Statistical curriculum learning: An elimination algorithm achieving an oracle risk

Omer Cohen OMER.COH@CAMPUS.TECHNION.AC.IL
Ron Meir RMEIR@EE.TECHNION.AC.IL
Nir Weinberger NIRWEIN@TECHNION.AC.IL
Department of Electrical and Computer Engineering, Technion – Israel Institute of Technology.

Editors: Shipra Agrawal and Aaron Roth

Abstract

We consider a statistical version of curriculum learning (CL) in a parametric prediction setting. The learner is required to estimate a target parameter vector, and can adaptively collect samples from either the target model, or other source models that are similar to the target model, but less noisy. We consider three types of learners, depending on the level of side-information they receive. The first two, referred to as strong/weak-oracle learners, receive high/low degrees of information about the models, and use these to learn. The third, a fully adaptive learner, estimates the target parameter vector without any prior information. In the single source case, we propose an elimination learning method, whose risk matches that of a strong-oracle learner. In the multiple source case, we advocate that the risk of the weak-oracle learner is a realistic benchmark for the risk of adaptive learners. We develop an adaptive multiple elimination-rounds CL algorithm, and characterize instance-dependent conditions for its risk to match that of the weak-oracle learner. We consider instance-dependent minimax lower bounds, and discuss the challenges associated with defining the class of instances for the bound. We derive two minimax lower bounds, and determine the conditions under which the performance weak-oracle learner is minimax optimal.

Keywords: Adaptive learning, best-arm identification, convergence rates, curriculum learning, elimination algorithm, lifelong learning, minimax lower bounds, multi-task learning.

1. Introduction

Curriculum learning (CL) (Bengio et al., 2009) is a machine learning paradigm that is successfully utilized in various applications, *e.g.*, weakly supervised object localization, object detection, neural machine translation, and reinforcement learning (Soviany et al., 2022; Wang et al., 2021). The distinctive feature of CL is the ability to control the learning *order*, *e.g.*, to choose which task to begin with, to begin from easy training examples, or to begin from a smooth loss function. This leads to an *optimization benefit.* In parallel, various *lifelong learning* methods also involve learning from multiple tasks, but aim to achieve a *statistical benefit*, *e.g.*, *multi-task learning, transfer learning*, and c*ontinual learning* (Caruana, 1997; Ruder, 2017; Zhang and Yang, 2021; Crawshaw, 2020; Zenke et al., 2017; Van de Ven and Tolias, 2019; De Lange et al., 2021), but without the challenging aspect of optimization of the order. Theoretical aspects of these methods were explored in, *e.g.* (Baxter, 2000; Argyriou et al., 2006; Maurer et al., 2016; Du et al., 2020; Tripuraneni et al., 2020, 2021; Evron et al., 2023; Goldfarb and Hand, 2023; Li et al., 2023; Lin et al., 2023; Goldfarb et al., 2024). In comparison, CL was much less explored Weinshall et al. (2018); Hacohen and Weinshall (2019); Weinshall and Amir (2020); Saglietti et al. (2022); Cornacchia and Mossel (2023), and much less from the statistical aspect. We survey these and other related works in detail in Appendix A.

Recently, Xu and Tewari (2022) advocated the *statistical benefits* of CL, by considering a setting in which T source models indexed by $t \in [T] := \{1, 2, \ldots, T\}$ are available to the learner, and model $t = 0$ is a *target* task to be learned. The learner has a budget of N samples for all models. Each model can be rated by both its *similarity* to the target task $t = 0$, and its *statistical difficulty*. A source model is useful for learning the target task if it is both similar to the target task and statistically easier to learn than the target task. Since the similarity of the tasks (and sometimes the difficulty) are unknown in advance, the algorithmic challenge is to set up a curriculum that optimally utilizes the source tasks to learn the target task. In this paper, we focus on a Gaussian mean estimation problem in $\mathbb{R}^d$, in which similarity between models is measured by the distance between parameter vectors, and difficulty by the noise variance. Since the mean parameters are unknown in advance, so is the similarity between them. Xu and Tewari (2022) introduced a *strong-oracle*, which reveals to a learner the source model that optimally balances between similarity and difficulty. They proved a lower bound on its risk (Xu and Tewari, 2022, Theorem 1), which is also a lower bound on the risk of any CL algorithm (which operates without this knowledge).

Contributions We address both the algorithmic and statistical challenges of CL, and show that even in a parametric learning setting, inherent challenges and intriguing aspects are encountered:

- We show that the risk of the strong-oracle learner can be achieved for $T = 1$. Interestingly, the learner does not know that its risk is *strictly* better than the risk obtained when just using samples from the target model, even if this is the case.
- We introduce a notion of a *weak-oracle learner,* and postulate that the risk of an ideal CL algorithm should match the risk of the weak-oracle learner when $T > 1$.
- We develop an adaptive *multiple-rounds source-elimination* CL algorithm, where in each round source models are eliminated based on estimated similarity. The possible elimination of source models allows for improved similarity estimation in the following round, and thus the elimination of additional source models. We introduce the *elimination curve*, which determines whether the risk of the weak-oracle learner will be achieved by the algorithm. We reveal that the risk is *not* monotonic in the similarity of the source models to the target model.
- We derive two lower bounds on the minimax risk of any CL algorithm. We reveal *intrinsic* hurdles in choosing a set of problem instances of *homogeneous* difficulty for the minimax bound. The first bound is for the set proposed by Xu and Tewari (2022, Theorem 2) and strictly improves it, and the second bound is for a newly introduced set, and matches the risk of the weak-oracle learner.

2. Problem setting

The statistical curriculum learning problem We use standard notation conventions outlined in Appendix B. Let $(\Omega, \mathfrak{B}(\Omega))$ be a Borel space, Φ_t an index set, and $\mathcal{M}_t := \{P_{\phi_t}\}_{\phi_t \in \Phi_t}$ a statistical model, where ϕ_t is an index in Φ_t and P_{ϕ_t} is a probability measure on the Borel space. Let $\mathcal{M}_0$ be the *target model*, and let $\{\mathcal{M}_t\}_{t \in [T]}$ be T *source models*. Let $\mathbf{\Phi} := \Phi_0 \times \Phi_1 \times \cdots \Phi_T$ be the product set of the index sets, and $\boldsymbol{\phi} := (\phi_0, \phi_1, \ldots, \phi_T) \in \mathbf{\Phi}$ be the collection of the indices of the $T + 1$ models. The learner may adaptively collect samples from each of these models, up to a total of N samples. To wit, the learner chooses the next model to sample from based on previous

model choices and observations. Let $A_i \in [\![T]\!] := \{0\} \cup [T]$ denote the model index chosen by the learner for the ith sample, and let S_i be the ith sample. Thus, the observations made by the learner are $H_N := (A_1, S_1, A_2, S_2, \ldots, A_N, S_N)$. We adopt the *random table model* from the closely-related *multi-armed bandit* problem (Lattimore and Szepesvári, 2020, Section 4.6), in which $N \cdot (T+1)$ independent r.v.'s $\boldsymbol{Z} := \{Z_{i,t}\}_{i\in[N],t\in[\![T]\!]}$ are defined over the Borel space $(\mathcal{Z}^{N(T+1)}, \mathfrak{B}(\mathcal{Z}^{N(T+1)}))$, so that the law of $Z_{i,t}$ is P_{ϕ_t}. The ith sample collected by the learner is $S_i = Z_{i,A_i}$, where A_i is $\sigma(H_{i-1})$-measurable, *i.e.*, A_i only depends on the *history* H_{i-1}, and S_i is sampled from the model $\mathcal{M}_{A_i}$. The policy π_i of the learner for the ith sample is the distribution of A_i conditioned on H_{i-1}, and its *policy* is $\pi := (\pi_i)_{i\in[N]}$. We consider the well-specified setting, in which the learner produces an hypothesis $\hat{\phi}\colon ([\![T]\!] \times \mathcal{Z})^N \to \Phi_0$ based on the collected samples $(A_i, S_i)_{i\in[N]}$. Let $\ell\colon \Phi_0 \times \Phi_0 \to \mathbb{R}_+$ be a loss function. A CL algorithm is $\mathcal{A} := (\pi, \hat{\phi})$, and its excess risk of after collecting N samples when the parameters are $\boldsymbol{\phi}$ is

$$L_N(\boldsymbol{\phi}, \mathcal{A}) := \mathbb{E}_{\boldsymbol{\phi},\mathcal{A}}\left[\ell\left(\hat{\phi}\left((A_i, S_i)_{i\in[N]}\right), \phi_0\right) - \ell\left(\phi_0, \phi_0\right)\right],$$

where the expectation is w.r.t. the randomness of the samples $\boldsymbol{Z}$, the policy, and possibly the estimator. The *minimax excess risk* of $\boldsymbol{\Psi} \subseteq \boldsymbol{\Phi}$ is given by (with a slight abuse of notation)

$$L_N(\boldsymbol{\Psi}) := \inf_{\mathcal{A}} \sup_{\boldsymbol{\phi}\in\boldsymbol{\Psi}} L_N(\boldsymbol{\phi}, \mathcal{A}). \tag{1}$$

The CL problem is to derive tight bounds on $L_N(\boldsymbol{\Psi})$, and a develop CL algorithms that achieve it.

The mean estimation statistical CL problem In order to focus on the intrinsic aspects of the statistical CL problem, we focus on a simple parametric mean estimation problem, prototypical to parametric learning problems. Specifically, the sample space is $\mathcal{Z} = \mathbb{R}^d$ and $\Phi_t := \mathbb{R}^d \times \mathbb{R}_+ \times \overline{\mathbb{S}}_d^{++}$, where $\overline{\mathbb{S}}_d^{++}$ is the positive semidefinite cone of matrices whose trace is normalized to d, and the parameters are $\phi_t := (\theta_t, \sigma_t^2, \overline{\Sigma}_t)$. The tth model, $t \in [\![T]\!]$, is given by

$$\mathcal{M}_t\colon \quad Y_t = \theta_t + \epsilon_t, \tag{2}$$

where $\theta_t \in \mathbb{R}^d$ is an unknown parameter vector and $\epsilon_t \sim \mathcal{N}(0, \sigma_t^2 \cdot \overline{\Sigma}_t)$ is a Gaussian noise with $\mathrm{Tr}[\overline{\Sigma}_t] = d$. The goal of the learner is to estimate the *target parameter* ϕ_0 under the squared error loss function $\ell(\phi_0, \phi) = \|\theta_0 - \theta\|^2$, that is, the parameters $(\sigma_t^2, \overline{\Sigma}_t)$ are *nuisance*. Thus, the excess risk is $L_N(\boldsymbol{\phi}, \mathcal{A}) = \mathbb{E}[\|\hat{\theta} - \theta_0\|^2]$. We assume for the most part that $\{(\sigma_t^2, \overline{\Sigma}_t)\}_{t\in[\![T]\!]}$ are known to the learner in advance (and extend to unknown $\{(\sigma_t^2, \overline{\Sigma}_t)\}_{t\in[\![T]\!]}$ in Appendix H). We assume w.l.o.g. that $\sigma_t^2 < \sigma_0^2$ for all $t \in [T]$, since otherwise there is no reason for the learner to sample from the tth source model – it is both mismatched, and has larger noise variance than the target model. We also make the mild assumption that $\max_{t\in[\![T]\!]} \|\theta_t\| \le C_\theta$, for some constant $C_\theta > 0$. As is well-known (*e.g.*, Rigollet and Hütter (2019, Theorem 2.2)), estimators based on N i.i.d. samples from a parametric model lead to a *parametric error rate*. Specifically, there exists an estimator $\bar{\theta}(N)$ and $g(\delta)\colon [0,1] \to \mathbb{R}_+$ so that for any $\delta \in (0,1)$,

$$\|\bar{\theta}(N) - \theta\|^2 \le g(\delta) \cdot \frac{d\sigma^2}{N} \tag{3}$$

holds with probability at least $1-\delta$. For the model (2), the empirical mean estimator achieves that with $g(\delta) = c\log(e/\delta)$ for some universal constant $c > 1$ (see Lemma 7 in Appendix C).

Xu and Tewari (2022) considered the closely-related linear regression problem, as we describe in Appendix D. We show that under the assumptions on the covariance matrix of the covariates made by Xu and Tewari (2022) (Assumption 1 in Appendix D), accurate estimation of the parameter vector is necessary for low risk (and not just sufficient). We then show using the small-ball method of Koltchinskii and Mendelson (2015), that (3) holds for the projected least squares (LS) estimator with $g(\delta) = \Theta(\log^2(\delta))$, under a mild condition on (N, d). So, under Assumption 1, the mean estimation and linear regression problems are not very different, and we thus focus on the former, which is simpler. Our results hold almost verbatim for the latter, with slight adjustments.

3. The similarity–difficulty balance and oracle-based learners

The gain in estimating θ_0 based on samples from a source model is related to their similarity and their difficulty. For parametric models, similarity can be gauged by the distances $\{Q_t\}_{t\in[T[}$ where $Q_t := \|\theta_t - \theta_0\|$ (and $Q_0 := 0$), and difficulty by the noise variances $\{\sigma_t^2\}_{t\in[\![T]\!]}$. Assuming, for now, that $\{\sigma_t^2\}_{t\in[\![T]\!]}$ are known to the learner, we follow Xu and Tewari (2022), and consider a *strong oracle* that knows $\{Q_t\}_{t\in[T]}$. We later contrast it with a more realistic *weak oracle*.

