{I}_p$, and $\|\mathbf{X} - \mathbf{X}'\|_{\mathrm{op}}, \|\mathbf{Y} - \mathbf{Y}'\|_{\mathrm{op}} \le \Delta \le \mu/3$ for some $\Delta > 0$. Then,*

$$\|\mathsf{CovBal}(\mathbf{X}', \mathbf{Y}') - \mathsf{CovBal}(\mathbf{X}, \mathbf{Y})\|_{\mathrm{op}} \le 4(M/\mu)^2\Delta.$$

Moreover, we have

$$\|\Psi_{\mathrm{bal}}(\mathbf{Y}'; \mathbf{X}')^{-\frac{1}{2}} - \Psi_{\mathrm{bal}}(\mathbf{Y}; \mathbf{X})^{-\frac{1}{2}}\|_{\mathrm{op}} \vee \|\Psi_{\mathrm{bal}}(\mathbf{Y}'; \mathbf{X}')^{\frac{1}{2}} - \Psi_{\mathrm{bal}}(\mathbf{Y}; \mathbf{X})^{\frac{1}{2}}\|_{\mathrm{op}} \le \frac{3}{2\mu}(M/\mu)^{3/4}\Delta.$$

Proof [Proof of Lemma L.5] We have

$$\begin{aligned}
&\|\mathsf{CovBal}(\mathbf{X}', \mathbf{Y}') - \mathsf{CovBal}(\mathbf{X}, \mathbf{Y})\|_{\mathrm{op}} \\
&= \|\Psi_{\mathrm{bal}}(\mathbf{Y}'; \mathbf{X}')^{\frac{1}{2}}\mathbf{Y}'\Psi_{\mathrm{bal}}(\mathbf{Y}'; \mathbf{X}')^{\frac{1}{2}} - \Psi_{\mathrm{bal}}(\mathbf{Y}; \mathbf{X})^{\frac{1}{2}}\mathbf{Y}\Psi_{\mathrm{bal}}(\mathbf{Y}; \mathbf{X})^{\frac{1}{2}}\|_{\mathrm{op}} \\
&\le \|\Psi_{\mathrm{bal}}(\mathbf{Y}'; \mathbf{X}')^{\frac{1}{2}}(\mathbf{Y}' - \mathbf{Y})\Psi_{\mathrm{bal}}(\mathbf{Y}'; \mathbf{X}')^{\frac{1}{2}}\|_{\mathrm{op}} \\
&\quad + \|\Psi_{\mathrm{bal}}(\mathbf{Y}'; \mathbf{X}')^{\frac{1}{2}} - \Psi_{\mathrm{bal}}(\mathbf{Y}; \mathbf{X})^{\frac{1}{2}}\|_{\mathrm{op}}\|\mathbf{Y}\|_{\mathrm{op}}(\|\Psi_{\mathrm{bal}}(\mathbf{Y}'; \mathbf{X}')^{\frac{1}{2}}\|_{\mathrm{op}} + \|\Psi_{\mathrm{bal}}(\mathbf{Y}; \mathbf{X})^{\frac{1}{2}}\|_{\mathrm{op}}) \\
&\le \Delta\sqrt{M/\mu} + 2M(M/\mu)^{1/4}\|\Psi_{\mathrm{bal}}(\mathbf{Y}'; \mathbf{X}')^{\frac{1}{2}} - \Psi_{\mathrm{bal}}(\mathbf{Y}; \mathbf{X})^{\frac{1}{2}}\|_{\mathrm{op}}.
\end{aligned}$$

We now require following perturbation inequality for the matrix square root.

Lemma L.6 (Perturbation of Matrix Square Root, Lemma 2.2. in Schmitt (1992)) *Let $\mathbf{A}_1, \mathbf{A}_2 \in \mathbb{S}^p_>$ satisfy $\mathbf{A}_1, \mathbf{A}_2 \succeq \gamma\mathbf{I}_p$. Then, $\|\mathbf{A}_1^{\frac{1}{2}} - \mathbf{A}_2^{\frac{1}{2}}\|_{\mathrm{op}} \le \frac{1}{2\sqrt{\gamma}}\|\mathbf{A}_1 - \mathbf{A}_2\|_{\mathrm{op}}$.*

Using $\Psi_{\mathrm{bal}}(\mathbf{Y}'; \mathbf{X}')^{\frac{1}{2}}, \Psi_{\mathrm{bal}}(\mathbf{Y}; \mathbf{X})^{\frac{1}{2}} \succeq \sqrt{\mu/M}\mathbf{I}_p$, Lemma L.6 followed by Lemma L.4 implies

$$\begin{aligned}
\|\Psi_{\mathrm{bal}}(\mathbf{Y}'; \mathbf{X}')^{\frac{1}{2}} - \Psi_{\mathrm{bal}}(\mathbf{Y}; \mathbf{X})^{\frac{1}{2}}\|_{\mathrm{op}} &\le \frac{1}{2}(M/\mu)^{1/4}\|\Psi_{\mathrm{bal}}(\mathbf{Y}'; \mathbf{X}') - \Psi_{\mathrm{bal}}(\mathbf{Y}; \mathbf{X})\|_{\mathrm{op}} \\
&\le \frac{3}{2\mu}(M/\mu)^{3/4}\Delta. \qquad \text{(L.5)}
\end{aligned}$$

Thus, we conclude the first part of the lemma:

$$\|\mathsf{CovBal}(\mathbf{X}', \mathbf{Y}') - \mathsf{CovBal}(\mathbf{X}, \mathbf{Y})\|_{\mathrm{op}} \le \Delta\sqrt{M/\mu} + 3(M/\mu)^2\Delta \le 4(M/\mu)^2\Delta.$$

The second bound in the lemma was derived above, and the bound on $\|\Psi_{\mathrm{bal}}(\mathbf{Y}'; \mathbf{X}')^{\frac{1}{2}} - \Psi_{\mathrm{bal}}(\mathbf{Y}; \mathbf{X})^{\frac{1}{2}}\|$ is precisely Eq. (L.5). The bound $\|\Psi_{\mathrm{bal}}(\mathbf{Y}'; \mathbf{X}')^{\frac{-1}{2}} - \Psi_{\mathrm{bal}}(\mathbf{Y}; \mathbf{X})^{\frac{-1}{2}}\|$ follows from the inverse symmetry of the balancing operator (Lemma L.1.(vi)). ■

L.3. Balancing of finite-dimensional embeddings

Lemma L.7 *Let $\mathcal{D}_{\mathcal{X}}$ be a distribution over $\mathcal{X}$, let $\boldsymbol{\Sigma} = \mathbb{E}_{\mathcal{D}_{\mathcal{X}}}[ff^\top]$, and let $\mathbf{P}$ be the orthogonal projection on* $\mathrm{range}(\boldsymbol{\Sigma})$. *Then $\mathbf{P}f = f$ $\mathcal{D}_{\mathcal{X}}$-almost surely; that is, $\mathbb{P}_{\mathcal{D}_{\mathcal{X}}}[f(x) \in \mathrm{range}(\boldsymbol{\Sigma})] = 1$.*

Proof It suffices to show $\mathbb{E}[\|(\mathbf{I}_p - \mathbf{P})f\|^2] = 0$. As $\mathbf{P}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}\mathbf{P}$, we have

$$\begin{aligned}
\mathbb{E}[\|(\mathbf{I}_p - \mathbf{P})f\|^2] &= \mathrm{tr}[\mathbb{E}[((\mathbf{I}_p - \mathbf{P})f)((\mathbf{I}_p - \mathbf{P})f)^\top]] = \mathrm{tr}(\boldsymbol{\Sigma} - \mathbf{P}\boldsymbol{\Sigma} - \boldsymbol{\Sigma}\mathbf{P}^\top + \mathbf{P}\boldsymbol{\Sigma}\mathbf{P}^\top) = 0 \\
&= \mathrm{tr}(\boldsymbol{\Sigma} - \boldsymbol{\Sigma} - \boldsymbol{\Sigma} + \boldsymbol{\Sigma}) = 0.
\end{aligned}$$

■

Lemma L.8 *For any pair of embeddings $\hat{f} : \mathcal{X} \to \mathbb{R}^r$ and $\hat{g} : \mathcal{Y} \to \mathbb{R}^r$, there exists embeddings $\tilde{f} : \mathcal{X} \to \mathbb{R}^r$ and $\tilde{g} : \mathcal{Y} \to \mathbb{R}^r$ such that* *(a)* $\langle \hat{f}, \hat{g} \rangle = \langle \tilde{f}, \tilde{g} \rangle$ *almost surely, and*

$$\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[\tilde{f}\tilde{f}^\top] = \mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}[\tilde{g}\tilde{g}^\top].$$

(b) *For all* $i \in \mathbb{N}$, $\sigma_i(\hat{f}, \hat{g}) = \sigma_i(\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[\tilde{f}\tilde{f}^\top]) = \sigma_i(\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[\tilde{f}\tilde{f}^\top])$, *where we recall* $\sigma_i(\cdot, \cdot)$ *defined in Eq.* (3.1)

(c) $(\tilde{f}, \tilde{g})$ *is full-rank if and only if* $(\hat{f}, \hat{g})$ *is, and in this case* $\mathbf{T}$ *is uniqely given by*

$$\mathbf{T} = \Psi_{\mathrm{bal}}(\mathbf{\Sigma}_g; \mathbf{\Sigma}_f)^{\frac{1}{2}} = \left(\mathbf{\Sigma}_f^{\frac{1}{2}} (\mathbf{\Sigma}_f^{\frac{1}{2}} \mathbf{\Sigma}_g \mathbf{\Sigma}_f^{\frac{1}{2}})^{-\frac{1}{2}} \mathbf{\Sigma}_f^{\frac{1}{2}} \right)^{\frac{1}{2}}, \tag{L.6}$$

where $\mathbf{\Sigma}_f = \mathbb{E}[\hat{f}\hat{f}^\top]$ *and* $\mathbf{\Sigma}_g = \mathbb{E}[\hat{g}\hat{g}^\top]$.

Proof Given $(\hat{f}, \hat{g})$, let us construct a sequence of embeddings $(\hat{f}_i, \hat{g}_i)_{i \geq 0}$, with covariances $\mathbf{\Sigma}_{f,i} := \mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[\hat{f}_i(\hat{f}_i)^\top]$ and $\mathbf{\Sigma}_{g,i} := \mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}[\hat{g}_i(\hat{g}_i)^\top]$, and minimum rank

$$r_i := \min\{\mathrm{rank}(\mathbf{\Sigma}_{f,i}, \mathbf{\Sigma}_{g,i}\}$$

Lastly, set $\mathbf{P}_{f,i}$ to be the orthogonal projection on the range of $\mathbf{\Sigma}_{f,i}$ and $\mathbf{P}_{g,i}$ the same for $\mathbf{\Sigma}_{g,i}$. We define

$$(\hat{f}_0, \hat{g}_0) = (\hat{f}, \hat{g}), \quad (\hat{f}_{i+1}, \hat{g}_{i+1}) = \begin{cases} (\mathbf{P}_{g,i}\hat{f}_i, \hat{g}_i) & \mathrm{rank}(\mathbf{\Sigma}_{f,i}) \geq \mathrm{rank}(\mathbf{\Sigma}_{g,i}) \\ (\hat{f}_i, \mathbf{P}_{g,i}\hat{g}_i) & \text{otherwise} \end{cases}$$

We establish three claims.

Claim L.2 *For any* n, $(\hat{f}_n, \hat{g}_n)$ *is full-rank if and only if* $(\hat{f}, \hat{g})$ *is, which is true if and only if* $\hat{f}_n = \hat{f}$ *and* $\hat{g}_n = g$.

Proof We argue by induction that $(\hat{f}_n, \hat{g}_n)$ is full-rank if and only if $(\hat{f}_{n+1}, \hat{g}_{n+1})$. The "if" follows since $\mathrm{rank}(\mathbf{\Sigma}_{\cdot,n}) \leq \mathrm{rank}(\mathbf{\Sigma}_{\cdot,n+1})$. The "only if" follows since if $(\hat{f}_n, \hat{g}_n)$ is full-rank, $\mathbf{P}_{f,n} = \mathbf{P}_{g,n}$ are the identity, and thus, $\hat{f}_{n+1} = \hat{f}_n$, $\hat{g}_{n+1} = \hat{g}_n$. ■

Claim L.3 *For any* n, *let holds that* $\langle \hat{f}_n, \hat{g}_n \rangle = \langle \hat{f}, \hat{g} \rangle$ *almost-surely under* $\mathcal{D}_{1 \otimes 1}$.

Proof We prove by induction on n. The base case $n = 0$ is immediate. Assume now that $\langle \hat{f}_n, \hat{g}_n \rangle = \langle \hat{f}, \hat{g} \rangle$ holds almost-surely under $\mathcal{D}_{1 \otimes 1}$. Assume that without los of generality $\mathrm{rank}(\mathbf{\Sigma}_{f,n}) \geq \mathrm{rank}(\mathbf{\Sigma}_{g,n})$, so that $(\hat{f}_{n+1}, \hat{g}_{n+1}) = (\mathbf{P}_{g,n}\hat{f}_n, \hat{g}_n)$. Then, by symmetry of the projection $\mathbf{P}_{g,n}$, we have

$$\langle \hat{f}_{n+1}, \hat{g}_{n+1} \rangle = \langle \mathbf{P}_{g,n}\hat{f}_n, \hat{g}_n \rangle = \langle \hat{f}_n, \mathbf{P}_{g,n}\hat{g}_n \rangle.$$

By Lemma L.7, $\mathbf{P}_{g,n}\hat{g}_n = \hat{g}_n$ almost surely, and the result follows. ■

Claim L.4 *Let*

$$n := \inf\{i \in \mathbb{N} : \mathrm{rank}(\mathbf{\Sigma}_{f,n}) = \mathrm{rank}(\mathbf{\Sigma}_{f,n+1}) \textit{ and } \mathrm{rank}(\mathbf{\Sigma}_{g,n}) = \mathrm{rank}(\mathbf{\Sigma}_{g,n+1})\}$$

Then n is finite, and $\mathrm{range}(\mathbf{\Sigma}_{f,n}) = \mathrm{range}(\mathbf{\Sigma}_{g,n})$.