Strong-oracle learners and their loss Let $\bar{\theta}_t(N)$ be the empirical mean estimator for θ_t using $\mathcal{M}_t$. Then, with high probability (w.h.p.), its loss for estimating θ_0 is upper bounded as

$$\|\bar{\theta}_t(N) - \theta_0\|^2 = \|\bar{\theta}_t(N) - \theta_t + \theta_t - \theta_0\|^2 \lesssim \|\bar{\theta}_t(N) - \theta_t\|^2 + \|\theta_t - \theta_0\|^2 \lesssim \frac{d\sigma_t^2}{N} + Q_t^2 \tag{4}$$

(which is typically tight in high dimensions, *e.g.*, Vershynin (2018, Remark 3.2.5)). Following Xu and Tewari (2022, Theorem 1), we define a strong-oracle learner as one that knows the task that minimizes this upper bound, *i.e.*, $t^* \in \operatorname{argmin}_{t\in[\![T]\!]}\{d\sigma_t^2/N + Q_t^2\}$, and thus knows the optimal trade-off between similarity and difficulty. Such a learner can allocate all its N samples to the model $\mathcal{M}_{t^*}$ (though it does not have to do so), and estimate θ_0 as $\bar{\theta}_{t^*}(N)$. This results in

$$\|\bar{\theta}_{t^*}(N) - \theta_0\|^2 \lesssim \min_{t\in[\![T]\!]} \left\{ \frac{d\sigma_t^2}{N} + Q_t^2 \right\}. \tag{5}$$

We refer to the r.h.s. of (5) as the loss of the *strong-oracle learner*. Xu and Tewari (2022, Theorem 1) showed that this loss is a lower bound on the risk of any CL algorithm (see Appendix E for a detailed comparison with Xu and Tewari (2022)). Since we focus on non-asymptotic analysis, we define for a constant $\kappa > 0$ the *strong oracle set (of source models)* as

$$\mathcal{T}_{\text{s.o.}} \equiv \mathcal{T}_{\text{s.o.}}(\kappa) := \left\{ t \in [\![T]\!] \colon Q_t^2 + \frac{d\sigma_t^2}{N} \le \kappa \cdot \left[\frac{d\sigma_{t^*}^2}{N} + Q_{t^*}^2 \right] \right\}, \tag{6}$$

so that any $t \in \mathcal{T}_{\text{s.o.}}(\kappa)$ is essentially as efficient as t^*. A natural question is whether the strong-oracle set can be identified by a learner that samples from $\{\mathcal{M}_t\}_{t\in[\![T]\!]}$, without prior knowledge of $\{Q_t\}_{t\in[T]}$. If this is possible then (5) can be achieved by a CL algorithm that splits half of the samples for finding $\mathcal{T}_{\text{s.o.}}(\kappa)$, and the rest for sampling from $\mathcal{M}_{\tilde{t}}$ with $\tilde{t} \in \mathcal{T}_{\text{s.o.}}(\kappa)$ and estimating θ_0 as $\bar{\theta}_{\tilde{t}}(N/2)$. We will argue that it is in generally impossible to identify $\mathcal{T}_{\text{s.o.}}(\kappa)$ from samples. We will propose a larger set of models, the *weak-oracle set*, and explain why identifying this set is a more realistic goal for CL algorithms. To motivate the weak oracle, we first address the simple case of $T = 1$.

A single source model and the source elimination lemma When $T = 1$, a CL algorithm only needs to decide whether $d\sigma_1^2/N + Q_1^2 \lesssim d\sigma_0^2/N$ or not. Since $\sigma_1^2 < \sigma_0^2$, this is equivalent to whether $Q_1^2 \lesssim d\sigma_0^2/N$ or not, so that if Q_1^2 is large then θ_0 and θ_1 are too different, and samples from $\mathcal{M}_0$ should be used to estimate θ_0, and otherwise from $\mathcal{M}_1$. To decide this, the learner can estimate $Q_1^2 = \|\theta_0 - \theta_1\|^2$ based on initial estimates $\bar{\theta}_0(N/2)$ and $\bar{\theta}_1(N/2)$, for which it holds w.h.p. that

$$\hat{Q_1^2} := \|\bar{\theta}_0(N/2) - \bar{\theta}_1(N/2)\|^2 \lesssim \frac{d\sigma_1^2}{N} + Q_1^2 + \frac{d\sigma_0^2}{N} \asymp Q_1^2 + \frac{d\sigma_0^2}{N}. \tag{7}$$

The learner then decides that if $\hat{Q_1^2} \gtrsim d\sigma_0^2/N$ then so is $Q_1^2 \gtrsim d\sigma_0^2/N$ and chooses $\mathcal{M}_0$. The source model $\mathcal{M}_1$ is thus *eliminated*, and the final estimate is $\hat{\theta} = \bar{\theta}_0(N/2)$. Otherwise, $\hat{\theta} = \bar{\theta}_1(N/2)$ and the loss is $O(Q_1^2 + d\sigma_1^2/N)$, which is guaranteed to be *not larger* (orderwise) than the loss of the target model estimate, given by $O(d\sigma_0^2/N)$. Optimistically, it *might* also hold that the estimation loss is *much smaller* when using the source model, that is, $Q_1^2 + d\sigma_1^2/N = o(d\sigma_0^2/N)$. However, the learner cannot know this, since it does not have any knowledge of θ_0 at a resolution smaller than $\sqrt{d\sigma_0^2/N}$. The following theorem is a rigorous theoretical guarantee on the loss and risk of this method.

Theorem 1 *Let $\tilde{\theta}_0 = \bar{\theta}_0(N/2)$ (resp. $\tilde{\theta}_1 = \bar{\theta}_1(N/2)$) be an initial estimate of θ_0 using $N/2$ i.i.d. samples from the target model $\mathcal{M}_0$ (resp. source model $\mathcal{M}_1$). Let $\delta \in (0,1)$ be given, and let*

$$\hat{\theta} = \begin{cases} \tilde{\theta}_0, & \|\tilde{\theta}_0 - \tilde{\theta}_1\|^2 \geq 10 g(\delta) \frac{d\sigma_0^2}{N/2} \\ \tilde{\theta}_1, & \text{otherwise} \end{cases} \tag{8}$$

be the final estimate of θ_0. Then, there exists $\nu \in [1/27, 1)$ such with probability at least $1 - \delta$

$$\|\hat{\theta} - \theta_0\|^2 \leq \frac{8g(\delta/2)}{\nu} \cdot \min\left\{ Q_1^2 + \frac{d\sigma_1^2}{N}, \frac{d\sigma_0^2}{N} \right\}. \tag{9}$$

Assume further that $Q_1^2 \leq 4C_\theta^2$ for some constant $C_\theta > 0$. Then, setting in (8)

$$\delta = \delta_* := \left[\frac{\sigma_1^2}{\sigma_0^2 + \sigma_1^2} \vee \frac{1}{4} \vee \frac{d\sigma_0^2}{8C_\theta^2 N} \right]^2$$

results in

$$\mathbb{E}\left[\|\hat{\theta} - \theta_0\|^2 \right] \leq \frac{28c}{\nu} \log\left(\frac{2}{\delta_*} \right) \cdot \min\left\{ Q_1^2 + \frac{d\sigma_1^2}{N}, \frac{d\sigma_0^2}{N} \right\}. \tag{10}$$

Remarkably, the high-probability loss achieved by this method has the same order as the loss achieved by the strong-oracle learner in (5), and their risk nearly match, up to a $\Theta(\log(N/d))$ factor.

The proof of Theorem 1 appears in Appendix F, and is based on a simple, yet crucial, Lemma 10, called the *source elimination lemma*. This lemma establishes that the learner can eliminate the worse of two models, with a constant loss factor (though larger than 2). Its proof is a consequence of the convexity and the approximate triangle-inequality of the squared Euclidean norm loss function. An equivalent rephrase of Lemma 10, which will be used for $T > 1$, is as follows:

Corollary 2 (to the source elimination lemma, Lemma 10) *Suppose an algorithm eliminates model $\mathcal{M}_t$ if $\|\tilde{\theta}_0 - \tilde{\theta}_t\|^2 \geq 10\lambda^2$ for some $\lambda > 0$. Then, there exists a numerical constant $\nu \in [1/27, 1)$ so that if $Q_t^2 \geq \lambda^2/\nu$ then $\mathcal{M}_t$ will be eliminated w.h.p.*

Letting $\lambda_t^2 \asymp d\sigma_t^2/N$, we use this corollary with $\lambda^2 = \lambda_{\max}^2 := \lambda_0^2 \vee \lambda_1^2 \asymp d\sigma_0^2/N$. The proof of the upper bound on the risk (expected loss) in (10) is more technical, and is obtained by controlling of the fourth-moment of the error.

Weak-oracle learners and their loss As is reasonably intuitive, and as proved by Ben-David et al. (2010b), transfer learning from a source to the target is impossible without samples from the target model. In our method, $N/2$ samples (say) were allocated to the target model. However, Q_t^2 cannot be estimated at a resolution smaller than $\Theta(\sqrt{d\sigma_0^2/N})$, even using N samples (see (7)). Thus, even if the learner is informed with the models for which $Q_t^2 \lesssim d\sigma_0^2/N$, it cannot further discriminate between them. Therefore, allocating $\Theta(N)$ samples to the source model with minimal $Q_t^2 + d\sigma_t^2/N$, as can be done using a strong oracle, cannot be guaranteed without prior knowledge of $\{Q_t\}_{t\in[T]}$.

Hence, it is reasonable that a CL algorithm will be able to match the performance of a *weak-oracle learner*. The weak oracle only knows which source models provide more accurate samples than the target model itself, that is, for some constant $\kappa > 0$, the models in the set

$$\mathcal{T}_{\text{w.o.}} \equiv \mathcal{T}_{\text{w.o.}}(\kappa) := \left\{ t \in [T] \colon Q_t^2 \leq \kappa \cdot \frac{d\sigma_0^2}{N} \right\}. \tag{11}$$

The weak-oracle learner will choose one of the source models in $\mathcal{T}_{\text{w.o.}}$ for its final estimate (if $\mathcal{T}_{\text{w.o.}} = \emptyset$ then it will simply use the target model). If $|\mathcal{T}_{\text{w.o.}}| > 1$, then the tightest uniform bound over all problem instances with the same $\mathcal{T}_{\text{w.o.}}$ is obtained by selecting the one with minimal noise variance. Concretely, for any given set $\mathcal{T} \subseteq [T]$, let

$$\bar{t}(\mathcal{T}) := \begin{cases} 0, & \mathcal{T} = \emptyset \\ \operatorname{argmin}^*_{t\in\mathcal{T}} \sigma_t^2, & |\mathcal{T}| \geq 1 \end{cases}, \tag{12}$$

where argmin^* pessimistically chooses the model with the maximal Q_t^2 within the set $\operatorname{argmin}_{t\in\mathcal{T}} \sigma_t^2$ (thus $\bar{t}(\mathcal{T})$ also depends on $\{Q_t^2\}_{t\in\mathcal{T}}$, but we omit this from the notation for brevity). Given $\mathcal{T}_{\text{w.o.}}$, the weak-oracle learner collects N samples from $\bar{t}(\mathcal{T}_{\text{w.o.}})$, and its final estimate is then $\hat{\theta}_{\text{w.o.}} = \overline{\theta}_{\bar{t}(\mathcal{T}_{\text{w.o.}})}(N)$. As in (4), the loss is bounded with w.h.p. as

$$\|\hat{\theta}_{\text{w.o.}} - \theta_0\|^2 \lesssim \frac{d\sigma^2_{\bar{t}(\mathcal{T}_{\text{w.o.}})}}{N} + Q^2_{\bar{t}(\mathcal{T}_{\text{w.o.}})}. \tag{13}$$

Since $\Theta(N)$ samples allow for Q_t to be estimated with resolution $\Theta(\sqrt{d\sigma_0^2/N})$, it is plausible that $\mathcal{T}_{\text{w.o.}}$ can be identified from samples without knowing $\{Q_t^2\}_{t\in[T]}$. We explore this possibility in the next section.

We finally remark that if $\sigma_t^2 \equiv \sigma^2$ for all source models, $t \in [T]$ one may argue that it is better to uniformly "combine the advice" of the models in $\mathcal{T}_{\text{w.o.}}$, and to choose the final estimator to be the average of $\hat{\theta}_{\text{w.o.,avg}} := \frac{1}{|\mathcal{T}_{\text{w.o.}}|} \sum_{t\in\mathcal{T}_{\text{w.o.}}} \overline{\theta}_t(N/|\mathcal{T}_{\text{w.o.}}|)$. Due to the convexity of the squared error, this reduces the loss, however, it does not lead to a reduced upper bound on the loss (since it is possible that θ_t is the same for all $t \in \mathcal{T}_{\text{w.o.}}$).

4. A CL elimination algorithm and risk upper bound

As for $T = 1$, source models can be eliminated by estimation of Q_t^2, but each estimate may be inaccurate, since there are possibly many sources ($T > 1$) and sampling all of them is necessary. For example, if one uniformly allocate $\Theta(N/T)$ samples to each of the source models, the number of samples will be smaller by a factor of T compared to the $T = 1$ case. However, if some source models have already been eliminated, the learner can re-estimate Q_t^2 only for the retained models,

and obtain more accurate estimates that will possibly allow to eliminate additional source models. We thus propose using *multiple* elimination rounds, as depicted in Algorithm 1. In this algorithm the average noise variances in a set of models $\mathcal{T} \subseteq [T]$ is denoted as $\overline{\sigma}^2(\mathcal{T}) := \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}}\sigma_t^2$.

Algorithm 1 A CL algorithm for a multiple source models

input: $(N, d, \{\sigma_t^2\}_{t\in[\![T]\!]}, \bar{r}, \delta)$

(1) Set $\bar{N} \leftarrow {}^{N}\!/_{(\bar{r}+2)}$ and $\bar{\delta} \leftarrow {}^{\delta}\!/_{(T\bar{r}+2)}$

Compute $\tilde{\theta}_0 = \bar{\theta}_0(\bar{N})$ by sampling $\bar{N}$ samples from the target model $\mathcal{M}_0$

Set $\mathcal{T}_0 \leftarrow [T]$ and $T_0 \leftarrow T$

for $r = 1$ to $\bar{r}$ **do**

  Set $\mathcal{T}_r \leftarrow \mathcal{T}_{r-1}$

  for $t \in \mathcal{T}_{r-1}$ **do**

(2) Compute $\tilde{\theta}_t = \bar{\theta}_t(\bar{N}_{t,r})$ by sampling $\bar{N}_{t,r} = \frac{\bar{N}}{T_{r-1}} \cdot \frac{\sigma_t^2}{\overline{\sigma}^2(\mathcal{T}_{r-1})}$ samples from $\mathcal{M}_t$

(3) **if** $\|\tilde{\theta}_0 - \tilde{\theta}_t\|^2 \geq 10g(\bar{\delta}) \cdot \left[\frac{d\overline{\sigma}^2(\mathcal{T}_{r-1})}{\bar{N}/T_{r-1}} \vee \frac{d\sigma_0^2}{\bar{N}}\right]$ **then**

    Set $\mathcal{T}_r \leftarrow \mathcal{T}_r \backslash \{t\}$

  end

  Set $\mathcal{T}_{\text{alg}} \leftarrow \mathcal{T}_r$ and $T_r \leftarrow |\mathcal{T}_r|$

(4) **if** $T_{r-1} \leq \sigma_0^2/\overline{\sigma}^2(\mathcal{T}_{r-1})$ or $T_r = 0$ or $T_r = T_{r-1}$ **then**

    Break out of the for loop

  end

end

(5) Return $\hat{\theta} = \overline{\theta}_{t^*}(\bar{N})$ for some $t^* \in \overline{t}(\mathcal{T}_{\text{alg}})$

Description of Algorithm 1 We assume that $N \gtrsim T$, so that all source models can be explored. The algorithm will use at most $T\overline{r} + 2$ empirical mean estimates (T for each of the $\overline{r}$ rounds, one for the initial target estimate, and one for the final estimate). So, setting $\bar{\delta} := {}^{\delta}\!/_{(T\bar{r}+2)}$ and requiring each estimate to be accurate with probability $1 - \bar{\delta}$, assures that all estimates are accurate with probability $1 - \delta$ (via the union bound). We next assume that this high probability event holds. Let $\overline{r}$ be a total number of elimination rounds and let $\bar{N} = {}^{N}\!/_{(\bar{r}+2)}$ be the number of samples per round (plus an initial and a final round). The learner first uses $\overline{N}$ samples from the target model, to obtain an initial estimate $\tilde{\theta}_0 = \overline{\theta}_0(\overline{N})$, and then starts to eliminate models. Let $\mathcal{T}_r \subseteq [T]$ be the set of retained source models *after* the rth round of elimination, initialized with $\mathcal{T}_0 := [T]$, and denote $T_r := |\mathcal{T}_r|$. Thus, $\mathcal{T}_r \subseteq \mathcal{T}_{r-1}$ and $T_r \leq T_{r-1}$. At each elimination round, $\bar{N}_{t,r}$ samples are allocated to the source models in $\mathcal{T}_r$, which are used to estimate θ_t as $\bar{\theta}_t(\bar{N}_{t,r})$, and then to estimate Q_t^2. Based on the estimated Q_t^2, it is decided (line 3) whether to eliminate the model from $\mathcal{T}_r$, using a condition similar to the source elimination lemma (Corollary 2). Then, it is decided (line 4) if to stop after the rth elimination round, based on several possible criteria: (1) If $T_{r-1} \leq \sigma_0^2/\overline{\sigma}^2(\mathcal{T}_{r-1})$ then this allows the algorithm to eliminate in the current round all models outside $\mathcal{T}_{\text{w.o.}}$. (2) If $\mathcal{T}_r = \emptyset$ then all source models have been eliminated, which implies that using the target model estimator $\overline{\theta}_0(\overline{N})$ is best. (3) If $T_r = T_{r-1}$ then no additional source model is eliminated. If none of these conditions is fulfilled then $\overline{r}$ elimination rounds are performed. Finally, in line 5, one of the models t^* in $\overline{t}(\mathcal{T}_{\text{alg}})$ is chosen (where $t^* = 0$ if $\mathcal{T}_{\text{alg}} = \emptyset$), and the final estimate is $\overline{\theta}_{t^*}(\bar{N})$.

Sample allocation at each round Assume that at round r the $\overline{N}$ samples are allocated to the retained source models in $\mathcal{T}_{r-1}$ as $\{\bar{N}_{t,r}\}_{t\in\mathcal{T}_{r-1}}$ with $\sum_{t\in\mathcal{T}_{r-1}}\bar{N}_{t,r}=\overline{N}$. Then, for any $t\in\mathcal{T}_{r-1}$

$$\|\overline{\theta}_t(N/T)-\theta_t\|^2 \le g(\bar{\delta})\frac{d\sigma_t^2}{\bar{N}_{t,r}} := \lambda_t^2$$

(with probability $1-\bar{\delta}$), and according to the source elimination lemma (Corollary 2), the tth model will be eliminated if $Q_t^2 \ge {}^{(\lambda_t^2\vee\lambda_0^2)}/_{\nu}$. Since the retained source models in $\mathcal{T}_{r-1}$ were not discriminated in round $r-1$, a plausible sample allocation will induce an equal elimination barrier for all $t\in\mathcal{T}_{r-1}$, *i.e.*, $\lambda_{t_1}^2=\lambda_{t_2}^2$ for any $t_1,t_2\in\mathcal{T}_{r-1}$. This results in the sample allocation $\bar{N}_{t,r}={}^{\bar{N}}/_{T_{r-1}}\cdot{}^{\sigma_t^2}/_{\overline{\sigma}^2(\mathcal{T}_{r-1})}$ used in line 2, for which $\lambda_t^2={}^{g(\bar{\delta})d\overline{\sigma}^2(\mathcal{T}_{r-1})}/_{(\bar{N}/T_{r-1})}$. For example, if $\sigma_t^2\equiv\sigma^2$ for all source models $t\in[T]$, then the sample allocation is uniform over T_{r-1} and $\lambda_t^2={}^{g(\bar{\delta})d\sigma^2}/_{(\bar{N}/T_{r-1})}$. The barrier ${}^{(\lambda_t^2\vee\lambda_0^2)}/_{\nu}$ thus decreases at each round, allowing for additional source models to be eliminated. Now, if the number of currently retained models T_{r-1} is *small* enough, specifically, $T_{r-1}\le{}^{\sigma_0^2}/_{\overline{\sigma}^2(\mathcal{T}_{r-1})}$, then the barrier for elimination is $\lambda_t^2\le{}^{g(\bar{\delta})}/_{\nu}\cdot{}^{d\sigma_0^2}/_{\overline{N}}$, which is the same barrier used by the weak oracle (11), with $\kappa\le{}^{g(\bar{\delta})}/_{\nu}\cdot(\overline{r}+2)$. Thus, if $\overline{r}$ is not very large, whenever $T_{r-1}\le{}^{\sigma_0^2}/_{\overline{\sigma}^2(\mathcal{T}_{r-1})}$, the set $\mathcal{T}_{\text{w.o.}}$ is identified by the algorithm w.h.p. In fact, if $T\le{}^{\sigma_0^2}/_{\overline{\sigma}^2(\mathcal{T}_0)}$ then this occurs already in the first round. Multiple rounds are needed for *large* number of source models.

Number of rounds and elimination dynamics The elimination dynamics depend in an intricate way on $\{Q_t^2,\sigma_t^2\}_{t\in[\![T]\!]}$. To allow for simple tracking, we introduce the *elimination curve* $\beta_\delta(\tau)$. Let $\tau\in(0,1)$, and assume that there are τT retained models. Then, $T\beta_\delta(\tau)$ is the number of models that will be eliminated next by the algorithm (w.h.p.). Concretely, letting $\tau_{\min}:=\frac{1}{T}\lceil{}^{\sigma_0^2}/_{\overline{\sigma}^2([T])}\rceil$, and setting the required reliability to $\delta\in(0,1]$, we obtain

$$\beta_\delta(\tau) := \begin{cases} \frac{\left|\left\{t\in[T]\colon Q_t^2\le\frac{g(\bar{\delta})}{\nu}\cdot\frac{d\overline{\sigma}^2([\tau T])}{N/(\tau T)}\right\}\right|}{T}, & \tau\in(\tau_{\min},1) \\ \frac{\left|\left\{t\in[T]\colon Q_t^2\le\frac{g(\bar{\delta})}{\nu}\cdot\frac{d\sigma_0^2}{N}\right\}\right|}{T}, & \tau\in(0,\tau_{\min}) \end{cases}. \tag{14}$$

Hence, at the first round $T_1=T\cdot\beta_\delta(1)$ source models will be retained by the algorithm w.h.p., after the second $T_2=T\cdot\beta_\delta(\beta_\delta(1))$, and so on. After $\bar{r}$ rounds, the number of retained source models is w.h.p. $T_{\bar{r}}=T\cdot\beta_\delta(\beta_\delta(\cdots(1))=T\cdot\beta_\delta^{(\bar{r})}(1)$, where $\beta_\delta^{(r)}$ denotes the rth order functional power of β_δ, and if $T_r\le\tau_{\min}T$ then the algorithm stops since it eliminates all source models outside $\mathcal{T}_{\text{w.o.}}$. That is, if the algorithm satisfies either $T_{r-1}\le{}^{\sigma_0^2}/_{\overline{\sigma}^2(\mathcal{T}_{r-1})}$ or $T_r=0$, then $\mathcal{T}_{\text{alg}}=\mathcal{T}_{\bar{r}}=\mathcal{T}_{\text{w.o.}}({}^{g(\bar{\delta})}/_{\nu})$, and the algorithm identifies the weak oracle set with constant ${}^{g(\bar{\delta})}/_{\nu}$. However, it may happen that the elimination stops becuase $T_r=T_{r-1}$. As we next show, this is easily identified by $\beta_\delta(\tau)$.

Example 1 *Consider $T=10^3$ and assume the normalized values ${}^{g(\bar{\delta})}/_{\nu}\cdot{}^{(d\sigma_t^2)}/_{(N/T)}=1$, for which $\beta_\delta(\tau)=\frac{1}{T}|\{t\in[\tau]\colon Q_t^2\le\tau\}|$ for $\tau\in(\tau_{min},1)$ and $\beta_\delta(\tau)=\beta_\delta(\tau_{min})$ for $\tau\in(0,\tau_{min})$. We assume $\tau_{min}=0.2$. We demonstrate the sensitivity of CL to the values of $\{Q_t^2\}_{t\in[T]}$ by comparing two very similar cases – in the first case $Q_t^2={}^{t^{1/3}}/_{9}$ and in the second case $Q_t^2={}^{t^{1/3}}/_{12}$. In both cases, if σ_0^2 is sufficiently large, then $\mathcal{T}_{w.o.}$ is not empty. Figure 1 (left panel) shows that in the first case (red curve) $\beta_\delta(1)\approx0.72$, $\beta_\delta(0.62)\approx0.28$, and $\beta_\delta(0.242)\approx0.017$, so that after 3 rounds, all models outside $\mathcal{T}_{w.o.}$ are eliminated. By contrast, in the second case (blue curve), $\beta_\delta(1)=1$, and while all source models are better than the target model,* none *of them is eliminated. Thus, the risk of the weak-oracle learner is* not *achieved by Algorithm 1.*

Example 1 reveals an intricate property: The distances $\{Q_t\}_{t\in[T]}$ are smaller in the second case compared to the first case, but Algorithm 1 has a lower loss guarantee in the first case – thus the loss is *not* monotonic with the distances. This appears to be an inherent property of the statistical CL problem: Models with low similarity (large Q_t) can be easily eliminated. Nonetheless, generalizing the observation of Example 1, if $\beta_\delta(\tau)$ is bounded away from the identity line (and thus does not have a fixed point), then Algorithm 1 will identify the weak-oracle set using $r = \Theta(\log T)$ rounds, as follows (the proof is obvious from the discussion above):

Proposition 3 *Suppose $\beta_\delta(\tau) \le \overline{\beta}\tau$ for some $\overline{\beta} \in (0,1)$ and all $\tau \in (\tau_{min}, 1]$. Then, $\mathcal{T}_{alg} = \mathcal{T}_{w.o.}(g(\bar{\delta})/\nu)$ if*

$$\bar{r} = \frac{\log \tau_{min}}{\log \overline{\beta}} = \frac{\log\left(T \cdot \overline{\sigma}^2([T])/\sigma_0^2\right)}{\log(1/\overline{\beta})}.$$

The final estimate Similarly to the strong/weak-oracle learners, we choose the final estimate to be based on $\overline{N}$ samples from $t_{\text{alg}} := \bar{t}(\mathcal{T}_{\text{alg}})$, which favors the model with minimal noise variance, see (12). The final estimate is thus $\hat{\theta}_{\text{alg}} = \overline{\theta}_{t_{\text{alg}}}(\overline{N})$, and the resulting loss is bounded w.h.p., as in (4), by $\|\hat{\theta}_{\text{alg}} - \theta_0\|^2 \lesssim d\sigma^2_{t_{\text{alg}}}/N + Q^2_{t_{\text{alg}}}$. The following theorem provides a rigorous high probability guarantee on the loss and the risk of Algorithm 1. The proof appears in Appendix G and is a direct consequence of the source elimination lemma.