Proof That n is finite follows since the ranks of the covariances $\mathrm{rank}(\mathbf{\Sigma}_{\cdot,i+1}) \le \mathrm{rank}(\mathbf{\Sigma}_{\cdot,i})\}$ are non-increasing. Next, without loss of generality, assume that $\mathrm{rank}(\mathbf{\Sigma}_{f,n}) \ge \mathrm{rank}(\mathbf{\Sigma}_{g,n})$, so that $(\hat{f}_{n+1}, \hat{g}_{n+1}) = (\mathbf{P}_{g,n}\hat{f}_n, \hat{g}_n)$. Then, $\mathbf{\Sigma}_{g,n} = \mathbf{\Sigma}_{g,n+1}$, and

$$\mathbf{\Sigma}_{f,n+1} = \mathbf{P}_{g,n}\mathbf{\Sigma}_{f,n}\mathbf{P}_{g,n}, \text{ so } \mathrm{range}(\mathbf{\Sigma}_{f,n+1}) \subset \mathrm{range}(\mathbf{P}_{g,n}) = \mathrm{range}(\mathbf{\Sigma}_{g,n}).$$

On the other hand,

$$\mathrm{rank}(\mathbf{\Sigma}_{g,n}) \le \mathrm{rank}(\mathbf{\Sigma}_{f,n}) = \mathrm{rank}(\mathbf{\Sigma}_{f,n+1}) = \mathrm{rank}(\mathbf{P}_{g,n}\mathbf{\Sigma}_{f,n}\mathbf{P}_{g,n}),$$

which implies that $\mathrm{range}(\mathbf{\Sigma}_{f,n+1}) \subset \mathrm{range}(\mathbf{\Sigma}_{g,n}) = \mathrm{range}(\mathbf{\Sigma}_{g,n+1})$. ■

Hence, let $\mathbf{T}$ denote the (symmetric) positive definite transformation assured by applying Lemma L.2 to

$$\mathbf{\Sigma} \leftarrow \mathbf{\Sigma}_{f,n+1}, \quad \mathbf{\Sigma}' \leftarrow \mathbf{\Sigma}_{g,n+1}; \tag{L.7}$$

these matrices have the same range by the above claim. Take $\tilde{f} := \mathbf{T}\hat{f}_{n+1}$ and $\tilde{g} := \mathbf{T}^{-1}\hat{g}_{n+1}$. We show all desired properties holds.

Part (a). The transformation $\mathbf{T}$ ensures that

$$\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[\tilde{f}\tilde{f}^\top] = \mathbf{T}\mathbf{\Sigma}_{f,n+1}\mathbf{T} = \mathbf{T}^{-1}\mathbf{\Sigma}_{g,n+1}\mathbf{T}^{-1} = \mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}[\tilde{g}\tilde{g}^\top],$$

Moreover, symmetry of $\mathbf{T}$ and Claim L.3 imply that, almost surely,

$$\langle \tilde{f}, \tilde{g} \rangle = \langle \mathbf{T}\hat{f}_{n+1}, \mathbf{T}^{-1}\hat{g}_{n+1} \rangle = \langle \hat{f}_{n+1}, \mathbf{T}^\top\mathbf{T}^{-1}\hat{g}_{n+1} \rangle = \langle \hat{f}_{n+1}, \hat{g}_{n+1} \rangle = \langle \hat{f}, \hat{g} \rangle,$$

Part (b). This is a consequence of Eq. (L.1) in Lemma L.2, noting that for $\mathbf{\Sigma}, \mathbf{\Sigma}'$ defined in Eq. (L.7) that $\sigma_r(\hat{f}, \hat{g}) := \sigma_r(\mathbf{\Sigma}^{1/2}(\mathbf{\Sigma}')^{1/2})$).

Part (c). Note that if $\hat{f}, \hat{g}$ are full-rank, $\mathbf{\Sigma}_{f,n+1} = \mathbf{\Sigma}_f$ and $\mathbf{\Sigma}_{g,n+1} = \mathbf{\Sigma}_g$, so that $\mathbf{T}$ is uniquely given by Eq. (L.6) due to Lemma L.1. Note that $(\tilde{f}, \tilde{g})$ is full-rank if and only if $(\hat{f}_{n+1}, \hat{g}_{n+1})$, which by Claim L.2 is full-rank if and only if $(\hat{f}, \hat{g})$ is. ■

L.4. Analysis of separation rank

Definition F.7 *Given $\mathbf{\Sigma} \in \mathbb{S}^p_{\ge}$, $r_0 \in [p]$, $\sigma > 0$, the* separated-rank *at (r_0, σ) (if it exists) is*

$$\textsf{sep-rank}(r_0, \sigma; \mathbf{\Sigma}) := \max\left\{ r \in [r_0] : \sigma_r(\mathbf{\Sigma}) \ge \sigma, \sigma_r(\mathbf{\Sigma}) - \sigma_{r+1}(\mathbf{\Sigma}) \ge \frac{\sigma_r(\mathbf{\Sigma})}{r_0} \right\}. \tag{F.9}$$

We say the separated-rank is well-defined *if the above maximum exists.*

Lemma L.9 (Properties of Separated Rank) *Given $r_0 \in [p]$ and $\sigma \in [\sigma_{r_0}(\mathbf{\Sigma}), \|\mathbf{\Sigma}\|_{\mathrm{op}}/e]$, $\mathsf{sep\text{-}rank}(r_0, \sigma; \mathbf{\Sigma})$ enjoys the following properties:*

(a) $\mathsf{sep\text{-}rank}(r_0, \sigma; \mathbf{\Sigma})$ *is well-defined: i.e. for some $r \in [r_0]$, it holds that $\sigma_r(\mathbf{\Sigma}) - \sigma_{r+1}(\mathbf{\Sigma}) \geq \frac{\sigma_r(\mathbf{\Sigma})}{r_0}$ and $\sigma_r(\mathbf{\Sigma}) \geq \sigma$.*

(b) For $r = \mathsf{sep\text{-}rank}(r_0, \sigma; \mathbf{\Sigma})$, we have $\sigma_{r+1}(\mathbf{\Sigma}) \leq e\sigma$.

Proof To prove part (a), we observe that since $\sigma \geq \sigma_{r_0}(\mathbf{\Sigma})$, there must exist some maximal $r_{\max} \in [r_0]$ for which $\sigma_{r_{\max}}(\mathbf{\Sigma}) \leq \sigma$. Now suppose that, for the sake of contradiction, for all $r \leq r_{\max}$, it holds that $\sigma_r(\mathbf{\Sigma}) - \sigma_{r+1}(\mathbf{\Sigma}) < \frac{\sigma_r(\mathbf{\Sigma})}{r_0}$. Then, $\|\mathbf{\Sigma}\|_{\mathrm{op}} = \sigma_1(\mathbf{\Sigma}) \leq (1 + 1/r_0)^{r_{\max}}\sigma_{r_{\max}+1}(\mathbf{\Sigma}) \leq (1 + 1/r_0)^{r_{\max}}\sigma \leq e\sigma$. This contradicts our condition that $\sigma \leq \|\mathbf{\Sigma}\|_{\mathrm{op}}/e$.

To prove part (b), again let $r_{\max} \leq r_0$ be as in the proof of part (a). We must have that $r = \mathsf{sep\text{-}rank}(r_0, \sigma; \mathbf{\Sigma}) \leq r_{\max}$. If $r = r_{\max}$, then $\sigma_{r+1}(\mathbf{\Sigma}) \leq \sigma$. Otherwise, for any $r' \in \{r + 1, r + 2, \dots, r_{\max}\}$, it holds that $\sigma_{r'}(\mathbf{\Sigma}) \leq (1 + 1/r_0)\sigma_{r'+1}(\mathbf{\Sigma})$. Hence, $\sigma_{r+1} \leq (1 + 1/r_0)^{r_{\max}-r}\sigma_{r_{\max}+1} \leq e\sigma_{r_{\max}+1} \leq e\sigma$. ■

L.5. Proof of Proposition F.3

Lemma L.10 *Fix $\mathbf{\Sigma}, \mathbf{\Sigma}'$, $r_0 \in [p]$ and suppose $\|\mathbf{\Sigma} - \mathbf{\Sigma}'\|_{\mathrm{op}} \leq \sigma/4r_0$. Lastly, assume that one of the two hold*

(i) $\sigma \in [\frac{4}{3}\sigma_{r_0}(\mathbf{\Sigma}), \frac{4}{5e}\|\mathbf{\Sigma}\|_{\mathrm{op}}]$.

(ii) There exists positive numbers $\bar{\sigma}_{r_0}$ and $\bar{\sigma}_1$ satisfying $\max\{|\bar{\sigma}_{r_0} - \sigma_{r_0}(\mathbf{\Sigma})|, |\bar{\sigma}_1 - \sigma_1(\mathbf{\Sigma})|\} \leq \sigma/4$ for which $\sigma \in [2\bar{\sigma}_{r_0}, \frac{2}{3e}\bar{\sigma}_1]$.

Let $r = \mathsf{sep\text{-}rank}(r_0, \sigma; \mathbf{\Sigma}')$, and let $\mathbf{P}_r$ and $\mathbf{P}'_r$ denote the projections onto the top-r singular spaces of $\mathbf{\Sigma}$ and $\mathbf{\Sigma}'$. Then, for any Schatten p-norm $\|\cdot\|_\circ$, $\mathbf{P}_r$ and $\mathbf{P}_{r'}$ are unique, and

$$\|\mathbf{P}_r - \mathbf{P}'_r\|_\circ \leq 4r_0\frac{\|\mathbf{\Sigma} - \mathbf{\Sigma}'\|_\circ}{\sigma}.$$

The lemma also holds under the following more general condition:

Proof Set $\Delta = \|\mathbf{\Sigma} - \mathbf{\Sigma}'\|_{\mathrm{op}}$. We shall show that under both conditions of the lemma, it holds that

$$\sigma \in [\sigma_{r_0}(\mathbf{\Sigma}'), \|\mathbf{\Sigma}'\|_{\mathrm{op}}/e],$$

so that the conditions of Lemma L.9 are met.

Condition (i). By Weyl's inequality and our assumption on Δ and the first assumption on σ,

$$\begin{aligned}\sigma - \sigma_{r_0}(\mathbf{\Sigma}') &\geq \sigma(1 - \tfrac{1}{4r_0}) - \sigma_{r_0}(\mathbf{\Sigma}) \geq \tfrac{3\sigma}{4} - \sigma_{r_0}(\mathbf{\Sigma}) \geq 0\\ \|\mathbf{\Sigma}'\|_{\mathrm{op}}/e - \sigma &\geq \|\mathbf{\Sigma}\|_{\mathrm{op}}/e - (1 + \tfrac{1}{4r_0})\sigma \geq 0,\end{aligned}$$

so that $\sigma \in [\sigma_{r_0}(\mathbf{\Sigma}'), \frac{1}{e}\|\mathbf{\Sigma}'\|_{\mathrm{op}}]$.

Condition (ii). We are given $\bar{\sigma}_{r_0}$ and $\bar{\sigma}_1$ satisfying $\max\{|\bar{\sigma}_{r_0} - \sigma_{r_0}(\mathbf{\Sigma})|, |\bar{\sigma}_1 - \sigma_1(\mathbf{\Sigma})|\} \le \sigma/4$ for which $\sigma \in [2\bar{\sigma}_{r_0}, \frac{2}{3e}\bar{\sigma}_1]$. Thus,

$$\begin{aligned}\sigma - \sigma_{r_0}(\mathbf{\Sigma}') &\ge \sigma(1 - \tfrac{1}{4r_0} - \tfrac{1}{4}) - \bar{\sigma}_{r_0}(\mathbf{\Sigma}_{\mathrm{ref}}) \ge \sigma/2 - \bar{\sigma}_{r_0} \ge 0\\ \|\mathbf{\Sigma}'\|_{\mathrm{op}}/e - \sigma &\ge \bar{\sigma}_1 - (1 + \tfrac{1}{4r_0} + \tfrac{1}{2})\sigma \ge 0.\end{aligned}$$

Next, set $\mu = \sigma_r(\mathbf{\Sigma}') - \Delta$. Then $\sigma_r(\mathbf{\Sigma}') \ge \mu$, and by Weyl's inequality, $\sigma_r(\mathbf{\Sigma}) \ge \sigma_r(\mathbf{\Sigma}') - \Delta = \mu$. Moreover, the definition of sep-rank ensures that $\sigma_r(\mathbf{\Sigma}') \ge \sigma$ as well as

$$\sigma_{r+1}(\mathbf{\Sigma}') \le \sigma_r(\mathbf{\Sigma}')(1 - 1/r_0) = \mu + \Delta - \sigma_r(\mathbf{\Sigma}')/r_0 \le \mu + \Delta - \sigma/r_0.$$

Again, by Weyl's inequality, $\sigma_{r+1}(\mathbf{\Sigma}) \le \mu + 2\Delta - \sigma/r_0$. Thus, for $\Delta \le \sigma/4r_0$, $\max\{\sigma_{r+1}(\mathbf{\Sigma}'), \sigma_{r+1}(\mathbf{\Sigma})\} \le \mu - \tau$, where $\tau = \sigma/2r_0$. It follows from Lemma D.1 that if $\mathbf{U}_r'$ is an orthonormal basis corresponding to the top r eigenvalues of $\mathbf{\Sigma}'$, and $\mathbf{U}_{>r}'$ is an orthonormal basis corresponding to eigenvalues $r+1, \dots, p$, and defining $\mathbf{U}_r, \mathbf{U}_{>r}$ analogously for $\mathbf{\Sigma}$. Then, one has that, for any Schatten-p norm $\|\cdot\|_\circ$

$$\max\{\|(\mathbf{U}_r')^\top \mathbf{U}_{>r}\|_\circ,\ \|(\mathbf{U}_{>r}')^\top \mathbf{U}_r\|_\circ\} \le 2r_0 \frac{\|\mathbf{\Sigma} - \mathbf{\Sigma}'\|_\circ}{\sigma}.$$

On the other hand, if $\mathbf{P}_r, \mathbf{P}_r'$ denote the projections onto the top-r eigenspaces of $\mathbf{\Sigma}, \mathbf{\Sigma}'$, we have

$$\begin{aligned}\|\mathbf{P}_r' - \mathbf{P}_r\|_\circ &\le \|\mathbf{P}_r'(\mathbf{P}_r' - \mathbf{P}_r)\|_\circ + \|(\mathbf{I}_p - \mathbf{P}_r')(\mathbf{P}_r' - \mathbf{P}_r)\|_\circ\\ &= \|\mathbf{P}_r' - \mathbf{P}_r'\mathbf{P}_r\|_\circ + \|(\mathbf{I}_p - \mathbf{P}_r')\mathbf{P}_r\|_\circ\\ &= \|\mathbf{P}_r'(\mathbf{I}_p - \mathbf{P}_r)\|_\circ + \|(\mathbf{I}_p - \mathbf{P}_r')\mathbf{P}_r\|_\circ\\ &= \|(\mathbf{U}_r')^\top \mathbf{U}_{>r}\|_\circ + \|(\mathbf{U}_{>r}')^\top \mathbf{U}_r\|_\circ\\ &\le 4r_0 \frac{\|\mathbf{\Sigma} - \mathbf{\Sigma}'\|_\circ}{\sigma}.\end{aligned}$$

■

Proposition F.3 (Perturbation of Balancing Projections) *Let $r_0 \in \mathbb{N}$, matrices $\mathbf{X}, \mathbf{X}', \mathbf{Y}, \mathbf{Y}' \in \mathbb{S}_>^p$, and positive numbers $\sigma > 0$ and $(\bar{\sigma}_i)_{i \in [r_0+1]}$ satisfy the following conditions:*

(a) For any $\mathbf{A} \in \{\mathbf{X}, \mathbf{X}', \mathbf{Y}, \mathbf{Y}'\}$, $\mu \mathbf{I}_p \preceq \mathbf{A} \preceq M\mathbf{I}_p$.