Theorem 4 *Assume w.l.o.g. that*[1] $0 = Q_0^2 \le Q_1^2 \le \cdots \le Q_T^2$. *Let $\beta_\delta(\tau)$ be as in (14) and let $\mathcal{T}_{alg} = [T_{\bar{r}}]$ with $T_{\bar{r}} = T\beta_\delta(\beta_\delta(\cdots(1))) = T\beta_\delta^{(\bar{r})}(1)$. Let $\bar{\delta} = \delta/(T\bar{r}+2)$ for $\delta \in (0,1)$ and let $\hat{\theta}$ be the output of Algorithm 1. Then, with probability at least $1-\delta$*

$$\|\hat{\theta} - \theta_0\|^2 \le 2(\bar{r}+2) \cdot \left(g(\bar{\delta}) \vee 1\right) \cdot \left[\frac{d\sigma^2_{\bar{t}(\mathcal{T}_{alg})}}{N} + Q^2_{\bar{t}(\mathcal{T}_{alg})}\right]. \tag{15}$$

Assume further that $\max_{t\in[T]} Q_t^2 \le 4C_\theta^2$ for some constant $C_\theta > 0$. Then, setting

$$\delta = \delta_* := \left[\frac{1}{2(T+1)} \cdot \left[\min_{t,t'\in[\![T]\!]} \frac{\sigma_t^2}{\sigma_{t'}^2} \vee \min_{t\in[\![T]\!]} \frac{d\sigma_t^2}{8NC_\theta^2}\right]\right]^2$$

as input to Algorithm 1 results in

$$\mathbb{E}\left[\|\hat{\theta} - \theta_0\|^2\right] \le 8c(\bar{r}+2) \cdot \log\left(\frac{(T\bar{r}+2)}{\delta_*}\right) \cdot \left[\frac{d\sigma^2_{\bar{t}(\mathcal{T}_{alg})}}{N} + Q^2_{\bar{t}(\mathcal{T}_{alg})}\right]. \tag{16}$$

For mean estimation $g(\bar{\delta}) = \Theta(\log(T\bar{r}/\delta))$, and so if $\mathcal{T}_{\text{alg}} = \mathcal{T}_{\text{w.o.}}(g(\bar{\delta})/\nu)$ then the loss/risk of the weak-oracle learner is achieved using $\bar{r} = \Theta(\log T)$ rounds, up to a mild $\Theta(\log^2(TN/d))$ factor.

Extension to unknown variances/covariance-matrices In Appendix H, we extend Algorithm 1 and Theorem 4 to the case the variances $\{\sigma_t^2\}_{t\in[\![T]\!]}$ are unknown (with $\overline{\Sigma}_t = I_d$ for all $t \in [\![T]\!]$), and then to unknown covariance matrices. The extension is based on a preliminary step of variance/covariance estimation, and plug-in of the estimated values instead of the exact ones in Algorithm 1. Assuming that the mild condition $dN = \Omega(T\log T)$ holds for unknown variances and $dN \gtrsim d\log d \cdot T\log T$ holds for unknown covariance matrices, the loss/risk of this extended Algorithm 1 is roughly the same as in Theorem 4 in both cases.

1. The learner does not know $\{Q_t^2\}_{t\in[T]}$, and does not know that this order holds.

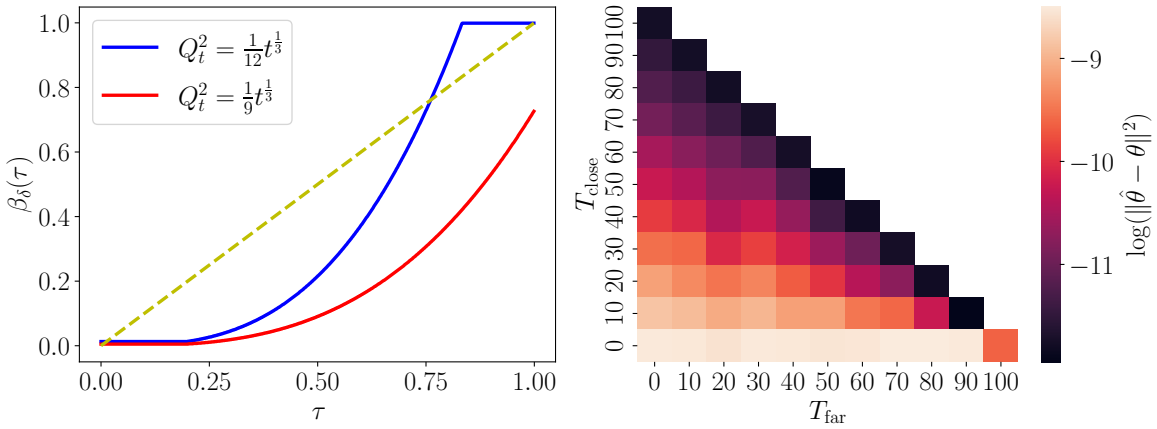

Figure 1: Left: The elimination curves $\beta_\delta(\tau)$ for Example 1 (the identity line in dashed yellow). Right: The loss $\|\hat{\theta} - \theta_0\|^2$ on a log-scale. Parameters are $N = 10^5$, $d = 2$, $\sigma^2 = 0.1$, $\sigma_0^2 = 1$, $\tilde{Q}^2_{\text{close}} = 0$, $\tilde{Q}^2_{\text{medium}} = 10$, $\tilde{Q}^2_{\text{far}} = 2 \cdot 10^4$ where $\tilde{Q}_t^2 := Q_t^2/(d\sigma_0^2/N)$ is the normalized distance.

Numerical experiments We empirically validated Algorithm 1 in simulations, and provide the results in Appendix I. Here, we highlight one experiment in which the sources are partitioned to three types, according to the distance of their parameter from the target parameter, designated as "close", "medium", and "far". Figure 1 (right panel) displays the loss of Algorithm 1 for different mixtures of types $(T_{\text{far}}, T_{\text{close}})$ and $T_{\text{med}} = T - T_{\text{close}} - T_{\text{far}}$, for $T = 100$ source models, averaged over 200 experiments. One can track the loss over various interesting paths. For example, since Algorithm 1 can eliminate the "far" models, the loss is consistently low when traversing the straight line from $(0, 100)$ to $(100, 0)$, as long as there is at least one close source model. The loss increases when traversing the straight line from $(50, 50)$ to $(0, 0)$ due to the additional "medium" parameters, which are not eliminated.

5. Minimax risk lower bounds

We next derive lower bounds on the minimax risk, which can be compared with the risk of the weak-oracle learner (13), and with Algorithm 1. Since the learner can typically learn the noise statistics (see Appendix H), we assume that $\{\overline{\Sigma}_t = I_d\}_{t\in[\![T]\!]}$ and $\{\sigma_t^2\}_{t\in[\![T]\!]}$ are known to the learner. In this case, we may assume w.l.o.g. that $\sigma_0^2 \geq \sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_T^2$, and let $\mathbf{\Gamma} := (\sigma_0^2, \sigma_1^2, \ldots, \sigma_T^2)$. We take the noise variances as fixed, thus specify a problem-instance by $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_T)$, and let $\mathbf{\Psi} \subset (\mathbb{R}^d)^{T+1}$ be a set of such parameters. The minimax risk is then

$$L_{N,d}(\mathbf{\Psi}, \mathbf{\Gamma}) := \inf_{\mathcal{A}} \sup_{\boldsymbol{\theta}\in\mathbf{\Psi}} \mathbb{E}_{\boldsymbol{\phi},\mathcal{A}} \left[\left\| \hat{\theta}\left((A_i, S_i)_{i\in[N]}\right) - \theta_0 \right\|^2 \right]. \tag{17}$$

To lower bound the minimax risk, we follow the standard method of reducing the learning problem to an hypothesis testing problem between K hypotheses $\mathbf{\Theta}_{\text{test}} := \{\theta^{(j)}\}_{j\in[K]} \subset \mathbf{\Psi}$ where $\theta^{(j)} = (\theta_0^{(j)}, \theta_1^{(j)}, \ldots, \theta_T^{(j)})$, and then bounding the error probability in the latter (Yang and Barron, 1999)

(Wainwright, 2019, Chapter 15). However, this reduction is different from the standard one in two aspects, which we next discuss, before deriving two minimax lower bounds.

A general reduction to hypothesis testing The algorithm $\mathcal{A}$ in (17) collects samples adaptively, and so its N samples are *not* i.i.d. Thus, in principle, the reduction to hypothesis testing should be made to a similarly adaptive tester. However, lower bounds for adaptive testers are not as widely available, compared to Le-Cam's and Fano's bounds for i.i.d. samples. There are a few strategies to circumvent this. First, we may naively assume that each of the $T+1$ models is sampled N times, and so there is no need for adaptive allocation. This leads to non-trivial lower bounds, but does not capture the correct dependence on T. Second, we may choose the set $\boldsymbol{\Theta}_{\text{test}}$ in a way that an optimal learner will surely only sample from one of the models. Since the noise is Gaussian, this is the case, *e.g.*, if the $\theta_{t_1}^{(j)} \propto \theta_{t_2}^{(j)}$ for some $t_1, t_2 \in [T]$ and all $j \in [K]$. Note that making an analogous claim for estimation, rather than testing, is non-trivial. We discuss this in detail in Appendix J.1.

The choice of the localization set and associated challenges The *localization set* $\boldsymbol{\Psi}$ localizes the lower bound on a set of problem instances. For the most informative bound, these instances should have equal difficulty, where here, as in (Mousavi Kalan et al., 2020; Xu and Tewari, 2022), the difficulty is determined by $\{Q_t^2\}_{t\in[\![T]\!]}$ (beyond$\{\sigma_t^2\}_{t\in[\![T]\!]}$, which are assumed fixed). In fact, for our problem, some localization is essential, since if $\boldsymbol{\Psi} = (\mathbb{R}^d)^{T+1}$ then the worst case is when $Q_t^2 \gg {}^{d\sigma_0^2}/_N$, and all source models are useless for the target task; this trivializes the bound. Xu and Tewari (2022) proposed a set we call a *semi-local set*, given by

$$\boldsymbol{\Psi} \equiv \boldsymbol{\Psi}_{\leq}(\boldsymbol{q}) := \bigcup_{\zeta\in\mathfrak{S}_T} \left\{ \boldsymbol{\theta} \in \boldsymbol{\Theta} \colon Q_t^2 := \|\theta_t - \theta_0\|^2 \leq q_{\zeta(t)}^2, \text{ for all } t \in [T] \right\}, \tag{18}$$

where $\mathfrak{S}_T$ is the set of all permutations of $[T]$ and $\boldsymbol{q} = \{q_t\}_{t\in[T]}$ satisfies, w.l.o.g., $0 < q_1 \leq \cdots \leq q_T$. So, the learner only knows an *unordered multiset* of upper bounds on $\{Q_t\}_{t\in[T]}$ (where $Q_t := \|\theta_0 - \theta_t\|$). However, the risks of the weak-oracle learner and Algorithm 1 are *not necessarily monotonic* with Q_t, and so the worst-case may achieved in the interior of this set, *i.e.*, for $Q_t < q_{\zeta(t)}$. Hence, the set $\boldsymbol{\Psi}_{\leq}(\boldsymbol{q})$ contains problem instances of *heterogeneous* difficulty. A reasonable solution is to replace the $\leq$ in (18) with $=$, thus obtaining $\boldsymbol{\Psi}_{=}(\boldsymbol{q})$, a *fully localized set*. However, restricting θ to $\boldsymbol{\Psi}_{=}(\boldsymbol{q})$ may reveal too much information to the learner, which allows unrealistically low risk, and thus to a loose bound. Indeed, assume that $\sigma_t^2 = 0$ for all $t \in [T]$, that T is odd, that ${}^{(T+1)}/_2$ of the source parameters are identical to the target parameter, and that the other ${}^{(T-1)}/_2$ source parameters are at a larger, equal distance, which is still within the target noise resolution. That is $q_0^2 = q_1^2 = \cdots q_{{}^{(T+1)}/_2}^2 = 0$ and $q_{{}^{(T+3)}/_2}^2 = \cdots = q_T^2 := \overline{q}^2 \lesssim {}^{d\sigma_0^2}/_N$. The loss of the weak-oracle learner in (13) is $\|\hat{\theta}_{\text{w.o.}} - \theta_0\|^2 \lesssim \overline{q}^2$, and is achieved by Algorithm 1. However, just knowing the *multiset* $\{q_t^2\}_{t\in[T]}$, but not the permutation ζ, the learner can collect one sample from each source model, and estimate θ_0 as the observed sample that is common to ${}^{(T+1)}/_2$ models (or more). Thus, no lower bound can be better than 0. In Appendix J.2 we provide two more examples of exceeding the weak-oracle learner risk, either by a dimensionality reduction to $d = 1$ (when $T = 2$), or by a reduction to two possibilities (when $T = 1$). We also explain why this occurs even for noise with a strictly positive variance. The problems of both $\boldsymbol{\Psi}_{\leq}(\boldsymbol{q})$ (lack of monotonicity) and $\boldsymbol{\Psi}_{=}(\boldsymbol{q})$ (possibly trivial or loose lower bound) reveal the delicate nature of the statistical CL problem.

Semi-local minimax risk lower bound in low dimensions For the *semi-local* set $\boldsymbol{\Psi}_{\leq}(\boldsymbol{q})$ we derive a lower bound based on a one-dimensional construction, which, as such, does not capture the

correct dependence on d. For brevity, we let $t_{\text{w.o.}} := \bar{t}(\mathcal{T}_{\text{w.o.}}(\kappa))$ where $\mathcal{T}_{\text{w.o.}}(\kappa)$ is defined for the identity permutation, *i.e.*, $Q_t^2 = \|\theta_t - \theta_0\|^2 = q_t^2$. We also let $q_{\text{w.o.}} := q_{t_{\text{w.o.}}}$ as well as $\sigma^2_{\text{w.o.}} := \sigma^2_{t_{\text{w.o.}}}$, and refer to $t_{\text{w.o.}}$ as the index of the *weak-oracle model*. We let the index of the task with median distance q_t among the "close" models $t \in [t_{\text{w.o.}} - 1]$ be $t_{\text{med}} := \lceil (t_{\text{w.o.}}+1)/2 \rceil$, and let $q_{\text{med}} := q_{t_{\text{med.}}}$.

Theorem 5 *Suppose w.l.o.g. that $q_0^2 = 0 \le q_1^2 \le \cdots \le q_T^2$ and $\sigma_0^2 \ge \sigma_1^2 \ge \cdots \ge \sigma_T^2$. Then*

$$L_{N,d}\left(\mathbf{\Psi}_{\le}(\boldsymbol{q}), \mathbf{\Gamma}\right) \ge \frac{1}{720} \cdot \left(\frac{\sigma_T^2}{N} + q_{med}^2 \right),$$

with $\kappa = 1/(4d)$ in $\mathcal{T}_{w.o.}(\kappa)$ of (11).

The lower bound is tighter than the strong-oracle learner risk of Xu and Tewari (2022, Theorem 1) and also tighter than its improved version (Xu and Tewari, 2022, Theorem 2) (see Appendix J.3 for a detailed comparison). However, in general it is lower than the weak-oracle learner risk since it uses q^2_{med} instead of the larger $q^2_{\text{w.o.}}$, and σ_T^2 instead of $\sigma^2_{t_{\text{med}}}$. Interestingly, the gap between them depends on $\{q_t^2\}_{t\in[T]}$ in an intricate way, just like the intricate way that the upper bound on the risk of Algorithm 1 depends on the distances (i.e., the elimination curve $\beta_\delta(\tau)$). In Appendix J.3 we discuss this in detail, and provide a proof based on Le-Cam's two-point method. We prove it for $T = 2$, and then reduce $T > 2$ to this case. Nonetheless, as the proof shows, the bound matches the weak-oracle learner risk for $T \le 2$.

Single localized-source minimax risk lower bound in high dimensions We propose

$$\mathbf{\Psi}_*(q) := \bigcup_{\zeta \in \mathfrak{S}_T} \left\{ \boldsymbol{\theta} \in \mathbf{\Theta} \colon \exists t \in [T] \text{ such that } Q_t^2 = \|\theta_t - \theta_0\|^2 = q^2 \right\},$$

which is *a single localized-source set*, and offers a reasonable balance between localization of the instance and the information revealed to the learner. Given that $\boldsymbol{\theta} \in \mathbf{\Psi}_*(q)$, the learner only knows that there exists one source model whose parameter is at distance q from the target parameter.

Theorem 6 *Assume that $d \ge 3$, that $q_0^2 = 0 \le q_1^2 \le \cdots \le q_T^2$ and that $\sigma_0^2 \ge \sigma_1^2 \ge \cdots \ge \sigma_T^2$. Set $\kappa = 1$ in the definition of $\mathcal{T}_{w.o.}(\kappa)$ in (11). There exists $c(d) \ge e^{-5}$ so that*

$$L_{N,d}(\mathbf{\Psi}_*(q_{w.o.}), \mathbf{\Gamma}) \ge \begin{cases} L_{N,d}(\mathbf{\Psi}_=(\boldsymbol{q}), \mathbf{\Gamma}) \ge \frac{c^2(d)}{2} \cdot \frac{d\sigma^2_{w.o.}}{N}, & d\sigma^2_{w.o.}/N \ge q^2_{w.o.} \\ \frac{c^2(d)}{8} \cdot q^2_{w.o.}, & d\sigma^2_{w.o.}/N < q^2_{w.o.} \end{cases}.$$

Thus, if the noise variance $d\sigma^2_{\text{w.o.}}/N$ of the weak oracle model dominates $q^2_{\text{w.o.}}$ the minimax risk lower bound holds even for the localized set $\mathbf{\Psi}_=(\boldsymbol{q})$. Otherwise, it holds for the single localized-source set $\mathbf{\Psi}_*(q_{\text{w.o.}})$. The proof is in Appendix J.4 and utilizes Fano's method to obtain a bound that correctly depends on d. The construction of $\mathbf{\Theta}_{\text{test}}$ is designed so that the optimal policy π is non-adaptive.

6. Future research

While the elimination approach is natural for CL, it would interesting to investigate other algorithmic approaches to compete with the weak-oracle learner. For lower bounds, it is interesting to develop and directly utilize bounds on adaptive hypothesis testing, or to utilize Bayesian methods. Additional interesting settings are the CL problem under a low-dimensional common structure of the parameters (as considered by Du et al. (2021); Xu and Tewari (2022)), or other similarity measures between parameters, *e.g.*, the sparsity level $\|\theta_0 - \theta_t\|_0$. Finally, it would be interesting to develop algorithms and extend the analysis to non-parametric and distribution-free settings.

Acknowledgments

The work of RM was partially supported by the Israel Science Foundation (ISF) grant 1693/22, by the Skillman chair in biomedical sciences and by the Ollendorf Center of the Viterbi Faculty of Electrical and Computer Engineering at the Technion. The work of OC and NW was supported by the Israel Science Foundation (ISF), grant 1782/22.

References

Eugene L Allgower and Kurt Georg. *Numerical continuation methods: An introduction*, volume 13. Springer Science & Business Media, 2012.

Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Multi-task feature learning. *Advances in neural information processing systems*, 19, 2006.

Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *COLT*, pages 41–53, 2010.

Jonathan Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12: 149–198, 2000.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010a.

Shai Ben-David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 129–136. JMLR Workshop and Conference Proceedings, 2010b.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Rich Caruana. Multitask learning. *Machine learning*, 28:41–75, 1997.

Elisabetta Cornacchia and Elchanan Mossel. A mathematical model for curriculum learning. *arXiv preprint arXiv:2301.13833*, 2023.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, Hoboken, NJ, USA, 2006.

Michael Crawshaw. Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*, 2020.

M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385, 2021.

Simon S Du, Wei Hu, Sham M Kakade, Jason D Lee, and Qi Lei. Few-shot learning via learning the representation, provably. *arXiv preprint arXiv:2002.09434*, 2020.

Simon Shaolei Du, Wei Hu, Sham M. Kakade, Jason D. Lee, and Qi Lei. Few-shot learning via learning the representation, provably. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=pW2Q2xLwIMD.

Eyal Even-Dar, Shie Mannor, Yishay Mansour, and Sridhar Mahadevan. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(6), 2006.

Itay Evron, Edward Moroshko, Gon Buzaglo, Maroun Khriesh, Badea Marjieh, Nathan Srebro, and Daniel Soudry. Continual learning in linear classification on separable data. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 9440–9484. PMLR, 23–29 Jul 2023.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027. PMLR, 2016.

Pascal Germain, Amaury Habrard, François Laviolette, and Emilie Morvant. A PAC-Bayesian approach for domain adaptation with specialization to linear classifiers. In *International conference on machine learning*, pages 738–746. PMLR, 2013.

Daniel Goldfarb and Paul Hand. Analysis of catastrophic forgetting for random orthogonal transformation tasks in the overparameterized regime. In *International Conference on Artificial Intelligence and Statistics*, pages 2975–2993. PMLR, 2023.

Daniel Goldfarb, Itay Evron, Nir Weinberger, Daniel Soudry, and PAul HAnd. The joint effect of task similarity and overparameterization on catastrophic forgetting – an analytical model. In *The Twelfth International Conference on Learning Representations*, 2024.

Alex Graves, Marc G Bellemare, Jacob Menick, Remi Munos, and Koray Kavukcuoglu. Automated curriculum learning for neural networks. In *international conference on machine learning*, pages 1311–1320. Pmlr, 2017.

Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks. In *International conference on machine learning*, pages 2535–2544. PMLR, 2019.

Steve Hanneke. A statistical theory of active learning. *Foundations and Trends in Machine Learning*, pages 1–212, 2013.

Steve Hanneke and Samory Kpotufe. On the value of target data in transfer learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Steve Hanneke et al. Theory of disagreement-based active learning. *Foundations and Trends® in Machine Learning*, 7(2-3):131–309, 2014.

Hideitsu Hino. Active learning: Problem settings and recent developments. *arXiv preprint arXiv:2012.04225*, 2020.

Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9): 5149–5169, 2021.

Kevin Jamieson and Robert Nowak. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *2014 48th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2014.

Vladimir Koltchinskii and Shahar Mendelson. Bounding the smallest singular value of a random matrix without concentration. *International Mathematics Research Notices*, 2015(23):12991–13008, 2015.

Mikhail Konobeev, Ilja Kuzborskij, and Csaba Szepesvári. A distribution-dependent analysis of meta learning. In *International Conference on Machine Learning*, pages 5697–5706. PMLR, 2021.

Samory Kpotufe and Guillaume Martinet. Marginal singularity and the benefits of labels in covariate-shift. *The Annals of Statistics*, 49(6):3299–3323, 2021.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of statistics*, pages 1302–1338, 2000.

Haoran Li, Jingfeng Wu, and Vladimir Braverman. Fixed design analysis of regularization-based continual learning. *arXiv preprint arXiv:2303.10263*, 2023.

Sen Lin, Peizhong Ju, Yingbin Liang, and Ness Shroff. Theory on forgetting and generalization of continual learning. *arXiv preprint arXiv:2302.05836*, 2023.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016.

Mehryar Mohri and Andres Muñoz Medina. New analysis and algorithm for learning with drifting distributions. In *Algorithmic Learning Theory: 23rd International Conference, ALT 2012, Lyon, France, October 29-31, 2012. Proceedings 23*, pages 124–138. Springer, 2012.

Mohammadreza Mousavi Kalan, Zalan Fabian, Salman Avestimehr, and Mahdi Soltanolkotabi. Minimax lower bounds for transfer learning with linear and one-hidden layer neural networks. *Advances in Neural Information Processing Systems*, 33:1959–1969, 2020.

Sanmit Narvekar, Bei Peng, Matteo Leonetti, Jivko Sinapov, Matthew E Taylor, and Peter Stone. Curriculum learning for reinforcement learning domains: A framework and survey. *The Journal of Machine Learning Research*, 21(1):7382–7431, 2020.

Rémy Portelas, Cédric Colas, Lilian Weng, Katja Hofmann, and Pierre-Yves Oudeyer. Automatic curriculum learning for deep RL: A short survey. *arXiv preprint arXiv:2003.04664*, 2020.

Philippe Rigollet and Jan-Christian Hütter. High dimensional statistics. *Lecture notes for course 18S997*, 2019.

Philippe Rigollet and Alexander Rakhlin. Mathematical statistics: A non-asymptotic approach. *Lecture notes for IDS.160*, 2020.

Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

Daniel Russo. Simple Bayesian algorithms for best arm identification. In *Conference on Learning Theory*, pages 1417–1418. PMLR, 2016.

Luca Saglietti, Stefano Mannelli, and Andrew Saxe. An analytical theory of curriculum learning in teacher-student networks. *Advances in Neural Information Processing Systems*, 35:21113–21127, 2022.

Mark J. Schervish and Morris H. DeGroot. *Probability and statistics*, volume 563. Pearson Education London, UK:, 2014.

Aleksandrs Slivkins. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

Marta Soare, Alessandro Lazaric, and Rémi Munos. Best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 27, 2014.

Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum learning: A survey. *International Journal of Computer Vision*, 130(6):1526–1565, 2022.

Nilesh Tripuraneni, Michael Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. *Advances in neural information processing systems*, 33:7852–7862, 2020.

Nilesh Tripuraneni, Chi Jin, and Michael Jordan. Provable meta-learning of linear representations. In *International Conference on Machine Learning*, pages 10434–10443. PMLR, 2021.

Gido M Van de Ven and Andreas S Tolias. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*, 2019.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Ricardo Vilalta and Youssef Drissi. A perspective view and survey of meta-learning. *Artificial intelligence review*, 18(2):77–95, 2002.

Martin J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4555–4576, 2021.

Daphna Weinshall and Dan Amir. Theory of curriculum learning, with convex loss functions. *Journal of Machine Learning Research*, 21(222):1–19, 2020.

Daphna Weinshall, Gad Cohen, and Dan Amir. Curriculum learning by transfer learning: Theory and experiments with deep networks. In *International Conference on Machine Learning*, pages 5238–5246. PMLR, 2018.

Ziping Xu. *On the Benefits of Multitask Learning: A Perspective Based on Task Diversity*. PhD thesis, 2023.

Ziping Xu and Ambuj Tewari. On the statistical benefits of curriculum learning. In *International Conference on Machine Learning*, pages 24663–24682. PMLR, 2022.

Liu Yang, Steve Hanneke, and Jaime Carbonell. A theory of transfer learning with applications to active learning. *Machine learning*, 90:161–189, 2013.

Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pages 1564–1599, 1999.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International conference on machine learning*, pages 3987–3995. PMLR, 2017.

Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609, 2021.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning, 2020.

Organization of the appendix In Appendix A, we review related work in detail. In Appendix B, we set up notation conventions. In Appendix C we derive, for completeness, a high probability bound on the empirical mean estimator. In Appendix D, we formulate the parametric linear regression setting, considered by Xu and Tewari (2022) instead of our similar parametric setting of Gaussian mean estimation, and derive, for completeness, a high probability bound on the projected LS estimator. In Appendix E, we make a detailed comparison with the results of Xu and Tewari (2022). In Appendix F, we provide the proof for Theorem 1, which bounds the loss/risk for our elimination method in the case of a single source model ($T = 1$), and to this end we introduce the *source elimination lemma* (Lemma 10). In Appendix G, we provide the proof of Theorem 4 that establishes theoretical guarantees on the loss/risk of Algorithm 1, used for multiple source models ($T > 1$). In Appendix H, we generalize Algorithm 1 and Theorem 4 to the case that the noise variances or covariance matrices are unknown, and show that the loss/risk only slightly increases, assuming mild conditions on the problem parameters hold. In Appendix I, we empirically validate Algorithm 1 on several simulated settings, and show that its risk matches the theoretical predictions. In Appendix J, we discuss in detail the two aspects of the minimax lower bound presented in the main text, and then prove the two minimax lower bounds (Theorem 5 and Theorem 6).