(b) $\max\{\|\mathbf{X} - \mathbf{X}'\|_{\mathrm{op}}, \|\mathbf{Y} - \mathbf{Y}'\|_{\mathrm{op}}\} \le \Delta$, where $\Delta \le \frac{\mu}{32r_0}(\mu/M)^2$.

(c) $\max_{i \in [r_0+1]} |\bar{\sigma}_i - \sigma_i(\mathbf{\Sigma})| \le \sigma/8r_0$, where $\mathbf{\Sigma} = \mathsf{CovBal}(\mathbf{X}, \mathbf{Y})$.

(d) $\sigma \in [\max\{\mu, 2\bar{\sigma}_{\hat{r}_0}\}, \frac{2}{3e}\bar{\sigma}_1]$.

Define $\mathbf{\Sigma}' = \mathsf{CovBal}(\mathbf{X}', \mathbf{Y}')$, $r = \mathsf{sep\text{-}rank}(r_0, \sigma; \mathbf{\Sigma}')$, $\mathbf{Q} = \mathsf{Proj}_{\mathrm{bal}}(r; \mathbf{X}, \mathbf{Y})$ and $\mathbf{Q}' = \mathsf{Proj}_{\mathrm{bal}}(r; \mathbf{X}', \mathbf{Y}')$. Then, r is well defined, $\mathbf{Q}$ and $\mathbf{Q}'$ are unique, and the following bounds hold:

$$\|\mathbf{Q}' - \mathbf{Q}\|_{\mathrm{op}} \le \frac{19r_0(M/\mu)^{5/2}\Delta}{\mu}, \quad \max\{\|\mathbf{Q}\|_{\mathrm{op}}, \|\mathbf{Q}'\|_{\mathrm{op}}\} \le \sqrt{M/\mu}.$$

Moreover, $\bar{\sigma}_r \ge \frac{3}{4}\sigma$, $\bar{\sigma}_{r+1} \le 3\sigma$, and $\bar{\sigma}_r - \bar{\sigma}_{r+1} \ge \frac{\bar{\sigma}_{\hat{r}}}{3r_0}$.

Proof [Proof of Proposition F.3] Throughout, we also set $\mathbf{\Sigma} = \mathsf{CovBal}(\mathbf{X}, \mathbf{Y})$, $\mathbf{W} = \Psi_{\mathrm{bal}}(\mathbf{Y}; \mathbf{X})$, and $\mathbf{W}' = \Psi_{\mathrm{bal}}(\mathbf{Y}'; \mathbf{X}')$. We further let $\mathbf{P}_r$ and $\mathbf{P}'_r$ denote the projections onto the top-r singular spaces of $\mathbf{\Sigma}$ and $\mathbf{\Sigma}'$, respectively.

By Lemma L.5 and that $\Delta \le \frac{\mu}{32r_0}(M/\mu)^2$ implies $\Delta \le \mu/3$, we have that

$$\|\mathbf{\Sigma} - \mathbf{\Sigma}'\|_{\mathrm{op}} \le 4(M/\mu)^2\Delta.$$

Hence, for $\Delta \le \frac{\mu}{32r_0}(M/\mu)^2$, it holds that

$$\|\mathbf{\Sigma} - \mathbf{\Sigma}'\|_{\mathrm{op}} \le \frac{\mu}{8r_0} \le \frac{\sigma}{8r_0}. \tag{L.8}$$

Moreover, condition (d) of the present proposition matches condition (ii) of Lemma L.10 (recall that lemma only requires one of conditions (i) or (ii) to be met). Thus, Lemma L.10 implies

$$\|\mathbf{P}_r - \mathbf{P}'_r\|_{\mathrm{op}} \le 4r_0\frac{\|\mathbf{\Sigma} - \mathbf{\Sigma}'\|_{\mathrm{op}}}{\sigma} \le \frac{16r_0(M/\mu)^2\Delta}{\sigma} \le \frac{16r_0(M/\mu)^2\Delta}{\mu}.$$

Thus, combining the above bound with Lemma L.5 and the norm bounds on $\mathbf{W}, \mathbf{W}'$, as well as their inverses due to Lemma L.4, it follows that

$$\begin{aligned}
&\|\mathbf{Q}' - \mathbf{Q}\|_{\mathrm{op}} \\
&= \|(\mathbf{W}')^{-\frac{1}{2}}\mathbf{P}'_r(\mathbf{W}')^{\frac{1}{2}} - \mathbf{W}^{-\frac{1}{2}}\mathbf{P}_r\mathbf{W}^{\frac{1}{2}}\|_{\mathrm{op}} \\
&\le \|(\mathbf{W}')^{-\frac{1}{2}}(\mathbf{P}'_r - \mathbf{P}_r)(\mathbf{W}')^{\frac{1}{2}}\|_{\mathrm{op}} + \|(\mathbf{W}')^{\frac{1}{2}}\mathbf{P}_r((\mathbf{W}')^{\frac{1}{2}} - (\mathbf{W})^{\frac{1}{2}})\|_{\mathrm{op}} + \|((\mathbf{W}')^{\frac{1}{2}} - \mathbf{W}^{\frac{1}{2}})\mathbf{P}_r(\mathbf{W})^{\frac{1}{2}}\|_{\mathrm{op}} \\
&\le \|\mathbf{P}'_r - \mathbf{P}_r\|_{\mathrm{op}}\|\mathbf{W}'\|_{\mathrm{op}} + 2\max\{\|\mathbf{W}\|_{\mathrm{op}}, \|\mathbf{W}'\|_{\mathrm{op}}\}^{1/2} \cdot \|(\mathbf{W}')^{\frac{1}{2}} - (\mathbf{W})^{\frac{1}{2}}\|_{\mathrm{op}} \\
&\le \frac{16r_0(M/\mu)^2\Delta}{\mu} \cdot (M/\mu)^{1/2} + 2 \cdot (M/\mu)^{1/4}\frac{3}{2\mu}(M/\mu)^{3/4}\Delta \\
&\le \frac{19r_0(M/\mu)^{5/2}\Delta}{\mu}.
\end{aligned}$$

Second, using $\|\mathbf{W}\|_{\mathrm{op}} \vee \|\mathbf{W}^{-1}\|_{\mathrm{op}} \le \sqrt{M/\mu}$ from Lemma L.4,

$$\|\mathbf{Q}\|_{\mathrm{op}} = \|\mathbf{W}^{-\frac{1}{2}}\mathbf{P}_r\mathbf{W}^{\frac{1}{2}}\|_{\mathrm{op}} \le \sqrt{\|\mathbf{W}\|_{\mathrm{op}}\|\mathbf{W}^{-1}\|_{\mathrm{op}}}\|\mathbf{P}_r\| \le \sqrt{M/\mu},$$

and similarly for $\|\mathbf{Q}'\|_{\mathrm{op}}$. Finally, we have from the assumption on $\sigma^\star_r$ and Weyl's inequality and Eq. (L.8) that

$$\begin{aligned}
\sigma^\star_r &\ge \sigma_r(\mathbf{\Sigma}) - \sigma/4 \ge \sigma_r(\mathbf{\Sigma}') - \|\mathbf{\Sigma} - \mathbf{\Sigma}'\|_{\mathrm{op}} - \frac{\sigma}{8r_0} \ge \sigma_r(\mathbf{\Sigma}') - \frac{\sigma}{4r_0} \\
\sigma^\star_{r+1} &\le \sigma_{r+1}(\mathbf{\Sigma}) + \sigma/4 \le \sigma_{r+1}(\mathbf{\Sigma}') + \|\mathbf{\Sigma} - \mathbf{\Sigma}'\|_{\mathrm{op}} + \frac{\sigma}{8r_0} \le \sigma_{r+1}(\mathbf{\Sigma}') + \frac{\sigma}{4r_0}.
\end{aligned}$$

From Lemma L.9, we have $\sigma_r(\mathbf{\Sigma}') \ge \sigma$ and $\sigma_{r+1}(\mathbf{\Sigma}') \le e\sigma$, so using $r_0 \ge 1$, we have $\sigma^\star_r \ge 3\sigma/4$ and $\sigma^\star_{r+1} \le (e + \frac{1}{4})\sigma \le 3\sigma$. Finally, from the definition of sep-rank (Definition F.7),

$$(1 - \frac{1}{r_0})\sigma_r(\mathbf{\Sigma}') - \sigma_{r+1}(\mathbf{\Sigma}') \ge 0.$$

Using the previous display this implies

$$(1 - \tfrac{1}{r_0})\sigma_r^\star - \sigma_{r+1}^\star \geq -(1 - \tfrac{1}{r_0})\frac{\sigma}{4r_0} - \frac{\sigma}{4r_0} \geq -\frac{\sigma}{2r_0},$$

so rearranging, and using $\sigma \leq 4\sigma_r^\star/3$ as derived above,

$$\sigma_r^\star - \sigma_{r+1}^\star \geq \frac{\sigma_r^\star}{r_0} - \frac{\sigma}{2r_0} \geq \frac{\sigma_r^\star}{r_0} - \frac{2\sigma_r^\star}{3r_0} = \frac{\sigma_r^\star}{3r_0}.$$

This completes the proof. ■

Appendix M. Supporting Proofs for Error Decomposition

In this section, we prove a slightly more specific statement of Proposition 4.1, which makes dependencies on the problem parameters explicit. We also state and prove an error decomposition result under more general assumptions which allow for additive slack, as described below.

The remainder of the section is structured as follows.

(a) Appendix M.1 states our main results, both under Assumptions 2.2 and 2.3, as well as under more general assumptions (Assumptions M.1, 2.2b and 2.3b) which allow for additive slack.

(b) Appendix M.2 sketches the main steps of the proof. The proofs of the constituent lemmas are deferred to subsequent sections. This section focuses on Proposition 4.1a, and mentions the modifications for Proposition 4.1b at its end in Subsubsection M.2.1.

(c) Appendix M.3 outlines helpful lemmas we refer to as "change of measure" lemmas. One key lemma uses the covariance-relation, Assumption 2.3, to relate certain expectation under $\mathcal{D}_{i\otimes 2}$ and $\mathcal{D}_{2\otimes j}$ to those under $\mathcal{D}_{i\otimes 1}$ and $\mathcal{D}_{1\otimes j}$.

(d) Appendices M.4 to M.6, prove the various lemmas given in Appendix M.2.

(e) Appendix M.7 derives Theorem 2 from Proposition 4.1 and Theorem 8.

M.1. Main results

We now give risk decomposition results which make dependencies on problem parameters explicit. Our granular guarantee under Assumptions 2.2 and 2.3 is as follows.

Proposition 4.1a (Final Error Decomposition, Explicit Dependence) *Suppose Assumptions 2.2 and 2.3 hold. For any $k \leq r$ with some fixed integer $r > 0$, and any aligned k-proxies (f, g) of the $\mathbb{R}^r$-embeddings $(\hat{f}, \hat{g})$, denote $\boldsymbol{\Delta}_0 = \boldsymbol{\Delta}_0(f, g, k)$ and $\boldsymbol{\Delta}_1 = \boldsymbol{\Delta}_1(f, g, k)$. Let $\sigma \leq \sigma_r(\hat{f}, \hat{g})$ be a lower bound on $\sigma_r(\hat{f}, \hat{g})$ as defined in Eq. (3.1), which satisfies $\sigma^2 \in (0, \mathsf{tail}_2^\star(k) + \boldsymbol{\Delta}_0 + \boldsymbol{\Delta}_{\mathrm{train}}]$. Then,*

$$\mathcal{R}(f, g; \mathcal{D}_{\mathrm{test}}) \lesssim \kappa_{\mathrm{tst}}\kappa_{\mathrm{cov}}^2 \left((\boldsymbol{\Delta}_1)^2 + \frac{1}{\sigma^2}\left(\boldsymbol{\Delta}_{\mathrm{apx}} + \boldsymbol{\Delta}_0 + \kappa_{\mathrm{cov}}\kappa_{\mathrm{trn}}\boldsymbol{\Delta}_{\mathrm{train}}\right)^2 \right).$$

In particular, suppressing polynomial dependence on $\kappa_{\mathrm{trn}}, \kappa_{\mathrm{cov}}$*, we recover*

$$\mathcal{R}(f, g; \mathcal{D}_{\mathrm{test}}) \lesssim_{\star} (\mathbf{\Delta}_1)^2 + \frac{1}{\sigma^2}(\mathsf{tail}_2^{\star}(k) + \mathbf{\Delta}_0 + \mathbf{\Delta}_{\mathrm{train}})^2.$$

The same bound holds more generally when $\mathsf{tail}_2^{\star}(k) + \mathbf{\Delta}_0 + \mathbf{\Delta}_{\mathrm{train}}$ *is replaced with an upper bound* M*, and when* σ^2 *need only satisfy* $\sigma^2 \le M$*. Moreover, it also holds that*

$$\mathcal{R}(f, g; \mathcal{D}_{1\otimes 1}) \le \kappa_{\mathrm{trn}} \mathcal{R}(f, g; \mathcal{D}_{\mathrm{train}}).$$

M.1.1. ERROR DECOMPOSITION UNDER ADDITIVE REMAINDERS

The above error decomposition holds under slightly more general conditions, which allow for additive *additive remainders* in the multiplicative approximations in Assumptions 2.3 and 2.4.