Appendix A. Related work

Bengio et al. (2009) explained the optimization benefits of CL by attributing to them regularization capabilities that reduce the generalization error. One way to achieve this is employing CL as a *continuation method* (Allgower and Georg, 2012), in which the difficulty of a task is defined through the smoothness of its objective function. A continuum of risk functions $\{L_\lambda(\theta)\}_{\lambda\in[0,1]}$ is defined, in which $L_0(\theta)$ is easiest to optimize over θ (*e.g.*, it is very smooth or convex), and $L_1(\theta)$ is the target risk function, which is difficult to optimize. The learner first optimizes $L_0(\theta)$ and then increases λ while continuously tracking the minimizer of the risk. Alternatively, improved generalization error can be achieved by placing weights on the training samples. At first, larger weights assigned to easy samples (*e.g.*, far from the margin, or without irrelevant inputs), and then gradually more weight is assigned to more difficult ones, until all samples are assigned a uniform weight. The experiments by Bengio et al. (2009) were the first to suggest that models trained with a simple CL strategy achieve better convergence rate and accuracy.

Following Bengio et al. (2009), the CL approach has been experimented and utilized in various applications, such as computer vision, natural language processing, computer vision, speech processing, medical imaging, reinforcement learning and others (Portelas et al., 2020; Narvekar et al., 2020). However, despite its intuitive plausibility, CL is not widely used in machine learning, and the empirical results occasionally exhibit only moderate effectiveness, if at all. In an attempt to map the advantages of CL, a few recent papers (Wang et al., 2021; Soviany et al., 2022) have comprehensively surveyed CL methods and their applications. Specifically, Soviany et al. (2022) classified CL methods to those applied on the data, the model, the task and the performance measure. A taxonomy of CL methods was created, based on seven categories representing aspects such as determinations of the CL pace, diversity in the training samples, the complexity of the model, and architectures such as teacher-student. That being said, CL was successfully recently used in foundation models (Brown et al., 2020), and this is a strong timely motivation to theoretically explore its benefits.

The statistical benefits of CL were considered by Xu and Tewari (2022), which introduced the parametric learning problem that we adopt here, showed that the risk of the strong-oracle learner is

a lower bound on the risk of a CL algorithm (Xu and Tewari, 2022, Theorem 1), and then offered an improved lower bound (Xu and Tewari, 2022, Theorem 2). Then, a single-round elimination CL algorithm was proposed. Under a restricted setting, an upper bound on the risk was derived (Xu and Tewari, 2022, Theorem 3), which was revised by Xu (2023, Theorem 3.3). We compare the results here with the results of Xu and Tewari (2022) both in the paper, and in Appendix E in detail.

The statistical benefits of CL rely on a quantitative measure of the similarity between the target task and the source task. Such measures were extensively addressed in *multitask learning and meta-learning* (*learning-to-learn*) (Vilalta and Drissi, 2002; Baxter, 2000; Argyriou et al., 2006; Maurer et al., 2016; Du et al., 2020; Tripuraneni et al., 2020, 2021; Hospedales et al., 2021; **?**), where similarity is obtained, *e.g.*, via a common low-dimensional representation. In these methods, improved performance on a target task is achieved by sampling multiple source tasks. Similarity between tasks is also considered in *transfer learning* (Ben-David et al., 2010a; Mohri and Muñoz Medina, 2012; Yang et al., 2013; Germain et al., 2013; Hanneke and Kpotufe, 2019; Zhuang et al., 2020; Kpotufe and Martinet, 2021), and *domain adaptation* (Mansour et al., 2009; Ben-David et al., 2010b), in which the goal is to transfer knowledge between tasks, and the resulting risk is measured by the task similarity. Finally, in *continual learning* (Allgower and Georg, 2012; Zenke et al., 2017; Van de Ven and Tolias, 2019; De Lange et al., 2021), tasks are presented to the learner in succession, and it should adapt to the new task while avoiding *forgetting* of previous tasks. CL refines these methods, as it allows optimizing the order in which source tasks are presented to the learner (the setting of a common low-dimensional representation was also explored by Xu and Tewari (2022)).

The ability of the learner to use its past experience in selecting an action – selecting the next model to sample from in the CL setting – directly relates the CL problem to online decision-making problems, such as the *multi-armed bandit* (MAB) problem (Bubeck et al., 2012; Slivkins, 2019; Lattimore and Szepesvári, 2020). However, in our CL setting, the outcome of the action is the sample of the chosen model, and unlike the MAB problem, this does not define an explicit or an immediate reward that is linearly accumulated. Also, unlike typical MAB problems, it is not obvious *a priori* that the optimal strategy is to sample from just a single model, even if the problem-instance is known to the learner (via an oracle). So the CL problem is also different from a best-arm identification problem (Even-Dar et al., 2006; Audibert et al., 2010; Soare et al., 2014; Jamieson and Nowak, 2014; Russo, 2016; Garivier and Kaufmann, 2016). Nonetheless, Graves et al. (2017) proposed MAB algorithms for mini-batch scheduling of training of deep neural networks (DNNs). They discussed reward signals driven by accuracy prediction and network complexity, and discussed CL for multitask representation learning. Xu and Tewari (2022) proposed prediction-gain driven task schedulers that are inspired by MAB approaches. In a similar spirit, the CL problem is also related to *active learning* (*e.g.*, (Hanneke, 2013; Hanneke et al., 2014; Hino, 2020)), say, in a supervised setting, in which the learner can choose the next input to observe a label from based on past observations. From the perspective of the CL setting, each input defines a task, and thus a continuum of tasks is available. In active learning, however, the goal is to minimize the average loss over all inputs (tasks), and typically there is no specific target task.

Direct theoretical investigations of CL methods are rather scarce. Beyond Xu and Tewari (2022), they include Weinshall et al. (2018); Hacohen and Weinshall (2019); Weinshall and Amir (2020) and the recent work by Cornacchia and Mossel (2023); Saglietti et al. (2022). In the sequence of papers (Weinshall et al., 2018; Hacohen and Weinshall, 2019; Weinshall and Amir, 2020), the difficulty of a training point is measured by the loss of the optimal hypothesis, and various relations between difficulty and convergence rates of a stochastic gradient descent (SGD) training algorithm

were derived, both for linear regression and binary classification. For example, it was proved that the convergence rate decreases with difficulty, and for a step size small enough and a fixed loss, the convergence rate increases with the loss. In practice, ranking the difficulty of training points is hardly available in advance, and so Weinshall et al. (2018); Hacohen and Weinshall (2019); Weinshall and Amir (2020) have proposed a teacher-student architecture, in which a teacher network transfers the training point difficulty to the student network. A training algorithm that is a variation of SGD was proposed, in which at the first steps the input training point is chosen randomly with bias towards easier examples, and this bias decays with the training iterations. They empirically showed in a DNN setup that CL increase the rate of convergence at the beginning of training, and improves generalization when either the task is difficult, the network is small, or when strong regularization is enforced.

Cornacchia and Mossel (2023) explored whether CL can learn concepts that are computationally hard to learn without curriculum. It thus focused on learning k-parity functions of d uniformly chosen (unbiased) input bits, which is difficult for iterative algorithms, such as SGD. Nonetheless, when the input bits are biased, SGD converges fast to low error, and thus implicitly identifies the support of the parity function. A CL approach is proposed based on domain adaptation, in which the input distribution is gradually shifted from deterministic (input bit is 1 with probability 1) to uniform (input bit is 1 with probability $1/2$). It is shown that the CL algorithm reduces the computational complexity from $d^{\Omega(k)}$ to $d^{\Omega(1)}$. Saglietti et al. (2022) proposed a CL model in which the task includes both relevant and irrelevant features, and a sample is difficult if it has irrelevant features with large variance. Analytical expressions were derived for the average learning trajectories of simple neural networks on this task, and it was shown that convergence is accelerated at the early iterations of training. A multi-stage CL algorithm in which the iterations of each stage are regularized with the solution of the previous stage was proposed, and the algorithm was analyzed in the high-dimensional limit. This showed the CL is most effective when there is a small number of relevant features.

Appendix B. Notation conventions

For a positive integer T we denote the sets $[T] := \{1, 2 \ldots, T\}$ as well as $[\![T]\!] := \{0\} \cup [T]$. For $a, b \in \mathbb{R}$, we denote $a \wedge b := \min\{a, b\}$ and $a \vee b := \max\{a, b\}$. We use standard Bachmann-Landau asymptotic notation, where, specifically, we use $\tilde{O}(\cdot)$ and $\tilde{\Omega}(\cdot)$ to hide poly-logarithmic factors. We also use the asymptotic relation $a \lesssim b$ to indicate that $a = O(b)$ and $a \asymp b$ to indicate that $a = \Theta(b)$. We denote the probability of an event $\mathcal{E}$ by $\mathbb{P}[\mathcal{E}]$, the indicator function for this event by $\mathbb{1}[\mathcal{E}]$, and the complement of this event by $\mathcal{E}^c$. We denote the d-dimensional Euclidean ball with radius r by $\mathbb{B}^d(r) := \{\theta \in \mathbb{R}^d \colon \|\theta\| \leq r\}$ and the $(d-1)$-dimensional unit sphere by $\mathbb{S}^{d-1} := \{\theta \in \mathbb{R}^d \colon \|\theta\| = 1\}$. We denote equivalence by $\equiv$ which is usually just a local simplification of notation. For a pair of probability distributions P and Q on a common alphabet $\mathcal{X}$ with densities p and q w.r.t. to a base measure ν, we denote the total variation distance by $\mathrm{d_{TV}}(P, Q) := \frac{1}{2} \int |p - q| \mathrm{d}\nu$ and the Kullback-Leibler (KL) divergence by $\mathrm{D_{KL}}(P \,||\, Q) := \int p \log(p/q) \mathrm{d}\nu$. For a pair of random variables (r.v.'s) $(X, Y) \sim P_{XY}$, the mutual information is denoted by $I(X; Y) := \mathrm{D_{KL}}(P_{XY} \,||\, P_X \otimes P_Y)$.

Appendix C. A high probability bound on the mean estimation loss

The following lemma is standard, and bounds the error of the empirical mean estimator. It is restated and proved here for completeness.

Lemma 7 *Consider N i.i.d. samples $\{Y(i)\}_{i\in[N]}$ from the model*

$$Y = \theta + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2\overline{\Sigma})$ where $\theta \in \mathbb{R}^d$ and $\overline{\Sigma} \in \overline{\mathbb{S}}_d^{++}$ ($\overline{\mathbb{S}}_d^{++}$ is the positive semidefinite cone of matrices whose trace is normalized to d). Let $\overline{\theta}(N) = \frac{1}{N}\sum_{i=1}^n Y(i)$ be the empirical mean estimator. Then, there exists a numerical constant $c > 1$ so that

$$\|\overline{\theta}(N) - \theta\|^2 \le c \log\left(\frac{e}{\delta}\right) \cdot \frac{d\sigma^2}{N}$$

with probability at least $1-\delta$, for any $\delta \in (0,1)$.

Proof It was shown by Rigollet and Rakhlin (2020, Lecture 4) that with probability at least $1-\delta$

$$\left| \|\overline{\theta}(N) - \theta\|^2 - \frac{\mathrm{Tr}[\sigma^2\overline{\Sigma}]}{N} \right| \le c\left(\frac{\|\sigma^2\overline{\Sigma}\|_{\mathrm{op}}}{N} \log\left(\frac{2}{\delta}\right) + \frac{\|\sigma^2\overline{\Sigma}\|_{\mathrm{F}}}{N}\sqrt{\log\left(\frac{2}{\delta}\right)} \right)$$

for some numerical constant $c > 0$, which can be taken to be $c \ge 1$. Since $\|\sigma^2\overline{\Sigma}\|_{\mathrm{op}} \le \|\sigma^2\overline{\Sigma}\|_{\mathrm{F}} \le \mathrm{Tr}[\sigma^2\overline{\Sigma}]$ it holds that

$$\begin{aligned}\|\overline{\theta}(N) - \theta\|^2 &\le \frac{\mathrm{Tr}[\sigma^2\overline{\Sigma}]}{N} \cdot c\left(1 + \log\left(\frac{e}{\delta}\right) + \sqrt{\log\left(\frac{e}{\delta}\right)}\right) \\ &\le \frac{\mathrm{Tr}[\sigma^2\overline{\Sigma}]}{N} \cdot 3c\log\left(\frac{e}{\delta}\right).\end{aligned}$$

■

Appendix D. The linear regression setting and high probability bounds on the loss

Rather than the parametric mean estimation problem considered in this paper, Xu and Tewari (2022) considered the parametric linear regression problem. We show that under the assumptions of Du et al. (2021); Xu and Tewari (2022), and for the purpose of the analysis in this paper, the parametric mean estimation and linear regression problems are similar, which allows to translate our results for the former to the latter, with slight modifications.