Assumption 2.2b (Coverage Decomposition with Additive Slack) *There exists* $\kappa_{\mathrm{tst}}, \kappa_{\mathrm{trn}} > 0$ *and* $\eta_{\mathrm{tst}}, \eta_{\mathrm{trn}} \in (0, 1]$ *such that* $\mathcal{D}_{\mathrm{train}}$ *covers all pairs* $\mathcal{D}_{i\otimes j}$ *with* $(i, j) = (1, 1)$*,* $(i, j) = (1, 2)$*, and* $(i, j) = (2, 1)$*, and* $\mathcal{D}_{\mathrm{test}}$ *is continuous with respect to the mixture of all pairs* $\mathcal{D}_{i\otimes j}$*. Formally, for all* $(x, y) \in \mathcal{X} \times \mathcal{Y}$*,*

$$\mathbb{P}_{\mathcal{D}_{i\otimes j}}\left[\frac{\mathrm{d}\mathcal{D}_{i\otimes j}(x, y)}{\mathrm{d}\mathcal{D}_{\mathrm{train}}(x, y)} > \kappa_{\mathrm{trn}}\right] \le \eta_{\mathrm{trn}}, \quad (i, j) \in \{(1, 1), (1, 2), (2, 1)\} \qquad \text{(Train Coverage)}$$

$$\mathbb{P}_{\mathcal{D}_{\mathrm{test}}}\left[\frac{\mathrm{d}\mathcal{D}_{\mathrm{test}}(x, y)}{\sum_{i,j\in\{1,2\}} \mathrm{d}\mathcal{D}_{i\otimes j}(x, y)} > \kappa_{\mathrm{tst}}\right] \le \eta_{\mathrm{tst}}. \qquad \text{(Test Coverage)}$$

Assumption 2.3b (Change of Covariance with Additive Slack) *There exists a* $\kappa_{\mathrm{cov}} \ge 1$ *and* $\eta_{\mathrm{cov}} \ge 0$ *such that, for any* $v \in \mathcal{H}$*,*

$$\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X},2}}[\langle f^{\star}(x), v\rangle_{\mathcal{H}}^2] \le \kappa_{\mathrm{cov}} \cdot \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X},1}}[\langle f^{\star}(x), v\rangle_{\mathcal{H}}^2] + \eta_{\mathrm{cov}}\|v\|_{\mathcal{H}}^2$$
$$\mathbb{E}_{y\sim\mathcal{D}_{\mathcal{Y},2}}[\langle g^{\star}(y), v\rangle_{\mathcal{H}}^2] \le \kappa_{\mathrm{cov}} \cdot \mathbb{E}_{y\sim\mathcal{D}_{\mathcal{Y},1}}[\langle g^{\star}(y), v\rangle_{\mathcal{H}}^2] + \eta_{\mathrm{cov}}\|v\|_{\mathcal{H}}^2.$$

For additive slack, we further require uniform boundedness of the embeddings (rather than just their inner products.)

Assumption M.1 (Boundedness) *There exists an upper bound* $B > 0$ *such that*

$$\max\left\{\sup_{x\in\mathcal{X}} \|f(x)\|_{\mathcal{H}} \vee \|f^{\star}(x)\|_{\mathcal{H}}, \sup_{y\in\mathcal{Y}} \|g(y)\|_{\mathcal{H}} \vee \|g^{\star}(y)\|_{\mathcal{H}}\right\} \le B.$$

We now state the general analogue of our error decomposition, Proposition 4.1b, which allows for additive slack terms. For simplicity, we assume $\eta_{\mathrm{cov}}, \eta_{\mathrm{tst}}, \eta_{\mathrm{trn}} \le 1$.

Proposition 4.1b (Final Error Decomposition with Additive Slack) *For any* $k \le r$ *with some fixed integer* $r > 0$*, and any aligned* k*-proxies* (f, g) *of the* $\mathbb{R}^r$*-embeddings* $(\hat{f}, \hat{g})$*, denote* $\mathbf{\Delta}_0 = \mathbf{\Delta}_0(f, g, k)$ *and* $\mathbf{\Delta}_1 = \mathbf{\Delta}_1(f, g, k)$*. Let* $\sigma = \sigma_r(\hat{f}, \hat{g})$ *(as in Eq. (3.1)), and let* σ *satisfy* $\sigma^2 \in (0, \mathsf{tail}_2^{\star}(k) + \mathbf{\Delta}_0]$*. Then, under Assumptions M.1, 2.2b and 2.3b, we have*

$$\mathcal{R}(f, g; \mathcal{D}_{\mathrm{test}}) \lesssim \kappa_{\mathrm{tst}} \kappa_{\mathrm{cov}}^2 \left((\mathbf{\Delta}_1)^2 + \frac{1}{\sigma^2}(\mathbf{\Delta}_{\mathrm{apx}} + \mathbf{\Delta}_0 + \kappa_{\mathrm{cov}}\kappa_{\mathrm{trn}}\mathbf{\Delta}_{\mathrm{train}})^2\right)$$
$$+ B^4 \mathrm{poly}(\kappa_{\mathrm{cov}}, \kappa_{\mathrm{tst}}, \kappa_{\mathrm{trn}}) \cdot (\eta_{\mathrm{tst}} + \eta_{\mathrm{trn}} + \eta_{\mathrm{cov}}).$$

M.2. Overview of proof

For now, we focus on Proposition 4.1a, which assumes Assumptions 2.2 and 2.3. The modification of Proposition 4.1b, under Assumptions M.1, 2.2b and 2.3b are described at the end. Each of the lemmas in this section is proved under these more general conditions.

Fix any embeddings $f : \mathcal{X} \to \mathcal{H}$ and $g : \mathcal{Y} \to \mathcal{H}$. We shall ultimately enforce that f, g are aligned k-proxies (Definition 4.2) for some $(\hat{f}, \hat{g})$, though this is only necessary for one step of the proof. We begin by recalling the error terms from Definition 4.3, and introducing a few other terms in our analysis.

Definition M.1 (Key Error Terms) *Given functions $f : \mathcal{X} \to \mathcal{H}$ and $g : \mathcal{X} \to \mathcal{H}$ and $k \in \mathbb{N}$, define*

$$\begin{aligned}
\boldsymbol{\Delta}_0(f,g,k) &:= \max\left\{\mathbb{E}_{\mathcal{D}_{1\otimes 1}}\left[\langle f_k^\star, g_k^\star - g\rangle^2\right], \mathbb{E}_{\mathcal{D}_{1\otimes 1}}\left[\langle f_k^\star - f, g_k^\star\rangle^2\right]\right\} && \text{(weighted error)}\\
\boldsymbol{\Delta}_1(f,g,k) &:= \max\left\{\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}\|f_k^\star - f\|^2, \mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}\|g_k^\star - g\|^2\right\} && \text{(unweighted error)}\\
\boldsymbol{\Delta}_2(f,g,k) &:= \max\left\{\mathbb{E}_{\mathcal{D}_{\mathcal{X},2}}\|f_k^\star - f\|, \mathbb{E}_{\mathcal{D}_{\mathcal{Y},2}}\|g_k^\star - g\|^2\right\} && \text{($\mathcal{D}_{2\otimes 2}$-recovery error)}\\
\boldsymbol{\Delta}_{\mathrm{apx}}(k) &:= \mathcal{R}(f_k^\star, g_k^\star; \mathcal{D}_{1\otimes 1}). && \text{(approximation error)}
\end{aligned}$$

When it is clear from the context, we will use the shorthand notation $\boldsymbol{\Delta}_0, \boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2$ and $\boldsymbol{\Delta}_{\mathrm{apx}}$, respectively, for convenience.

Above, $\boldsymbol{\Delta}_0(f,g,k)$ captures differences $g_k^\star - g$ (resp $f_k^\star - f$) weighted by $f_k^\star$ (resp. $g_k^\star$) under $\mathcal{D}_{1\otimes 1}$. $\boldsymbol{\Delta}_1$ captures the unweighted differences (i.e. in $\|\cdot\|$) under $\mathcal{D}_{1\otimes 1}$, and $\boldsymbol{\Delta}_2$ does the same under $\mathcal{D}_{2\otimes 2}$. Since $f^\star$ and $g^\star$ have spectral decay, we expect the unweighted errors $\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2$ to be larger than the weighted one $\boldsymbol{\Delta}_0$.

Bounding $\mathcal{D}_{\mathrm{test}}$-risk with $\mathcal{D}_{2\otimes 2}$-risk. We begin the proof with a lemma which bounds the risk under $\mathcal{D}_{\mathrm{test}}$ by the risk under the bottom-right block $\mathcal{D}_{2\otimes 2}$, plus the risk under $\mathcal{D}_{\mathrm{train}}$ (i.e. $\boldsymbol{\Delta}_{\mathrm{train}}$). The following is proved in Appendix M.4.

Lemma M.1 (Error Decomposition on $\mathcal{D}_{\mathrm{test}}$) *Under Assumption 2.2, the following holds for any $f : \mathcal{X} \to \mathcal{H}$ and $g : \mathcal{Y} \to \mathcal{H}$:*

$$\mathcal{R}(f,g;\mathcal{D}_{\mathrm{test}}) \le \kappa_{\mathrm{tst}}\left(\mathcal{R}(f,g;\mathcal{D}_{2\otimes 2}) + 3\kappa_{\mathrm{trn}}\mathcal{R}(f,g;\mathcal{D}_{\mathrm{train}})\right).$$

Bounding the $\mathcal{D}_{2\otimes 2}$-risk. The difficulty is now in handling the error on the bottom-right block $\mathcal{D}_{2\otimes 2}$. Our analysis reveals that the leading order term is precisely the weighted error $\boldsymbol{\Delta}_0$, with the unweighted errors $\boldsymbol{\Delta}_1$ and $\boldsymbol{\Delta}_2$ entering only in a quadratic way (i.e. at most second order) into the error. The following is proved in Appendix M.5.

Lemma M.2 (Error Decomposition on $\mathcal{D}_{2\otimes 2}$) *Under Assumptions 2.2 and 2.3, for any $f : \mathcal{X} \to \mathcal{H}$, $g : \mathcal{Y} \to \mathcal{H}$, and $k \in \mathbb{N}$*

$$\mathcal{R}(f,g;\mathcal{D}_{2\otimes 2}) \lesssim \kappa_{\mathrm{cov}}^2(\boldsymbol{\Delta}_0 + (\boldsymbol{\Delta}_1)^2 + \boldsymbol{\Delta}_{\mathrm{apx}}) + (\boldsymbol{\Delta}_2)^2 + \kappa_{\mathrm{cov}}\kappa_{\mathrm{trn}}\boldsymbol{\Delta}_{\mathrm{train}},$$

where above we suppress error term dependence on f, g, k.

Bounding $\mathbf{\Delta}_2$. We now turn to bounding $\mathbf{\Delta}_2$. This requires making full use of the assumption that (f, g) are aligned k-proxies of $(\hat{f}, \hat{g})$. Going forward, recall from Eq. (3.1) and the construction in Definition 4.2 that

$$\sigma_r(\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[ff^\top]) = \sigma_r(\hat{f}, \hat{g}) > 0,$$

where the positivity is a consequence of the assumption that $(\hat{f}, \hat{g})$ are full-rank. We may now bound $\mathbf{\Delta}_2$. The following is proved in Appendix M.6.

Lemma M.3 (Decomposition of $\mathbf{\Delta}_2$) *Suppose (f, g) are aligned k-proxies of $(\hat{f}, \hat{g})$. Then, we have*

$$\sigma_r(\hat{f}, \hat{g})\mathbf{\Delta}_2 \lesssim \left(\frac{\mathbf{\Delta}_{\text{train}}}{\omega_{\min}}\right) + \kappa_{\text{cov}}(\mathbf{\Delta}_0 + \mathbf{\Delta}_{\text{apx}}(k))$$

where above we suppress dependence on f, g, k in all error terms.

Concluding the proof of Proposition 4.1a. Lastly, we observe that the rank-k approximation error under $\mathcal{D}_{1\otimes 1}$ is precisely the tail term $\mathsf{tail}_2^\star(k)$.

Lemma M.4 *We have $\mathbf{\Delta}_{\text{apx}}(k) = \mathcal{R}(f_k^\star, g_k^\star; \mathcal{D}_{1\otimes 1}) = \mathsf{tail}_2^\star(k)$.*

Proof Using that projection matrices are self-adjoint and idempotent,

$$\begin{aligned}
\mathcal{R}(f_k^\star, g_k^\star; \mathcal{D}_{1\otimes 1}) &= \mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle f_k^\star(x), g_k^\star(y)\rangle - \langle f^\star(x), g^\star(y)\rangle)^2] \\
&= \mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle \mathbf{P}_k^\star f^\star(x), \mathbf{P}_k^\star g^\star(y)\rangle - \langle f^\star(x), g^\star(y)\rangle)^2] \\
&= \mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle \mathbf{P}_k^\star f^\star(x), g^\star(y)\rangle - \langle f^\star(x), g^\star(y)\rangle)^2] \\
&= \mathbb{E}_{\mathcal{D}_{1\otimes 1}}[\langle (\mathbf{I} - \mathbf{P}_k^\star) f^\star(x), g^\star(y)\rangle^2] \\
&= \mathrm{tr}[\mathbf{\Sigma}_{1\otimes 1}^\star(\mathbf{I} - \mathbf{P}_k^\star)\mathbf{\Sigma}_{1\otimes 1}^\star(\mathbf{I} - \mathbf{P}_k^\star)] = \sum_{j>k} \lambda_j(\mathbf{\Sigma}_{1\otimes 1}^\star)^2 := \mathsf{tail}_2^\star(k),
\end{aligned}$$

where in the last line, we use that $\mathbf{P}_k^\star$ projects onto the top k eigenvalues of $\mathbf{\Sigma}_{1\otimes 1}^\star$. ■

Putting these terms together reveals our final error decomposition result.