In the linear regression setting, the sample space is $\mathcal{Z} = \mathbb{R}^d \times \mathbb{R}$ and

$$\Phi_t := \mathbb{R}^d \times \mathbb{R}_+ \times \overline{\mathbb{S}}_d^{++}$$

where $\phi_t := (\theta_t, \sigma_t^2, \overline{\Sigma}_t)$, and $\overline{\mathbb{S}}_d^{++}$ is the positive semidefinite cone of matrices whose trace is normalized to d. The tth model, $t \in [\![T]\!]$, is given by

$$\mathcal{M}_t: \quad Y_t = \langle X_t, \theta_t\rangle + \epsilon_t, \tag{19}$$

where $X_t \sim \mathcal{N}(0, \overline{\Sigma}_t)$ is a random feature vector, where $\mathrm{Tr}[\overline{\Sigma}_t] = d$, $\theta_t \in \mathbb{R}^d$ is the unknown parameter vector, and $\epsilon_t \sim \mathcal{N}(0, \sigma_t^2)$ is a Gaussian noise. Upon samplinf from $\mathcal{M}_t$, the learner observes the $Z_t = (X_t, Y_t)$. The loss function is given by the squared prediction error

$$\ell\left(\phi, \tilde{\phi}\right) := \left(Y_t - \langle X_t^\top, \tilde{\theta}\rangle\right)^2 = \left(\langle X_t^\top, \theta - \tilde{\theta}\rangle + \epsilon_t\right)^2,$$

and the excess risk is

$$L_N(\phi, \mathcal{A}) = \mathbb{E}\left[\left(\langle X_t^\top, \theta - \hat{\theta}\rangle + \epsilon_0\right)^2 - \epsilon_0^2\right] = (\hat{\theta} - \theta_0)^\top \Sigma_0 (\hat{\theta} - \theta_0),$$

where $\hat{\theta} \equiv \hat{\theta}((A_i, S_i)_{i\in[N]})$. In addition to the assumption that $\|\theta_t\| \le C_\theta$ for all $t \in [\![T]\!]$, we follow Du et al. (2021); Xu and Tewari (2022) and make the following assumption:

Assumption 1 *For the linear regression setting, there exists constants $0 < \underline{C}_\Sigma < \overline{C}_\Sigma$ such that*

$$\underline{C}_\Sigma \cdot I_d \preceq \Sigma_t \preceq \overline{C}_\Sigma \cdot I_d$$

holds for any $t \in [\![T]\!]$.

Under Assumption 1 it can be asserted that (see the proof of Lemma 9 in what follows) that

$$L_N(\phi, \mathcal{A}) \asymp \mathbb{E}\left[\|\hat{\theta} - \theta_0\|^2\right], \tag{20}$$

that is, the excess risk is proportional to the MSE of the parameter, as in the mean estimation problem, for which $L_N(\phi, \mathcal{A}) = \mathbb{E}[\|\hat{\theta} - \theta_0\|^2]$. Therefore, while, in general, an accurate estimation of the parameter *suffices* for a low prediction risk, under Assumption 1, low prediction risk also *necessitates* accurate estimation. Therefore, we focused in this paper on the mean estimation setting in which the risk itself is the MSE of an estimator, rather than the linear regression setting. This estimation problem is simpler because it does not involve the randomness of the features, only the noise.

We next show that, similarly to the mean estimation setting, there exists an estimator $\bar{\theta}(N)$ and $g(\delta)\colon [0,1] \to \mathbb{R}_+$ such that

$$\|\bar{\theta}(N) - \theta\|^2 \le g(\delta) \cdot \frac{d\sigma^2}{N}$$

holds with probability at least $1 - \delta$, for any $\delta \in (0,1)$. Specifically, we will show next (Lemma 9) that if $\bar{\theta}(N)$ is the LS estimator, projected on $\mathbb{B}^d(C_\theta)$, then there exist universal constants $c, c_1, c_2 > 0$ such that if

$$N \ge \frac{1}{c_1} \vee \frac{c_0 \overline{C}_\Sigma}{\underline{C}_\Sigma} d \tag{21}$$

then (3) holds with $g(\delta) = {}^{c}/_{\underline{C}_\Sigma} \log^2({}^{4}/_{\delta})$. Evidently, the error in linear regression and mean estimation behaves roughly the same, where the former requires the condition (21), which can be assumed for the target model and all the T source models. The dependence on the error is also slightly different, and is $\Theta(\log^2({}^{1}/_{\delta}))$ for the linear regression setting compared to $\Theta(\log({}^{1}/_{\delta}))$ (sub-exponential) for the mean estimation setting.

To prove this, we will utilize the following high probability lower bound on the minimal singular value of a random matrix, due to Koltchinskii and Mendelson (2015, Theorem 3.1):

Theorem 8 (Koltchinskii and Mendelson (2015, Theorem 3.1)) *Let $\boldsymbol{X} \in \mathbb{R}^{N\times d}$ be a random matrix comprised of i.i.d. rows $\{X^\top(i)\}_{i\in[N]}$, which are equal in distribution to $X \in \mathbb{R}^N$. Assume that there exist constants a, A, B so that for every $u \in \mathbb{S}^{d-1}$*

$$a \le \sqrt{\mathbb{E}\left[\langle u, X\rangle^2\right]} \le A, \quad \sqrt{\mathbb{E}\left[\langle u, X\rangle^2\right]} \le B \cdot \mathbb{E}\left|\langle u, X\rangle\right|.$$

Then, there exist constants $c_0, c_1, c_2 > 0$ so that for any $N \ge c_0 B^4 (\frac{A}{a})^2 d$

$$\lambda_{\min}\left(\frac{1}{N}\boldsymbol{X}^\top\boldsymbol{X}\right) \ge c_2^2 \frac{a^2}{B^4}$$

with probability at least $1 - \exp(-c_1 B^4 N)$.

We may now prove the high probability bound on the loss of the projected LS estimator:

Lemma 9 *Consider N i.i.d. samples $\{X(i), Y(i)\}_{i\in[N]}$ from the model*

$$Y = \langle X, \theta\rangle + \epsilon$$

where $X \sim \mathcal{N}(0, \overline{\Sigma})$, with $\overline{\Sigma} \in \overline{\mathbb{S}}_d^{++}$, $\theta \in \mathbb{B}^d(C_\theta)$, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Let

$$\boldsymbol{X} = \begin{bmatrix} - & X^\top(1) & - \\ - & X^\top(2) & - \\ & \vdots & \\ - & X^\top(N) & - \end{bmatrix} \in \mathbb{R}^{N\times d}$$

and $\boldsymbol{Y} = (Y(1), Y(2), \cdots, Y(N))^\top \in \mathbb{R}^N$ and $\boldsymbol{\epsilon} = (\epsilon(1), \epsilon(2), \cdots, \epsilon(N))^\top \in \mathbb{R}^N$ so that $\boldsymbol{Y} = \boldsymbol{X}\theta + \boldsymbol{\epsilon}$. Consider the LS estimator

$$\overline{\theta}_{LS}(N) = \left(\boldsymbol{X}^\top\boldsymbol{X}\right)^\dagger \boldsymbol{X}^\top\boldsymbol{Y},$$

and let $\overline{\theta}(N)$ be $\overline{\theta}_{LS}(N)$ projected on $\mathbb{B}^d(C_\theta)$. Then, there exist numerical constants $c > 0$ and $c_0 \vee c_2 < 1 < c_1$ such that if

$$N \ge \underline{N} := \frac{1}{c_1} \vee \frac{c_0 \overline{C}_\Sigma}{\underline{C}_\Sigma} d \tag{22}$$

then

$$\|\overline{\theta}(N) - \theta\|^2 \le \frac{c}{c_2^2 \cdot \underline{C}_\Sigma} \log^2\left(\frac{4}{\delta}\right) \frac{d\sigma^2}{N} \tag{23}$$

with probability at least $1 - \delta$, for any $\delta \in (0, 1)$.

Proof It was shown by Rigollet and Rakhlin (2020, Lecture 6) that with probability at least $1 - \delta/2$

$$\left|\frac{1}{N}\|\boldsymbol{X}\overline{\theta}(N) - \boldsymbol{X}\theta\|^2 - \frac{\sigma^2 \cdot \mathrm{rank}(\boldsymbol{X}^\top\boldsymbol{X})}{N}\right| \le c\frac{\sigma^2}{N}\log\left(\frac{2}{\delta}\right)$$

for some numerical constant $c > 0$, which can be taken to be $c \ge 1$. Thus, with probability at least $1 - \delta/2$

$$\frac{1}{N}\left\|\boldsymbol{X}\left(\overline{\theta}(N) - \theta\right)\right\|^2 \le 2c\frac{d\sigma^2}{N}\log\left(\frac{2}{\delta}\right).$$

Moreover,

$$\frac{1}{N}\left\|\boldsymbol{X}\left(\overline{\theta}(N)-\theta\right)\right\|^2=\frac{1}{N}\left(\overline{\theta}(N)-\theta\right)^\top\boldsymbol{X}^\top\boldsymbol{X}\left(\overline{\theta}(N)-\theta\right)$$
$$\geq\frac{1}{N}\lambda_{\min}(\boldsymbol{X}^\top\boldsymbol{X})\cdot\left\|\overline{\theta}(N)-\theta\right\|^2$$

and so with probability at least $1-\delta/2$

$$\left\|\overline{\theta}(N)-\theta\right\|^2\leq\frac{2c\frac{\sigma^2 d}{N}\log\left(\frac{2}{\delta}\right)}{\frac{1}{N}\lambda_{\min}(\boldsymbol{X}^\top\boldsymbol{X})}. \tag{24}$$

Let $\sigma_{\min}(\boldsymbol{X})$ be the minimal singular value of a matrix $\boldsymbol{X}$. We next use (Koltchinskii and Mendelson, 2015, Theorem 3.1), reproduced as Theorem 8 above. To use this bound, we first verify the qualifying conditions of Koltchinskii and Mendelson (2015, Assumption 3.1). For brevity, let X be distributed as $X(1)\sim\mathcal{N}(0,\Sigma)$. Now, first, for any $u\in\mathbb{S}^{d-1}$

$$\mathbb{E}\left[\langle X,u\rangle^2\right]=\mathbb{E}\left[u^\top XX^\top u\right]=u^\top\Sigma_t u$$

and so Assumption 1 results

$$\sqrt{\underline{C}_\Sigma}\leq\sqrt{\mathbb{E}\left[\langle X,u\rangle^2\right]}\leq\sqrt{\overline{C}_\Sigma}.$$

Second, since $X_t\sim\mathcal{N}(0,\Sigma)$ it holds that $\langle X,u\rangle\sim\mathcal{N}(0,u^\top\Sigma u)$ and so $\sqrt{\mathbb{E}[\langle X,u\rangle^2]}=\sqrt{u^\top\Sigma u}$ and $\mathbb{E}|\langle X,u\rangle|=\sqrt{\frac{2}{\pi}u^\top\Sigma u}$. Hence, for any $u\in\mathbb{S}^{d-1}$ and $B\geq\sqrt{2/\pi}$ it holds that

$$\sqrt{\mathbb{E}[\langle X,u\rangle^2]}\leq B\cdot\mathbb{E}\left|\langle X,u\rangle\right|.$$

Given these properties, Koltchinskii and Mendelson (2015, Theorem 3.1) proved that there exist numerical constants $c_0,c_1,c_2>0$, so that the following holds: If

$$N\geq c_0\frac{\overline{C}_\Sigma}{\underline{C}_\Sigma}B^4 d \tag{25}$$

then

$$\frac{1}{N}\lambda_{\min}(\boldsymbol{X}^\top\boldsymbol{X})\geq\frac{c_2^2\cdot\underline{C}_\Sigma}{B^4} \tag{26}$$

with probability at least $1-\exp(-c_1B^4N)$.

Assume that $N\geq\underline{N}=(1/c_1)\vee(c_0\overline{C}_\Sigma d/\underline{C}_\Sigma)$. Let us choose

$$B^4=1\vee\left[\frac{1}{c_1N}\log\left(\frac{2}{\delta}\right)\right]$$

for which $B\geq\sqrt{2/\pi}$ holds. Consider two cases based on the value of $\delta\in(0,1)$: First, assume that δ is large enough so that

$$N\geq\sqrt{\frac{c_0\overline{C}_\Sigma}{c_1\underline{C}_\Sigma}\log\left(\frac{2}{\delta}\right)d}.$$

Then, the condition (25) of Koltchinskii and Mendelson (2015, Theorem 3.1), that is,

$$N \geq \frac{c_0 \overline{C}_\Sigma}{\underline{C}_\Sigma} \left\{ 1 \vee \left[\frac{1}{c_1 N} \log\left(\frac{2}{\delta}\right) \right] \right\} d,$$

holds and so (26) also holds, with probability at least $1 - \delta/2$. Combining with (24) and the union bound implies that for all $N \geq 1/c_1$

$$\begin{aligned} \left\| \overline{\theta}(N) - \theta \right\|^2 &\leq 2c \frac{1 \vee \left[\frac{1}{c_1 N} \log\left(\frac{2}{\delta}\right) \right]}{c_2^2 \cdot \underline{C}_\Sigma} \log\left(\frac{2}{\delta}\right) \frac{\sigma^2 d}{N} \\ &\leq \frac{2c}{c_2^2 \cdot \underline{C}_\Sigma} \log^2\left(\frac{4}{\delta}\right) \frac{\sigma^2 d}{N}, \end{aligned}$$

with probability at least $1 - \delta$. Second, assume that δ is too small, so that

$$N \leq \sqrt{\frac{c_0 \overline{C}_\Sigma}{c_1 \underline{C}_\Sigma} \log\left(\frac{2}{\delta}\right) d}.$$

then the condition (25) of Koltchinskii and Mendelson (2015, Theorem 3.1) does not hold. However, this case is equivalent to

$$\log\left(\frac{2}{\delta}\right) \geq \frac{N^2}{d} \frac{c_1 \underline{C}_\Sigma}{c_0 \overline{C}_\Sigma}. \tag{27}$$

Now, since $\|\theta\| \vee \|\overline{\theta}(N)\| \leq C_\theta$ it holds with probability 1 that

$$\left\| \overline{\theta}(N) - \theta \right\|^2 \leq 4 C_\theta^2.$$

If we further assume the condition

$$N \geq 2 \cdot \frac{c_0^{3/4} c_2^{1/2}}{c_1^{3/4} c^{1/2}} \frac{\overline{C}_\Sigma^{3/4}}{\underline{C}_\Sigma^{1/4}} C_\theta \frac{d^{1/4}}{\sigma},$$

then, it holds that

$$\begin{aligned} 4C_\theta^2 &\leq \frac{c_1^{3/2} c \underline{C}_\Sigma^{1/2}}{c_0^{3/2} c_2 \overline{C}_\Sigma^{3/2}} \cdot \frac{N^3}{d^{3/2}} \cdot \frac{\sigma^2 d}{N} \\ &\leq \frac{c_1^{3/2} c \underline{C}_\Sigma^{1/2}}{c_0^{3/2} c_2 \overline{C}_\Sigma^{3/2}} \frac{c_0^{3/2} \overline{C}_\Sigma^{3/2}}{c_1^{3/2} \underline{C}_\Sigma^{3/2}} \log^{3/2}\left(\frac{2}{\delta}\right) \cdot \frac{\sigma^2 d}{N} \\ &= \frac{c}{c_2 \underline{C}_\Sigma} \log^{3/2}\left(\frac{2}{\delta}\right) \cdot \frac{\sigma^2 d}{N} \end{aligned}$$

where the second inequality follows from (27).