Proof [Proof of Proposition 4.1a] Let $\sigma \le \sigma_r(\hat{f}, \hat{g})$. We write

$$\begin{aligned}
&\mathcal{R}(f, g; \mathcal{D}_{\text{test}}) \\
&\overset{(i)}{\lesssim} \kappa_{\text{tst}}\left(\mathcal{R}(f, g; \mathcal{D}_{2\otimes 2}) + \kappa_{\text{trn}}\mathbf{\Delta}_{\text{train}}\right) \\
&\overset{(ii)}{\lesssim} \kappa_{\text{tst}}\left(\kappa_{\text{cov}}^2(\mathbf{\Delta}_0 + \mathbf{\Delta}_{\text{apx}} + (\mathbf{\Delta}_1)^2) + (\mathbf{\Delta}_2)^2 + \kappa_{\text{cov}}\kappa_{\text{trn}}\mathbf{\Delta}_{\text{train}}\right) + \kappa_{\text{trn}}\kappa_{\text{den}}\mathbf{\Delta}_{\text{train}} \\
&\lesssim \kappa_{\text{tst}}\left(\kappa_{\text{cov}}^2(\mathbf{\Delta}_0 + \mathbf{\Delta}_{\text{apx}} + (\mathbf{\Delta}_1)^2) + (\mathbf{\Delta}_2)^2 + \kappa_{\text{cov}}\kappa_{\text{trn}}\mathbf{\Delta}_{\text{train}}\right),
\end{aligned}$$

where in the last line, we absorb terms using $\kappa_{\text{cov}} \ge 1$. Continuing the string of inequalities,

$$\begin{aligned}
&\overset{(iii)}{\le} \kappa_{\text{tst}}\left(\kappa_{\text{cov}}^2(\mathbf{\Delta}_0 + \mathbf{\Delta}_{\text{apx}} + (\mathbf{\Delta}_1)^2) + \frac{\kappa_{\text{cov}}^2}{\sigma^2}(\mathbf{\Delta}_{\text{apx}} + \mathbf{\Delta}_0 + \kappa_{\text{cov}}\kappa_{\text{trn}}\mathbf{\Delta}_{\text{train}})^2 + \kappa_{\text{cov}}\kappa_{\text{trn}}\mathbf{\Delta}_{\text{train}}\right) \\
&\overset{(iv)}{\lesssim} \kappa_{\text{tst}}\left(\kappa_{\text{cov}}^2(\mathbf{\Delta}_1)^2 + \frac{\kappa_{\text{cov}}^2}{\sigma^2}(\mathbf{\Delta}_{\text{apx}} + \mathbf{\Delta}_0 + \kappa_{\text{cov}}\kappa_{\text{trn}}\mathbf{\Delta}_{\text{train}})^2\right) \\
&= \kappa_{\text{tst}}\kappa_{\text{cov}}^2\left((\mathbf{\Delta}_1)^2 + \frac{1}{\sigma^2}(\mathbf{\Delta}_{\text{apx}} + \mathbf{\Delta}_0 + \kappa_{\text{cov}}\kappa_{\text{trn}}\mathbf{\Delta}_{\text{train}})^2\right),
\end{aligned}$$

where (i) uses Lemma M.1, (ii) uses Lemma M.2, (iii) invokes Lemma M.3, and (iv) applies the assumption $\sigma^2 \le \mathbf{\Delta}_{\mathrm{apx}} + \mathbf{\Delta}_0 + \mathbf{\Delta}_{\mathrm{train}}$ and $\kappa_{\mathrm{cov}}, \kappa_{\mathrm{trn}} \ge 1$ to absorb the term

$$\kappa_{\mathrm{cov}}^2(\mathbf{\Delta}_0 + \mathbf{\Delta}_{\mathrm{apx}}) + \kappa_{\mathrm{cov}}\kappa_{\mathrm{trn}}\mathbf{\Delta}_{\mathrm{train}} \le \frac{\kappa_{\mathrm{cov}}^2}{\sigma^2}\left(\mathbf{\Delta}_{\mathrm{apx}} + \mathbf{\Delta}_0 + \kappa_{\mathrm{cov}}\kappa_{\mathrm{trn}}\mathbf{\Delta}_{\mathrm{train}}\right)^2,$$

loosing at most a constant factor of 2. The last inequality uses $\kappa_{\mathrm{cov}} \ge 1$. Note that the simplification also holds when likewise $\mathbf{\Delta}_{\mathrm{apx}} + \mathbf{\Delta}_0 + \mathbf{\Delta}_{\mathrm{train}}$ by an upper bound M, such that $\sigma^2 \le M$.

Lastly, the final statement of the proposition, namely the bound $\mathcal{R}(f, g; \mathcal{D}_{1\otimes 1}) \le \kappa_{\mathrm{trn}}\mathcal{R}(f, g; \mathcal{D}_{\mathrm{train}})$, is a direct consequence of Assumption 2.2. ■

M.2.1. MODIFICATIONS FOR ADDITIVE SLACK

Like their analogues in proving Proposition 4.1a, the following lemmas are proved in Appendices M.4 to M.6, respectively.

Lemma M.1b (Error Decomposition on $\mathcal{D}_{\mathrm{test}}$ with Additive Slack) *Under Assumptions M.1 and 2.2b,*

$$\mathcal{R}(f, g; \mathcal{D}_{\mathrm{test}}) \le \kappa_{\mathrm{tst}}\left(\mathcal{R}(f, g; \mathcal{D}_{2\otimes 2}) + 3\kappa_{\mathrm{trn}}\mathcal{R}(f, g; \mathcal{D}_{\mathrm{train}})\right) + 4B^4(\eta_{\mathrm{tst}} + 3\kappa_{\mathrm{tst}}\eta_{\mathrm{trn}}).$$

Lemma M.2b (Error Decomposition on $\mathcal{D}_{2\otimes 2}$ with Additive Slack) *Under Assumptions M.1, 2.2b and 2.3b, for any $k \in \mathbb{N}$,*

$$\mathcal{R}(f, g; \mathcal{D}_{2\otimes 2}) \lesssim \kappa_{\mathrm{cov}}^2(\mathbf{\Delta}_0 + (\mathbf{\Delta}_1)^2 + \mathbf{\Delta}_{\mathrm{apx}}) + (\mathbf{\Delta}_2)^2 + \kappa_{\mathrm{cov}}\kappa_{\mathrm{trn}}\mathbf{\Delta}_{\mathrm{train}} + B^4\kappa_{\mathrm{cov}}(\eta_{\mathrm{cov}} + \eta_{\mathrm{trn}}),$$

where above we suppress error term dependence on f, g, k.

Lemma M.3b (Decomposition of $\mathbf{\Delta}_2$ with Additive Slack) *Suppose $k \le r$, and (f, g) are aligned k-proxies for $(\hat{f}_r, \hat{g}_r)$ with $\mathbf{P}_k^\star$. Then for $k \le r$,*

$$\sigma_r(\hat{f}, \hat{g})\mathbf{\Delta}_2 \lesssim \left(\frac{\mathbf{\Delta}_{\mathrm{train}}}{\omega_{\mathrm{min}}}\right) + \kappa_{\mathrm{cov}}(\mathbf{\Delta}_0 + \mathbf{\Delta}_{\mathrm{apx}}(k)) + B^2(\eta_{\mathrm{cov}} + \eta_{\mathrm{tst}}),$$

where above we suppress dependence on f, g, k in all error terms.

Deriving Proposition 4.1b from the previous lemmas follows in much the same way as Proposition 4.1a, and is omitted for brevity.

M.3. Key change-of-measure lemmas

We begin by establishing some important change-of-measure results.

Lemma M.5 (Change of Covariance) *Under either the boundedness assumption Assumption M.1, or assumping $\eta_{\mathrm{cov}} = 0$, the following holds for any $i, j \in \{1, 2\}$ and any $(\tilde{f}, \tilde{g})$, under Assumption 2.3b,*

- $\mathbb{E}_{\mathcal{D}_{i\otimes 2}}[\langle \tilde{f}, g_k^\star\rangle^2] \le \kappa_{\mathrm{cov}}\mathbb{E}_{\mathcal{D}_{i\otimes 1}}[\langle \tilde{f}, g_k^\star\rangle^2] + B^2\eta_{\mathrm{cov}}$
- $\mathbb{E}_{\mathcal{D}_{2\otimes j}}[\langle f_k^\star, \tilde{g}\rangle^2] \le \kappa_{\mathrm{cov}}\mathbb{E}_{\mathcal{D}_{1\otimes j}}[\langle f_k^\star, \tilde{g}\rangle^2] + B^2\eta_{\mathrm{cov}}.$

The same holds if $g^\star_k$ (resp. $f^\star_k$) are replaced by $g^\star_{>k} := g^\star - g^\star_k$ (resp. $f^\star_{>k} := f^\star - f^\star_k$) or $g^\star$ (resp. $f^\star$), and under Assumption 2.3, the above holds with $\eta_{\mathrm{cov}} = 0$.

Proof Since Assumption 2.3 is stronger than Assumption 2.3b, we focus on the proofs under Assumption 2.3b. Let's begin by proving the first item under Assumption 2.3b; the extension to the second item is similar. We have

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_{i\otimes 2}}[\langle \tilde{f}, g^\star_k\rangle^2] &= \mathbb{E}_{\mathcal{D}_{\mathcal{X},i}}[\mathbb{E}_{\mathcal{D}_{\mathcal{Y},2}}[\langle \tilde{f}, g^\star_k\rangle^2]] && \text{(Fubini)}\\
&= \mathbb{E}_{\mathcal{D}_{\mathcal{X},i}}[\mathbb{E}_{\mathcal{D}_{\mathcal{Y},2}}[\langle \tilde{f}, \mathbf{P}^\star_k g^\star\rangle^2]]\\
&\le \mathbb{E}_{\mathcal{D}_{\mathcal{X},i}}[\kappa_{\mathrm{cov}}\mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}[\langle \mathbf{P}^\star_k\tilde{f}, g^\star\rangle^2] + \eta_{\mathrm{cov}}\|\mathbf{P}^\star_k\tilde{f}\|^2] && \text{(Assumption 2.3b)}\\
&\le \mathbb{E}_{\mathcal{D}_{\mathcal{X},i}}[\kappa_{\mathrm{cov}}\mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}[\langle \mathbf{P}^\star_k\tilde{f}, g^\star\rangle^2] + \eta_{\mathrm{cov}}\|\tilde{f}\|^2] && (\mathbf{P}^\star_k \text{ is a projection})\\
&\le \mathbb{E}_{\mathcal{D}_{\mathcal{X},i}}[\kappa_{\mathrm{cov}}\mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}[\langle \mathbf{P}^\star_k\tilde{f}, g^\star\rangle^2]] + B^2\eta_{\mathrm{cov}} && \text{(Assumption M.1)}\\
&= \kappa_{\mathrm{cov}}\mathbb{E}_{\mathcal{D}_{\mathcal{X},i}\otimes\mathcal{D}_{\mathcal{Y},1}}[\langle \mathbf{P}^\star_k\tilde{f}, g^\star\rangle^2] + B^2\eta_{\mathrm{cov}} && \text{(Fubini)}\\
&= \kappa_{\mathrm{cov}}\mathbb{E}_{\mathcal{D}_{\mathcal{X},i}\otimes\mathcal{D}_{\mathcal{Y},1}}[\langle \tilde{f}, g^\star_k\rangle^2] + B^2\eta_{\mathrm{cov}}.
\end{aligned}$$

As mentioned, the second item is similar. To derive the similar bounds for $g^\star - g^\star_k$, we use that $g^\star - g^\star_k = (\mathbf{I} - \mathbf{P}^\star_k)g^\star$, and $\mathbf{I} - \mathbf{P}^\star_k$ is also a projection operator; the bound for $f^\star - f^\star_k$ can be derived similarly. Finally, the bounds for $f^\star, g^\star$ are slightly simpler to establish, because we need not commute the projection operator. ■

Lemma M.6 (Change of Risk) *The following bounds hold:*

- *The risk on the "off-diagonal" product distribution is bounded by*

$$\mathcal{R}(f^\star_k, g^\star_k; \mathcal{D}_{1\otimes 2}) \vee \mathcal{R}(f^\star_k, g^\star_k; \mathcal{D}_{2\otimes 1}) \le \kappa_{\mathrm{cov}}\boldsymbol{\Delta}_{\mathrm{apx}}(k) + \eta_{\mathrm{cov}}B^2.$$

- *The risk on the "bottom-right" product distribution is bounded by*

$$\mathcal{R}(f^\star_k, g^\star_k; \mathcal{D}_{2\otimes 2}) \le \kappa_{\mathrm{cov}}^2\boldsymbol{\Delta}_{\mathrm{apx}}(k) + 2\kappa_{\mathrm{cov}}\eta_{\mathrm{cov}}B^2.$$

Proof Introduce the shorthand $f^\star_{>k} := f^\star - f^\star_k$ and $g^\star_{>k} := g^\star - g^\star_k$. Note that

$$f^\star_{>k} = (\mathbf{I} - \mathbf{P}^\star_k)f^\star, \quad g^\star_{>k} = (\mathbf{I} - \mathbf{P}^\star_k)g^\star,$$

and hence $f^\star_{>k}, g^\star_{>k}$ are B-bounded under Assumption M.1.

Observe that since $\mathbf{P}^\star_k$ is an orthogonal projection, so is $\mathbf{I} - \mathbf{P}^\star_k$. Since orthogonal projections are self-adjoint and idempotent,

$$\begin{aligned}
h^\star - \langle f^\star_k, g^\star_k\rangle &= \langle f^\star, g^\star\rangle - \langle \mathbf{P}^\star_k f^\star, g^\star\rangle = \langle (\mathbf{I} - \mathbf{P}^\star_k)f^\star, g^\star\rangle\\
&= \langle (\mathbf{I} - \mathbf{P}^\star_k)^{\mathsf{H}}(\mathbf{I} - \mathbf{P}^\star_k)f^\star, g^\star\rangle = \langle (\mathbf{I} - \mathbf{P}^\star_k)f^\star, (\mathbf{I} - \mathbf{P}^\star_k)g^\star\rangle = \langle f^\star_{>k}, g^\star_{>k}\rangle.
\end{aligned} \tag{M.1}$$

Thus, by Lemma M.5 and the fact that $f^\star_{>k}$ is B-bounded, we have

$$\mathcal{R}(f^\star_k, g^\star_k; \mathcal{D}_{1\otimes 2}) = \mathbb{E}_{\mathcal{D}_{1\otimes 2}}(h^\star - \langle f^\star_k, g^\star_k\rangle)^2 = \mathbb{E}_{\mathcal{D}_{1\otimes 2}}\langle f^\star_{>k}, g^\star_{>k}\rangle^2 \le \kappa_{\mathrm{cov}}\mathbb{E}_{\mathcal{D}_{1\otimes 1}}\langle f^\star_{>k}, g^\star_{>k}\rangle^2 + \eta_{\mathrm{cov}}B^2.$$

Similarly, $\mathcal{R}(f_k^\star, g_k^\star; \mathcal{D}_{2\otimes 1}) \le \kappa_{\mathrm{cov}} \mathbb{E}_{\mathcal{D}_{1\otimes 1}} \langle f_{>k}^\star, g_{>k}^\star \rangle^2 + \eta_{\mathrm{cov}} B^2$. Finally, by two applications of Lemma M.5, we have

$$\begin{aligned}
\mathcal{R}(f_k^\star, g_k^\star; \mathcal{D}_{2\otimes 2}) = \mathbb{E}_{\mathcal{D}_{2\otimes 2}} (h^\star - \langle f_k^\star, g_k^\star \rangle)^2 &= \mathbb{E}_{\mathcal{D}_{2\otimes 2}} \langle f_{>k}^\star, g_{>k}^\star \rangle^2 \\
&\le \kappa_{\mathrm{cov}} \mathbb{E}_{\mathcal{D}_{1\otimes 2}} \langle f_{>k}^\star, g_{>k}^\star \rangle^2 + \eta_{\mathrm{cov}} B^2 \\
&\le \kappa_{\mathrm{cov}}^2 \mathbb{E}_{\mathcal{D}_{1\otimes 1}} \langle f_{>k}^\star, g_{>k}^\star \rangle^2 + (\kappa_{\mathrm{cov}} \eta_{\mathrm{cov}} + \eta_{\mathrm{cov}}) B^2 \\
&\le \kappa_{\mathrm{cov}}^2 \mathbb{E}_{\mathcal{D}_{1\otimes 1}} \langle f_{>k}^\star, g_{>k}^\star \rangle^2 + 2\kappa_{\mathrm{cov}} \eta_{\mathrm{cov}} B^2 \\
&= \kappa_{\mathrm{cov}}^2 \mathbb{E}_{\mathcal{D}_{1\otimes 1}} (h^\star - \langle f_k^\star, g_k^\star \rangle)^2 + 2\kappa_{\mathrm{cov}} \eta_{\mathrm{cov}} B^2 \qquad \text{(Eq. (M.1))} \\
&= \kappa_{\mathrm{cov}}^2 \mathcal{R}(f_k^\star, g_k^\star; \mathcal{D}_{1\otimes 1}) + 2\kappa_{\mathrm{cov}} \eta_{\mathrm{cov}} B^2.
\end{aligned}$$

This completes the proof. ■

Lemma M.7 *Suppose the boundedness assumption, i.e. Assumption M.1 holds for some* (f, g)*. Then, for any* $(i, j) \in \{(1,1), (1,2), (2,1)\}$,

$$\mathcal{R}(f, g; \mathcal{D}_{i\otimes j}) \le 4B^4 \epsilon_{\mathrm{trn}} + \kappa_{\mathrm{trn}} \mathcal{R}(f, g; \mathcal{D}_{\mathrm{train}}).$$

The same also holds without Assumption M.1, with $\epsilon_{\mathrm{trn}} = 0$ *(ignoring the* $4B^2\epsilon_{\mathrm{trn}}$ *term).*

Proof Define the event

$$\mathcal{E}_{\mathrm{train}, i\otimes j} := \left\{ \frac{\mathrm{d}\mathcal{D}_{i\otimes j}(x,y)}{\mathrm{d}\mathcal{D}_{\mathrm{train}}(x,y)} \le \kappa_{\mathrm{trn}} \right\}.$$

We first consider the case where Assumption M.1 holds. To this end, we consider any function $F : \mathcal{X} \times \mathcal{Y} \to [0, M]$. We then have

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_{i\otimes j}}[F(x,y)] &\le M\, \mathbb{P}_{\mathcal{D}_{i\otimes j}}[\neg \mathcal{E}_{\mathrm{train}, i\otimes j}] + \mathbb{E}_{\mathcal{D}_{i\otimes j}}[F(x,y) \mathbb{I}\{\mathcal{E}_{\mathrm{train}, i\otimes j}\}] \\
&= M\, \mathbb{P}_{\mathcal{D}_{i\otimes j}}[\neg \mathcal{E}_{\mathrm{train}, i\otimes j}] + \mathbb{E}_{\mathcal{D}_{\mathrm{train}}}[F(x,y) \mathbb{I}\{\mathcal{E}_{\mathrm{train}, i\otimes j}\} \cdot \frac{\mathrm{d}\mathcal{D}_{i\otimes j}(x,y)}{\mathrm{d}\mathcal{D}_{\mathrm{train}}(x,y)}] \\
&\le M\, \mathbb{P}_{\mathcal{D}_{i\otimes j}}[\neg \mathcal{E}_{\mathrm{train}, i\otimes j}] + \mathbb{E}_{\mathcal{D}_{\mathrm{train}}}[F(x,y) \frac{1}{\kappa_{\mathrm{trn}}}] \\
&= \kappa_{\mathrm{trn}} M\, \mathbb{P}_{\mathcal{D}_{i\otimes j}}[\neg \mathcal{E}_{\mathrm{train}, i\otimes j}] + \mathbb{E}_{\mathcal{D}_{\mathrm{train}}}[F(x,y)] \le M\epsilon_{\mathrm{trn}} + \kappa_{\mathrm{trn}} \mathbb{E}_{\mathcal{D}_{\mathrm{train}}}[F(x,y)].
\end{aligned}$$

The result follows by setting $F(x,y) = (\langle f(x), g(y) \rangle - h^\star(x,y))^2$, which lies in $[0, 4B^4]$ by Lemma M.8, stated just below.

For the case Assumption M.1 does not hold but with $\epsilon_{\mathrm{trn}} = 0$, the first term on the right-hand side above does not appear, which completes the proof. ■

Lemma M.8 *Under Assumption M.1, given any* B*-bounded functions* f, g*, the function* $F(x,y) = (\langle f(x), g(y) \rangle - h^\star(x,y))^2$ *satisfies* $0 \le F(x,y) \le 4B^4$.

Proof Since f, g are B-bounded $|\langle f, g \rangle| \le \|f\|\|g\| \le B^2$. Similarly, $|h^\star| \le \|f^\star\|\|g^\star\| \le B^2$ The bound follows. ■

M.4. Proof of Lemmas M.1 and M.1b

Proof We prove the more general statement under Assumption 2.2b, and explain the modification to Assumption 2.2 afterward. Define the event

$$\mathcal{E}_{\text{test}} := \left\{ \frac{\mathrm{d}\mathcal{D}_{\text{test}}(x,y)}{\sum_{i,j\in\{1,2\}} \mathrm{d}\mathcal{D}_{i\otimes j}(x,y)} \le \kappa_{\text{tst}} \right\}.$$

Then, for any bounded, nonnegative function $F(x,y) : \mathcal{X} \times \mathcal{Y} \to [0, M]$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_{\text{test}}}[F(x,y)] \\
&\le M\,\mathbb{P}_{\mathcal{D}_{\text{test}}}[\neg\mathcal{E}_{\text{test}}] + \mathbb{E}_{\mathcal{D}_{\text{test}}}[\mathbb{I}\{\mathcal{E}_{\text{test}}\}F(x,y)] \\
&= M\,\mathbb{P}_{\mathcal{D}_{\text{test}}}[\neg\mathcal{E}_{\text{test}}] + \int_{(x,y)} \left(F(x,y) \cdot \left(\sum_{i,j\in\{1,2\}} \mathrm{d}\mathcal{D}_{i\otimes j}(x,y) \right) \cdot \frac{\mathrm{d}\mathcal{D}_{\text{test}}(x,y)}{\sum_{i,j\in\{1,2\}} \mathrm{d}\mathcal{D}_{i\otimes j}(x,y)} \mathbb{I}\{\mathcal{E}_{\text{test}}\} \right) \\
&\le M\,\mathbb{P}_{\mathcal{D}_{\text{test}}}[\neg\mathcal{E}_{\text{test}}] + \kappa_{\text{tst}} \int_{(x,y)} F(x,y) \cdot \left(\sum_{i,j\in\{1,2\}} \mathrm{d}\mathcal{D}_{i\otimes j}(x,y) \right) \\
&= M\,\mathbb{P}_{\mathcal{D}_{\text{test}}}[\neg\mathcal{E}_{\text{test}}] + \kappa_{\text{tst}} \sum_{i,j=1}^{2} \mathbb{E}_{\mathcal{D}_{i\otimes j}}[F(x,y)] \\
&\le M\eta_{\text{tst}} + \kappa_{\text{tst}} \sum_{i,j=1}^{2} \mathbb{E}_{\mathcal{D}_{i\otimes j}}[F(x,y)].
\end{aligned}$$

Taking $F(x,y) = (\langle f(x), g(y)\rangle - h^\star(x,y))$, which takes values in $[0, 4B^4]$ by Lemma M.8, we find

$$\mathcal{R}(f,g;\mathcal{D}_{\text{test}}) \le 4B^4\eta_{\text{tst}} + \kappa_{\text{tst}} \sum_{i,j=1}^{2} \mathbb{E}_{\mathcal{D}_{i\otimes j}}[F(x,y)].$$

By Lemma M.7, we bound

$$\sum_{i,j\neq(2,2)} \mathcal{R}(f,g;\mathcal{D}_{i\otimes j}) \le 12B^4\eta_{\text{trn}} + 3\kappa_{\text{trn}}\mathcal{R}(f,g;\mathcal{D}_{\text{train}}).$$

Therefore,

$$\mathcal{R}(f,g;\mathcal{D}_{\text{test}}) \le \kappa_{\text{tst}} \left(\mathcal{R}(f,g;\mathcal{D}_{2\otimes 2}) + 3\kappa_{\text{trn}}\mathcal{R}(f,g;\mathcal{D}_{\text{train}})\right) + 4B^4(\eta_{\text{tst}} + 3\kappa_{\text{tst}}\eta_{\text{trn}}).$$

The bound follows. To obtain the simpler statement with Assumption 2.2, under which we can take $\eta_{\text{tst}} = \eta_{\text{trn}} = 0$, and complete the proof. ■

M.5. Proof of Lemmas M.2 and M.2b

We begin with an elementary algebraic lemma which helps us expand the risk $\mathcal{R}(f,g;\mathcal{D}_{2\otimes 2})$.

Lemma M.9 *For any $h^\star : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, $f_1, f_2 : \mathcal{X} \to \mathcal{H}$, and $g_1, g_2 : \mathcal{Y} \to \mathcal{H}$, we have*

$$(\langle f_1, g_1\rangle - h^\star)^2 \le 2(\langle f_2, g_2\rangle - h^\star)^2 + 6\langle f_1 - f_2, g_2\rangle^2 + 6\langle f_2, g_1 - g_2\rangle^2 + 6\|f_1 - f_2\|^2\|g_1 - g_2\|^2.$$

Proof [Proof of Lemma M.9] Set $h_1 = \langle f_1, g_1\rangle$ and $h_2 = \langle f_2, g_2\rangle$. Then,

$$\begin{aligned}(h_1 - h^\star)^2 - (h_2 - h^\star)^2 &= (h_1 - h^\star + h_2 - h^\star)(h_1 - h_2)\\ &= (h_1 - h_2)^2 + 2(h_2 - h^\star)(h_1 - h_2)\\ &\le 2(h_1 - h_2)^2 + (h_2 - h^\star)^2.\end{aligned}$$

Hence, we have $(h_1 - h^\star)^2 \le 2(h_1 - h_2)^2 + 2(h_2 - h^\star)^2$. To conclude, we bound

$$\begin{aligned}(h_1 - h_2)^2 &= (\langle f_1, g_1\rangle - \langle f_2, g_2\rangle)^2\\ &= (\langle f_1 - f_2, g_2\rangle + \langle f_2, g_1 - g_2\rangle + \langle f_1 - f_2, g_1 - g_2\rangle)^2\\ &\le 3\langle f_1 - f_2, g_2\rangle^2 + 3\langle f_2, g_1 - g_2\rangle^2 + 3\langle f_1 - f_2, g_1 - g_2\rangle^2\\ &\le 3\langle f_1 - f_2, g_2\rangle^2 + 3\langle f_2, g_1 - g_2\rangle^2 + 3\|f_1 - f_2\|^2\|g_1 - g_2\|^2.\end{aligned}$$

Combining the two displays completes the proof. ■

Step 1. Change of covariance under $\mathcal{D}_{2\otimes 2}$. Taking $f_1 = f$, $g_1 = g$, $f_2 = f_k^\star$ and $g_2 = g_k^\star$, Lemma M.9 implies

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}_{2\otimes 2}}[(\langle f, g\rangle - h^\star)^2] - 2\mathbb{E}_{\mathcal{D}_{2\otimes 2}}[(\langle f_k^\star, g_k^\star\rangle - h^\star)^2]\\ &\le 6\left(\mathbb{E}_{\mathcal{D}_{2\otimes 2}}\langle f - f_k^\star, g_k^\star\rangle^2 + \mathbb{E}_{\mathcal{D}_{2\otimes 2}}\langle f_k^\star, g - g_k^\star\rangle^2 + \mathbb{E}_{\mathcal{D}_{2\otimes 2}}[\|f - f_k^\star\|^2\|g - g_k^\star\|^2]\right)\\ &\overset{(i)}{\le} 6\kappa_{\mathrm{cov}}\left(\mathbb{E}_{\mathcal{D}_{2\otimes 1}}\langle f - f_k^\star, g_k^\star\rangle^2 + \mathbb{E}_{\mathcal{D}_{1\otimes 2}}\langle f_k^\star, g - g_k^\star\rangle^2\right) + 6\mathbb{E}_{\mathcal{D}_{2\otimes 2}}\|f - g_k^\star\|^2\|g - g_k^\star\|^2 + 48B^2\eta_{\mathrm{cov}}\\ &\overset{(ii)}{=} 6\kappa_{\mathrm{cov}}\left(\mathbb{E}_{\mathcal{D}_{2\otimes 1}}\langle f - f_k^\star, g_k^\star\rangle^2 + \mathbb{E}_{\mathcal{D}_{1\otimes 2}}\langle f_k^\star, g - g_k^\star\rangle^2\right) + 6\mathbb{E}_{\mathcal{D}_{\mathcal{X},2}}\|f - f_k^\star\|^2 \cdot \mathbb{E}_{\mathcal{D}_{\mathcal{Y},2}}\|f - f_k^\star\|^2 + 48B^4\eta_{\mathrm{cov}}\\ &\overset{(iii)}{\le} 6\kappa_{\mathrm{cov}}\left(\mathbb{E}_{\mathcal{D}_{2\otimes 1}}\langle f - f_k^\star, g_k^\star\rangle^2 + \mathbb{E}_{\mathcal{D}_{1\otimes 2}}\langle f_k^\star, g - g_k^\star\rangle^2\right) + 6(\boldsymbol{\Delta}_2)^2 + 48B^4\eta_{\mathrm{cov}},\end{aligned} \tag{M.2}$$

where in (i) we apply Lemma M.5 to the terms $\mathbb{E}_{\mathcal{D}_{2\otimes 2}}\langle f - f_k^\star, g_k^\star\rangle^2$ and $\mathbb{E}_{\mathcal{D}_{2\otimes 2}}\langle f_k^\star, g - g_k^\star\rangle^2$, for with $\tilde{B} = 2B$, in (ii) we use that $\mathcal{D}_{2\otimes 2} = \mathcal{D}_{\mathcal{X},2} \otimes \mathcal{D}_{\mathcal{Y},2}$ is a product measure, and in (iii) we recall the definition of $\boldsymbol{\Delta}_2 = \boldsymbol{\Delta}_2(f, g, k)$.

Step 2. Expansion of $\mathcal{D}_{1\otimes 2}$ and $\mathcal{D}_{2\otimes 1}$. Next, we expand the first two terms in Eq. (M.2). First, $\langle f - f_k^\star, g_k^\star\rangle = \langle f, g_k^\star\rangle - h_k^\star = \langle f, g\rangle - h_k^\star + \langle f, g_k^\star - g\rangle = \langle f, g\rangle - h_k^\star + \langle f - f_k^\star, g_k^\star - g\rangle + \langle f_k^\star, g_k^\star - g\rangle$. Hence,

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}_{2\otimes 1}}\langle f - f_k^\star, g_k^\star\rangle^2\\ &\le 3\mathbb{E}_{\mathcal{D}_{2\otimes 1}}[(\langle f, g\rangle - h_k^\star)^2] + 3\mathbb{E}_{\mathcal{D}_{2\otimes 1}}\langle f_k^\star, g_k^\star - g\rangle^2 + 3\mathbb{E}_{\mathcal{D}_{2\otimes 1}}\langle f - f_k^\star, g_k^\star - g\rangle^2\\ &\overset{(i)}{\le} 3\mathbb{E}_{\mathcal{D}_{2\otimes 1}}[(\langle f, g\rangle - h_k^\star)^2] + 3\kappa_{\mathrm{cov}}\mathbb{E}_{\mathcal{D}_{1\otimes 1}}\langle f_k^\star, g_k^\star - g\rangle^2 + 3\mathbb{E}_{\mathcal{D}_{2\otimes 1}}\|f_k^\star - f\|^2\|g_k^\star - g\|^2 + 12B^4\eta_{\mathrm{cov}},\end{aligned}$$

where in (i) we again apply Lemma M.5. We can further expand

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_{2\otimes 1}}\|f_k^\star - f\|^2\|g_k^\star - g\|^2 &= \mathbb{E}_{\mathcal{D}_{\mathcal{X},2}}\|f_k^\star - f\|^2 \cdot \mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}\|g_k^\star - g\|^2 \\ &\le \frac{1}{2\kappa_{\mathrm{cov}}}(\mathbb{E}_{\mathcal{D}_{\mathcal{X},2}}\|f_k^\star - f\|^2)^2 + \frac{\kappa_{\mathrm{cov}}}{2}(\mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}\|g_k^\star - g\|^2)^2 \\ &\le \frac{1}{2\kappa_{\mathrm{cov}}}(\mathbf{\Delta}_2)^2 + \frac{\kappa_{\mathrm{cov}}}{2}(\mathbf{\Delta}_1)^2,\end{aligned}$$

where again, we recall the definition of $\mathbf{\Delta}_2$ and $\mathbf{\Delta}_1$ in Definition M.1. In sum, we find

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}_{2\otimes 1}}\langle f - f_k^\star, g_k^\star\rangle^2 \\ &\le 3\mathbb{E}_{\mathcal{D}_{2\otimes 1}}[(\langle f, g\rangle - h_k^\star)^2] + 3\kappa_{\mathrm{cov}}\left(\mathbb{E}_{\mathcal{D}_{1\otimes 1}}\langle f_k^\star, g_k^\star - g\rangle^2 + \frac{1}{2}(\mathbf{\Delta}_1)^2\right) + \frac{3}{2\kappa_{\mathrm{cov}}}(\mathbf{\Delta}_2)^2 + 12B^4\eta_{\mathrm{cov}}.\end{aligned}$$

A similar analysis bounds

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}_{1\otimes 2}}\langle f_k^\star, g - g_k^\star\rangle^2 \\ &\overset{(i)}{\le} 3\mathbb{E}_{\mathcal{D}_{1\otimes 2}}[(\langle f, g\rangle - h_k^\star)^2] + 3\kappa_{\mathrm{cov}}(\mathbb{E}_{\mathcal{D}_{1\otimes 1}}\langle f - f_k^\star, g_k^\star\rangle^2 + \frac{1}{2}(\mathbf{\Delta}_1)^2) + \frac{3}{2\kappa_{\mathrm{cov}}}(\mathbf{\Delta}_2)^2 + 12B^4\eta_{\mathrm{cov}}.\end{aligned}$$

Thus, defining

$$\Delta_{\mathrm{off}} = \mathbb{E}_{\mathcal{D}_{1\otimes 2}}[(\langle f, g\rangle - h_k^\star)^2] + \mathbb{E}_{\mathcal{D}_{2\otimes 1}}[(\langle f, g\rangle - h_k^\star)^2],$$

we have

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}_{2\otimes 1}}\langle f - f_k^\star, g_k^\star\rangle^2 + \mathbb{E}_{\mathcal{D}_{1\otimes 2}}\langle f_k^\star, g - g_k^\star\rangle^2 \\ &\le 3\mathbb{E}_{\mathcal{D}_{1\otimes 2}}[(\langle f, g\rangle - h_k^\star)^2] + 3\mathbb{E}_{\mathcal{D}_{2\otimes 1}}[(\langle f, g\rangle - h_k^\star)^2] \\ &\quad + 3\kappa_{\mathrm{cov}}(\mathbb{E}_{\mathcal{D}_{1\otimes 1}}\langle f - f_k^\star, g_k^\star\rangle^2 + \mathbb{E}_{\mathcal{D}_{1\otimes 1}}\langle f_k^\star, g - g_k^\star\rangle^2 + (\mathbf{\Delta}_1)^2) + \frac{3}{\kappa_{\mathrm{cov}}}(\mathbf{\Delta}_2)^2 + 24B^4\eta_{\mathrm{cov}} \\ &= 3\Delta_{\mathrm{off}} + 3\kappa_{\mathrm{cov}}(2\mathbf{\Delta}_0 + (\mathbf{\Delta}_1)^2) + \frac{3}{\kappa_{\mathrm{cov}}}(\mathbf{\Delta}_2)^2 + 24B^4\eta_{\mathrm{cov}}.\end{aligned} \tag{M.3}$$

Step 3. Intermediate simplification. Combining Eqs. (M.3) and (M.2), we find

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}_{2\otimes 2}}[(\langle f, g\rangle - h^\star)^2] - 2\mathbb{E}_{\mathcal{D}_{2\otimes 2}}[(\langle f_k^\star, g_k^\star\rangle - h^\star)^2] \\ &\le 6\kappa_{\mathrm{cov}}\left(\mathbb{E}_{\mathcal{D}_{2\otimes 1}}\langle f - f_k^\star, g_k^\star\rangle^2 + \mathbb{E}_{\mathcal{D}_{1\otimes 2}}\langle f_k^\star, g - g_k^\star\rangle^2\right) + 6(\mathbf{\Delta}_2)^2 + 48B^4\eta_{\mathrm{cov}} \\ &\le 18\kappa_{\mathrm{cov}}\Delta_{\mathrm{off}} + 18\kappa_{\mathrm{cov}}^2(2\mathbf{\Delta}_0 + (\mathbf{\Delta}_1)^2) + 24(\mathbf{\Delta}_2)^2 + (144\kappa_{\mathrm{cov}} + 48)B^4\eta_{\mathrm{cov}}.\end{aligned}$$

That is, by rearranging

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_{2\otimes 2}}[(\langle f, g\rangle - h^\star)^2] \le 18\kappa_{\mathrm{cov}}^2(2\mathbf{\Delta}_0 + (\mathbf{\Delta}_1)^2) + 24(\mathbf{\Delta}_2)^2 + (144\kappa_{\mathrm{cov}} + 48)B^4\eta_{\mathrm{cov}} \\ + 18\kappa_{\mathrm{cov}}\Delta_{\mathrm{off}} + 2\mathbb{E}_{\mathcal{D}_{2\otimes 2}}[(\langle f_k^\star, g_k^\star\rangle - h^\star)^2].\end{aligned} \tag{M.4}$$

Step 4. Concluding the proof. To conclude, we upper bound Eq. (M.4). We begin by noting that, by Lemma M.6,

$$2\mathbb{E}_{\mathcal{D}_{2\otimes 2}}[(\langle f_k^\star, g_k^\star\rangle - h^\star)^2] = 2\mathcal{R}(f_k^\star, g_k^\star; \mathcal{D}_{2\otimes 2}) \le 2\kappa_{\mathrm{cov}}^2 \mathcal{R}(f_k^\star, g_k^\star; \mathcal{D}_{1\otimes 1}) + 4\kappa_{\mathrm{cov}}\eta_{\mathrm{cov}} B^4.$$

Similarly, again by Lemma M.6

$$\begin{aligned}
\Delta_{\mathrm{off}} &:= \mathbb{E}_{\mathcal{D}_{1\otimes 2}}[(\langle f, g\rangle - h_k^\star)^2] + \mathbb{E}_{\mathcal{D}_{2\otimes 1}}[(\langle f, g\rangle - h_k^\star)^2] \\
&\le 2\mathbb{E}_{\mathcal{D}_{1\otimes 2}}[(\langle f, g\rangle - h^\star)^2] + 2\mathbb{E}_{\mathcal{D}_{2\otimes 1}}[(\langle f, g\rangle - h^\star)^2] \\
&\quad + 2\underbrace{\mathbb{E}_{\mathcal{D}_{1\otimes 2}}[(h_k^\star - h^\star)^2]}_{=\mathcal{R}(f_k^\star, g_k^\star; \mathcal{D}_{1\otimes 2})} + 2\underbrace{\mathbb{E}_{\mathcal{D}_{2\otimes 1}}[(h_k^\star - h^\star)^2]}_{=\mathcal{R}(f_k^\star, g_k^\star; \mathcal{D}_{2\otimes 1})} \\
&\le 2\mathbb{E}_{\mathcal{D}_{1\otimes 2}}[(\langle f, g\rangle - h^\star)^2] + 2\mathbb{E}_{\mathcal{D}_{2\otimes 1}}[(\langle f, g\rangle - h^\star)^2] \\
&\quad + 4\kappa_{\mathrm{cov}}\mathcal{R}(f_k^\star, g_k^\star; \mathcal{D}_{1\otimes 1}) + 4\eta_{\mathrm{cov}} B^4.
\end{aligned}$$

Hence,

$$\begin{aligned}
&18\kappa_{\mathrm{cov}}\Delta_{\mathrm{off}} + 2\mathbb{E}_{\mathcal{D}_{2\otimes 2}}[(\langle f_k^\star, g_k^\star\rangle - h^\star)^2] \\
&\le (4\cdot 18 + 2)\kappa_{\mathrm{cov}}^2 \mathcal{R}(f_k^\star, g_k^\star; \mathcal{D}_{1\otimes 1}) + (4\cdot 18 + 4)\kappa_{\mathrm{cov}}\eta_{\mathrm{cov}} B^5 \\
&\qquad + (2\cdot 18)\kappa_{\mathrm{cov}}\left(\mathbb{E}_{\mathcal{D}_{1\otimes 2}}[(\langle f, g\rangle - h^\star)^2] + \mathbb{E}_{\mathcal{D}_{2\otimes 1}}[(\langle f, g\rangle - h^\star)^2]\right). \\
&= 74\kappa_{\mathrm{cov}}^2 \mathcal{R}(f_k^\star, g_k^\star; \mathcal{D}_{1\otimes 1}) + 76\kappa_{\mathrm{cov}}\eta_{\mathrm{cov}} B^5 + 36\kappa_{\mathrm{cov}}\left(\mathcal{R}(f, g; \mathcal{D}_{1\otimes 2}) + \mathcal{R}(f, g; \mathcal{D}_{2\otimes 1})\right).
\end{aligned}$$

By Lemma M.7, and using that $0 \le (\langle f, g\rangle - h^\star)^2 \le 4B^4$

$$(\mathcal{R}(f, g; \mathcal{D}_{1\otimes 2}) + \mathcal{R}(f, g; \mathcal{D}_{2\otimes 1})) \le 2\kappa_{\mathrm{trn}}\mathcal{R}(f, g; \mathcal{D}_{\mathrm{train}}) + 8B^4\eta_{\mathrm{trn}},$$

Thus,

$$\begin{aligned}
&18\kappa_{\mathrm{cov}}\Delta_{\mathrm{off}} + 2\mathbb{E}_{\mathcal{D}_{2\otimes 2}}[(\langle f_k^\star, g_k^\star\rangle - h^\star)^2] \\
&\le 74\kappa_{\mathrm{cov}}^2 \mathcal{R}(f_k^\star, g_k^\star; \mathcal{D}_{1\otimes 1}) + 72\kappa_{\mathrm{cov}}\kappa_{\mathrm{trn}}\mathcal{R}(f, g; \mathcal{D}_{\mathrm{train}}) + (8\cdot 36)\eta_{\mathrm{trn}}\kappa_{\mathrm{cov}} B^4 + 76\kappa_{\mathrm{cov}}\eta_{\mathrm{cov}} B^4.
\end{aligned}$$

In sum

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}_{2\otimes 2}}[(\langle f, g\rangle - h^\star)^2] \\
&\le 18\kappa_{\mathrm{cov}}^2(2\boldsymbol{\Delta}_0 + (\boldsymbol{\Delta}_1)^2) + 24(\boldsymbol{\Delta}_2)^2 + (144\kappa_{\mathrm{cov}} + 48)B^4\eta_{\mathrm{cov}} \\
&\qquad + 18\kappa_{\mathrm{cov}}\Delta_{\mathrm{off}} + 2\mathbb{E}_{\mathcal{D}_{2\otimes 2}}[(\langle f_k^\star, g_k^\star\rangle - h^\star)^2] \\
&\le 18\kappa_{\mathrm{cov}}^2(2\boldsymbol{\Delta}_0 + (\boldsymbol{\Delta}_1)^2) + 24(\boldsymbol{\Delta}_2)^2 + 72\kappa_{\mathrm{cov}}\kappa_{\mathrm{trn}}\mathcal{R}(f, g; \mathcal{D}_{\mathrm{train}}) \\
&\qquad + 74\kappa_{\mathrm{cov}}^2 \mathcal{R}(f_k^\star, g_k^\star; \mathcal{D}_{1\otimes 1}) + (144\kappa_{\mathrm{cov}} + 48)B^4\eta_{\mathrm{cov}} + 288\eta_{\mathrm{trn}}\kappa_{\mathrm{cov}} B^4 + 76\kappa_{\mathrm{cov}}\eta_{\mathrm{cov}} B^4 \\
&\le 18\kappa_{\mathrm{cov}}^2(2\boldsymbol{\Delta}_0 + (\boldsymbol{\Delta}_1)^2) + 24(\boldsymbol{\Delta}_2)^2 + 72\kappa_{\mathrm{cov}}\kappa_{\mathrm{trn}}\underbrace{\mathcal{R}(f, g; \mathcal{D}_{\mathrm{train}})}_{=\boldsymbol{\Delta}_{\mathrm{train}}} \\
&\qquad + 74\kappa_{\mathrm{cov}}^2 \underbrace{\mathcal{R}(f_k^\star, g_k^\star; \mathcal{D}_{1\otimes 1})}_{=\boldsymbol{\Delta}_{\mathrm{apx}}} + 268\kappa_{\mathrm{cov}} B^4\eta_{\mathrm{cov}} + 288\eta_{\mathrm{trn}}\kappa_{\mathrm{cov}} B^4,
\end{aligned}$$

where in the last line, we used $\kappa_{\mathrm{cov}} \ge 1$ and $144 + 48 + 76 = 268$. Dropping constants and simplifying,

$$\begin{aligned}
\mathcal{R}(f, g; \mathcal{D}_{2\otimes 2}) &= \mathbb{E}_{\mathcal{D}_{2\otimes 2}}[(\langle f, g\rangle - h^\star)^2] \\
&\lesssim \kappa_{\mathrm{cov}}^2(\boldsymbol{\Delta}_0 + (\boldsymbol{\Delta}_1)^2 + \boldsymbol{\Delta}_{\mathrm{apx}}) + (\boldsymbol{\Delta}_2)^2 + \kappa_{\mathrm{cov}}\kappa_{\mathrm{trn}}\boldsymbol{\Delta}_{\mathrm{train}} + B^4\kappa_{\mathrm{cov}}(\eta_{\mathrm{cov}} + \eta_{\mathrm{trn}}).
\end{aligned}$$

The proof for Lemma M.2 follows by setting $\eta_{\mathrm{trn}} = \eta_{\mathrm{cov}} = 0$. ■

M.6. Proof of Lemmas M.3 and M.3b

Recall the definitions

$$\begin{aligned}\boldsymbol{\Delta}_0 &:= \max\left\{\mathbb{E}_{\mathcal{D}_{1\otimes 1}}\left[\langle f_k^\star, g_k^\star - g\rangle^2, \mathbb{E}_{\mathcal{D}_{1\otimes 1}}\langle f_k^\star, g_k^\star - g\rangle^2\right]\right\}\\ \boldsymbol{\Delta}_1 &:= \max\left\{\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}\|f_k^\star - f\|^2, \mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}\|g_k^\star - g\|^2\right\}\\ \boldsymbol{\Delta}_2 &:= \max\left\{\mathbb{E}_{\mathcal{D}_{\mathcal{X},2}}\|f_k^\star - f\|^2, \mathbb{E}_{\mathcal{D}_{\mathcal{Y},2}}\|g_k^\star - g\|^2\right\}\\ \boldsymbol{\Delta}_{\mathrm{train}}(f,g) &:= \mathcal{R}(f,g;\mathcal{D}_{\mathrm{train}})\\ \boldsymbol{\Delta}_{\mathrm{apx}}(k) &:= \mathcal{R}(f_k^\star, g_k^\star; \mathcal{D}_{1\otimes 1})\end{aligned}$$

where the dependence of f, g, k is suppressed in all $\boldsymbol{\Delta}_{(\cdot)}$ terms. Our aim is to bound $\boldsymbol{\Delta}_2$. We focus on bounding $\mathbb{E}_{\mathcal{D}_{\mathcal{Y},2}}\|g_k^\star - g\|^2$, for the bound on $\mathbb{E}_{\mathcal{D}_{\mathcal{X},2}}\|f_k^\star - f\|^2$ is analogous.

Further, let us recall what it measn for (f,g) to be aligned k-proxies. This means that (a) $f = (\iota_r \circ \mathbf{T}^{-1})\hat{f}$, $g = (\iota_r \circ \mathbf{T})\hat{g}$, where $\iota_r : \mathbb{R}^r \to \mathcal{H}$ is an isometric inclusion, and $\mathbf{T}$ is the balancing operator of Lemma 4.3, and (b) for $\mathbf{P}_k^\star$ projection onto the top k-eigenvectors of $\boldsymbol{\Sigma}_{1\otimes 1}^\star$, we have

$$\mathrm{range}(\mathbf{P}_k^\star) \subseteq \mathrm{range}(\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[ff^\top]). \tag{M.5}$$

In particular, let $\mathcal{V} := \mathrm{range}(\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[ff^\top])$. Since $\hat{f}, \hat{g}$ are full-rank, $\mathcal{V} = \mathrm{range}(\iota_r) = \mathrm{range}(\mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}[gg^\top])$. Moreover, $\mathrm{range}(\mathbb{E}_{\mathcal{D}_{\mathcal{Y},1}}[g_k^\star(y)g_k^\star(y)^\top]) = \mathrm{range}(\mathbf{P}_k^\star) \subseteq \mathcal{V}_r$. Hence, By Lemma L.7 , $g(x)\ g_k^\star(y) \in \mathcal{V}$ almost surely, and thus, $g_k^\star(y) - g(y) \in \mathcal{V}$ with probability one. In addition, since $\mathcal{V} = \mathrm{range}(\iota_r)$ has dimension r, it follows that for any $\mathbf{v} \in \mathcal{V}$, and since $\sigma_r(\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[ff^\top]) = \sigma_r(\hat{f},\hat{g})$ in view of the construction in Definition 4.2,

$$\mathbf{v}^\top \mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}[ff^\top]\mathbf{v} \geq \|\mathbf{v}\|^2 \cdot \sigma_r(\hat{f},\hat{g}).$$

Therefore,

$$\begin{aligned}\sigma_r(\hat{f},\hat{g})\mathbb{E}_{\mathcal{D}_{\mathcal{Y},2}}\|g_k^\star - g\|^2 &\leq \mathbb{E}_{\mathcal{D}_{\mathcal{Y},2}}\left[\frac{1}{\sigma_r(\hat{f},\hat{g})}\mathbb{E}_{\mathcal{D}_{\mathcal{X},1}}\langle f, g_k^\star - g\rangle^2\right]\\ &= \frac{1}{\sigma_r(\hat{f},\hat{g})}\mathbb{E}_{\mathcal{D}_{1\otimes 2}}\left[\langle f, g - g_k^\star\rangle^2\right].\end{aligned}$$

In other words, we bound $\mathbb{E}_{\mathcal{D}_{\mathcal{Y},2}}\|g_k^\star - g\|^2$ by relating an expectation involving $\mathcal{D}_{\mathcal{X},1}$. Now, we can further expand

$$\begin{aligned}\langle f, g - g_k^\star\rangle &= \langle f, g\rangle - \langle f, g_k^\star\rangle = \langle f, g\rangle - \langle f_k^\star, g_k^\star\rangle - \langle f - f_k^\star, g_k^\star\rangle\\ &= (\langle f, g\rangle - h^\star) - (\langle f_k^\star, g_k^\star\rangle - h^\star) - \langle f - f_k^\star, g_k^\star\rangle.\end{aligned}$$

Hence,

$$\begin{aligned}\sigma_r(\hat{f},\hat{g})\mathbb{E}_{\mathcal{D}_{\mathcal{Y},2}}\|g_k^\star - g\|^2 &\leq 3\mathbb{E}_{\mathcal{D}_{1\otimes 2}}(\langle f, g\rangle - h^\star)^2 + 3\mathbb{E}_{\mathcal{D}_{1\otimes 2}}(\langle f_k^\star, g_k^\star\rangle - h^\star)^2 + 3\mathbb{E}_{\mathcal{D}_{1\otimes 2}}\langle f - f_k^\star, g_k^\star\rangle^2\\ &\leq 3\mathcal{R}(f,g;\mathcal{D}_{1\otimes 2}) + 3\mathcal{R}(f_k^\star, g_k^\star; \mathcal{D}_{1\otimes 2}) + 3\mathbb{E}_{\mathcal{D}_{1\otimes 2}}\langle f - f_k^\star, g_k^\star\rangle^2.\end{aligned}$$

By Lemma M.5 and the fact that $f - f_k^\star$ is $2B$-bounded,

$$\mathbb{E}_{\mathcal{D}_{1\otimes 2}}\langle f - f_k^\star, g_k^\star\rangle^2 \leq \kappa_{\mathrm{cov}}\mathbb{E}_{\mathcal{D}_{1\otimes 1}}\langle f - f_k^\star, g_k^\star\rangle^2 + 4B^4\eta_{\mathrm{cov}} = \kappa_{\mathrm{cov}}\boldsymbol{\Delta}_0 + 4B^4\eta_{\mathrm{cov}}.$$

By Lemma M.6,

$$\mathcal{R}(f_k^\star, g_k^\star; \mathcal{D}_{1\otimes 2}) \le \kappa_{\mathrm{cov}} \mathcal{R}(f_k^\star, g_k^\star; \mathcal{D}_{1\otimes 1}) + \eta_{\mathrm{cov}} B^4 = \kappa_{\mathrm{cov}} \mathbf{\Delta}_{\mathrm{apx}} + \eta_{\mathrm{cov}} B^4.$$

Finally, by applying Lemma M.7,

$$\mathcal{R}(f, g; \mathcal{D}_{1\otimes 2}) \le 4B^4 \eta_{\mathrm{tst}} + \kappa_{\mathrm{trn}} \mathcal{R}(f, g; \mathcal{D}_{\mathrm{train}}) = 4B^4 \eta_{\mathrm{tst}} + \kappa_{\mathrm{trn}} \mathbf{\Delta}_{\mathrm{train}}.$$

Thus,

$$\sigma_r(\hat{f}, \hat{g}) \mathbb{E}_{\mathcal{D}_{\mathcal{Y},2}} \| g_k^\star - g \|^2 \le 3 \left(\frac{\mathbf{\Delta}_{\mathrm{train}}}{\omega_{\mathrm{min}}} \right) + 3\kappa_{\mathrm{cov}} (\mathbf{\Delta}_0 + \mathbf{\Delta}_{\mathrm{apx}}(k)) + 12B^2(\eta_{\mathrm{cov}} + \eta_{\mathrm{tst}}).$$

This completes the proof. ■

M.7. Proof of Theorem 2

Set $\epsilon^2 = \epsilon_{1\otimes 1}^2$. For any $s \in \mathbb{N}, \epsilon > 0$ satisfying $s < \|\mathbf{\Sigma}_{1\otimes 1}^\star\|_{\mathrm{op}}/40\epsilon$ and

$$\epsilon^2 \ge \inf_{s' \ge s-1} \mathbb{E}_{\mathcal{D}_{1\otimes 1}}[(\langle f, g \rangle - \langle f_{s'}^\star, g_{s'}^\star \rangle^2)],$$

by Theorem 8, we can always find a k for which

$$\begin{aligned} (\mathbf{\Delta}_0(f, g, k) + \mathsf{tail}_2^\star(k) + \mathbf{\Delta}_{\mathrm{train}}(f, g))^2 &\lesssim s^6 \epsilon^4 + s^2 (\boldsymbol{\sigma}_s^\star)^4 + \mathsf{tail}_2^\star(s)^2 + \epsilon_{\mathrm{trn}}^4 \\ \mathbf{\Delta}_1(f, g, k)^2 &\lesssim (r + s^4)\epsilon^2 + s^2 (\boldsymbol{\sigma}_s^\star)^2 + \mathsf{tail}_1^\star(s)^2. \end{aligned} \tag{M.6}$$

Proposition 4.1a ensures that, with the choice of $s = r + 1$,

$$\begin{aligned}
&\mathcal{R}(f, g; \mathcal{D}_{\mathrm{test}}) \\
&\lesssim_\star (\mathbf{\Delta}_1)^2 + \frac{1}{\hat{\sigma}_r^2} (\mathsf{tail}_2^\star(k) + \mathbf{\Delta}_0 + \mathbf{\Delta}_{\mathrm{train}})^2 \\
&\le (\mathbf{\Delta}_1)^2 + \frac{\alpha}{(\boldsymbol{\sigma}_r^\star)^2} (\mathsf{tail}_2^\star(k) + \mathbf{\Delta}_0 + \mathbf{\Delta}_{\mathrm{train}})^2 \\
&\lesssim \left((r + s^4)\epsilon^2 + s^2 (\boldsymbol{\sigma}_s^\star)^2 + \mathsf{tail}_1^\star(s)^2 \right) + \alpha \cdot \frac{s^6 \epsilon^4 + \epsilon_{\mathrm{trn}}^4 + s^2 (\boldsymbol{\sigma}_s^\star)^4 + \mathsf{tail}_2^\star(s)^2}{(\boldsymbol{\sigma}_r^\star)^2} \\
&\qquad\qquad (s = r + 1 \lesssim r) \\
&\lesssim \left(r^4 \epsilon^2 + r^2 (\boldsymbol{\sigma}_{r+1}^\star)^2 + \mathsf{tail}_1^\star(r+1)^2 \right) + \alpha \cdot \frac{r^6 \epsilon^4 + \epsilon_{\mathrm{trn}}^4 + r^2 (\boldsymbol{\sigma}_{r+1}^\star)^4 + \mathsf{tail}_2^\star(r+1)^2}{(\boldsymbol{\sigma}_r^\star)^2} \\
&\le \left(r^4 \epsilon^2 + (1+\alpha) r^2 (\boldsymbol{\sigma}_{r+1}^\star)^2 + \mathsf{tail}_1^\star(r+1)^2 \right) + \alpha \cdot \frac{r^6 \epsilon^4 + \epsilon_{\mathrm{trn}}^4 + \mathsf{tail}_2^\star(r+1)^2}{(\boldsymbol{\sigma}_r^\star)^2} \\
&\le \left(r^4 \epsilon^2 + (1+\alpha) r^2 (\boldsymbol{\sigma}_{r+1}^\star)^2 + \mathsf{tail}_1^\star(r)^2 \right) + \alpha \cdot \frac{r^6 \epsilon^4 + \epsilon_{\mathrm{trn}}^4 + \mathsf{tail}_2^\star(r)^2}{(\boldsymbol{\sigma}_r^\star)^2} \qquad \text{(Monoticity of } \mathsf{tail}_q^\star) \\
&\lesssim \left(r^4 \epsilon^2 + \alpha r^2 (\boldsymbol{\sigma}_{r+1}^\star)^2 + \mathsf{tail}_1^\star(r)^2 \right) + \alpha \cdot \frac{r^6 \epsilon^4 + \epsilon_{\mathrm{trn}}^4 + \mathsf{tail}_2^\star(r)^2}{(\boldsymbol{\sigma}_r^\star)^2}. \qquad (\alpha \ge 1)
\end{aligned}$$

The last statement of the theorem - upper bounding $\alpha \le 2$, is precisely the last statement of Theorem 8. ■