Combining both cases, we obtain

$$\left\| \overline{\theta}(N) - \theta \right\|^2 \leq \frac{2c}{c_2^2 \cdot \underline{C}_\Sigma} \log^2\left(\frac{4}{\delta}\right) \frac{\sigma^2 d}{N}$$

using $c_2 < 1$ (so that $1/c_2^2 \leq 1/c_2$). This implies that (23) holds with probability at least $1 - \delta$. ■

Appendix E. Comments on Xu and Tewari (2022)

E.1. The minimax risk lower bound of Xu and Tewari (2022, Theorem 1)

Xu and Tewari (2022, Theorem 1) derived a lower bound on the minimax risk of the strong-oracle learner. The bound is derived for the linear regression model, but is essentially the same for the mean estimation problem we consider, and so we describe that result in the former setting (using our notation). This bound addresses the minimax risk of a CL algorithm for the class $\mathbf{\Psi}_{\leq}(\boldsymbol{q})$ (defined in (18)). The minimax risk is derived for the strong oracle, which knows $\{Q_t^2\}_{t\in[T]}$, where $Q_t := \|\theta_t - \theta_0\|$, and is given by

$$\frac{d\sigma_0^2}{N} \wedge \min_{t\in[T]} \left\{ \frac{d\sigma_t^2}{N} + Q_t^2 \right\}. \tag{28}$$

This manifests that it is rate optimal for a strong oracle to allocate all samples to a single model. The proof of this result appears in Xu and Tewari (2022, Appendix A), and is based on the proof of Mousavi Kalan et al. (2020), which provided a lower bound for the risk of a target task based on a single source model, in a transfer learning problem. Both proofs reduce the learning problem to an hypothesis testing problem, and use Fano's method to obtain the lower bound. Nonetheless, Mousavi Kalan et al. (2020) considered a transfer learning problem, in which the number of samples allocated to each of the models is fixed in advanced. By contrast, the strong oracle utilized in the proof of Xu and Tewari (2022, Theorem 1) may optimize the number of samples allocated to each of the models based on a (partial) knowledge of $\{\theta_t\}_{t\in[\![T]\!]}$. Thus, the CL problem should be reduced to an hypothesis testing problem in which the number of samples allocated to each of the models is determined adaptively. In turn, Fano's lower bound should be adapted to this setting, as its standard version (Cover and Thomas, 2006; Wainwright, 2019) provides a lower bound on the error probability in an hypothesis testing problem with a fixed number of samples for each of the models. This is not reflected in the proof of Xu and Tewari (2022, Appendix A), which writes the lower bound resulting from Fano's method as (in our notation)

$$U^2 \left(1 - \frac{\log 2 + \sum_{t=1}^{T} N_t \cdot I(J; Y_t)}{\log K} \right) \tag{29}$$

where U^2 is the claimed lower bound on the minimax rate, K is the number of hypotheses, $J \sim \text{Uniform}[K]$. Importantly, in this bound, N_t, which is the number of samples allocated to each of the models, statistically depends on J, but then it is not obvious that (29) is a valid application of Fano's bound. That being said, the specific construction of hypotheses used in the proof, indeed results the lower bound (28), and we adopted a similar construction in our proofs. In short, source task parameters with $q_t \gtrsim U$ are chosen to be the same for all hypotheses, making them independent of J, and source task parameters with $q_t \lesssim U$ are chosen as a U-packing set, with equal parameters across these "close" models, for each hypothesis. Thus, the allocation of samples to these source models can be chosen w.l.o.g. as allocating all samples to just one of them. This reduces the problem to a single source task, and then it can be assumed, with only a constant factor in the lower bound, that the source task and the target task are each allocated $N/2$ of the samples (with probability 1). Thus, the lower bound is intact.

In Appendix J.3 we compare our minimax lower bound (Theorem 5) with the bound of Xu and Tewari (2022, Theorem 2).

E.2. The algorithm and risk upper bound of Xu and Tewari (2022, Theorem 3)

Xu and Tewari (2022, Section 3.3) proposed and analyzed a two-step CL algorithm. The setting assumes that there is at least one source model t^* whose parameter equals exactly to the target parameter, namely $Q_{t^*} = 0$. Translating their algorithm for linear regression to our mean estimation setting, the algorithm can be described as first allocating $N/(2T)$ samples to each of the T source models, to obtain an initial estimates $\{\overline{\theta}_t(N/(2T)\}_{t\in[T]}$, using $N/2$ to obtain an initial estimate for the target parameter $\overline{\theta}_0(N/2)$ and then choosing

$$t^* = \operatorname*{argmin}_{t\in[T]} \left\|\overline{\theta}_t(N/(2T)) - \overline{\theta}_0(N/2)\right\|$$

so t^* is an r.v. This algorithm was analyzed by Xu and Tewari (2022, Appendix C), though its statement in (Xu and Tewari, 2022, Theorem 3) is inaccurate, and thus was revised by Xu (2023, Theorem 3.3) to the bound

$$\|\overline{\theta}_{t^*}(N/(2T) - \theta_0\| \lesssim \log\left(\frac{Td}{\delta}\right) \cdot \left(\frac{\sigma_0^2}{N} + \frac{d\sigma_{t^*}^2}{N/T} + \sqrt{\frac{d}{N}}\right), \tag{30}$$

which holds with probability at least $1-\delta$ (In (Xu and Tewari, 2022, Theorem 3) only the first two terms appear). In (30), the third term, $\sqrt{d/N}$, prevents this bound from achieving the typical fast parametric rate $d\sigma^2/N$ associated with risks under mean squared error (MSE), and one which does not vanish when the noise variances vanish, i.e., $\max_{t\in[\![T]\!]}\sigma_t \downarrow 0$. The first term, σ_0^2/N, does not appear in our bound, and its proof of Xu and Tewari (2022, Appendix C, Lemma 4) is somewhat unclear. In our setting and notation, the claim therein can be roughly rephrased as "*martingale concentration inequality on the sum* $\|\hat{\theta}_t - \theta_0^*\| \cdot \sum_{i=1}^{N/2}\epsilon_{0,i}$" where $\{\epsilon_{0,i}\}$ are i.i.d. samples from $\mathcal{N}(0,\sigma_0^2)$. However, $\sum_{i=1}^{N/2}\epsilon_{0,i}$ are actually bounded w.h.p. as $O(\sigma_0/\sqrt{N})$, and not as $O(\sigma_0^2/N)$. We next discuss the second term in (30), $d\sigma_{t^*}^2/(N/T)$. Due to the assumption that there exists t with $Q_t = 0$ one may speculate that this term would generalize to

$$\tilde{O}\left(Q_{t^*}^2 + \frac{d\sigma_{t^*}^2}{N/T}\right)$$

when this assumption does not hold. It is not clear what can be assured on t^*, and so it cannot be directly compared to our bound, in which t^* is chosen from a subset $\mathcal{T} \subseteq [\![T]\!]$, where under favorable conditions it holds that $\mathcal{T} = \mathcal{T}_{\text{w.o.}}$. More importantly, our bound is

$$\min_{t\in\mathcal{T}} \tilde{O}\left(Q_t^2 + \frac{d\sigma_t^2}{N}\right)$$

and the noise term in our bound is $d\sigma_{t^*}^2/N$, which is a factor of T smaller than the second term in (30). Finally, we prove a high probability bound as well as a bound that holds for the risk (expected loss).

Appendix F. Proof of Theorem 1 (a single source model $T = 1$)

The success of the method proposed in Theorem 1 it is based on the ability of the learner to reliably eliminate the source model whenever Q_1^2 is large. This ability is proved in the following *source*

elimination lemma. It is stated in slightly greater generality than needed for the $T = 1$ setting, in the sense that it also considers the case in which the noise variance in the source model is *larger* than the noise variance in the target model (that is, the case $\lambda_1^2 > \lambda_0^2$ in what follows). This form will be necessary for the analysis of the multiple source models $T > 1$.

Lemma 10 (The source elimination lemma) *For $t = 0, 1$, let $\theta_t, \tilde{\theta}_t \in \mathbb{R}^d$ and $\lambda_t \in \mathbb{R}_+$ be such that*

$$\|\tilde{\theta}_t - \theta_t\|^2 \le \lambda_t^2.$$

Set $Q_1^2 := \|\theta_1 - \theta_0\|^2$ and $\lambda_{max} := \lambda_0 \vee \lambda_1$. Then, there exists a numerical constant $\nu \in [1/27, 1]$ so that

$$\begin{cases} Q_1^2 \ge \nu \cdot \lambda_{max}^2, & \text{if } \|\tilde{\theta}_0 - \tilde{\theta}_1\|^2 \ge 10\lambda_{max}^2 \\ Q_1^2 \le \frac{\lambda_{max}^2}{\nu}, & \text{otherwise} \end{cases}. \tag{31}$$

Proof We consider three cases. First suppose that $Q_1^2 \le \lambda_{\max}^2$. Then,

$$\begin{aligned} \|\tilde{\theta}_0 - \tilde{\theta}_1\|^2 &= \|\tilde{\theta}_0 - \theta_0 + \theta_1 - \tilde{\theta}_1 + \theta_0 - \theta_1\|^2 \\ &\overset{(*)}{\le} 2\|\tilde{\theta}_0 - \theta_0\|^2 + 2\|\theta_1 - \tilde{\theta}_1 + \theta_0 - \theta_1\|^2 \\ &\overset{(*)}{\le} 2\|\tilde{\theta}_0 - \theta_0\|^2 + 4\|\theta_1 - \tilde{\theta}_1\|^2 + 4\|\theta_0 - \theta_1\|^2 \\ &\le 2\lambda_0^2 + 4\lambda_1^2 + 4Q_1^2 \\ &\le 10\lambda_{\max}^2, \end{aligned}$$

using $(a+b)^2 \le 2a^2 + 2b^2$ in $(*)$.[2] This shows that the first case in (31) holds even for $\nu = 1$. Second, suppose that $Q_1^2 \ge 27 \cdot \lambda_{\max}^2$. Then,

$$\begin{aligned} \|\tilde{\theta}_0 - \tilde{\theta}_1\| &\overset{(a)}{\ge} \|\tilde{\theta}_1 - \theta_0\| - \|\tilde{\theta}_0 - \theta_0\| \\ &\overset{(a)}{\ge} \|\theta_1 - \theta_0\| - \|\tilde{\theta}_1 - \theta_1\| - \|\tilde{\theta}_0 - \theta_0\| \\ &\ge Q_1 - \lambda_1 - \lambda_0 \\ &\overset{(b)}{\ge} \sqrt{27} \cdot \lambda_{\max} - 2\lambda_{\max} \\ &\ge \sqrt{10}\lambda_{\max}, \end{aligned}$$

where (a) follows from the triangle inequality (twice), (b) follows since $Q_1^2 \ge 27 \cdot \lambda_{\max}^2$ was assumed. This shows that the second case in (31) holds for $\nu = 1/27$. Thirdly, if $\lambda_{\max}^2 \le Q_1^2 \le 27\lambda_{\max}^2$ then both statements in (31) hold (for $\nu = 1/27$). Note that the numerical constant ν is not optimized, and so there exists some maximal $\nu \in [1/27, 1]$ for which the statement also holds. ■

We next prove Theorem 1 using the source elimination lemma, and show that the estimated output essentially achieves the loss (5) of the strong-oracle learner.

2. The constant can be slightly improved if we use the refined inequality $(a+b)^2 \le (1+\zeta^{-1})a^2 + (1+\zeta)b^2$ and optimize over $\zeta > 0$. Note also that $\mathbb{E}[\|\tilde{\theta}_0 - \tilde{\theta}_1\|^2] = \lambda_0^2 + \lambda_1^2 + Q_1^2 \le 3\lambda_{\max}^2$, and since $\langle \tilde{\theta}_0 - \theta_0, \tilde{\theta}_1 - \theta_1 \rangle \lesssim \lambda_0^2/\sqrt{d}$ w.h.p. (and similarly for the other mixed terms), the random error concentrates fast around this expected value. Thus, the constant 10 appearing in the bound can be improved by at most a factor of $10/3$.

Proof [of Theorem 1] Let $\mathcal{E}_\delta$ be the event in which

$$\|\bar{\theta}_t(N/2) - \theta\|^2 = \|\tilde{\theta}_t - \theta\|^2 \leq g\left(\frac{\delta}{2}\right)\frac{d\sigma_t^2}{N/2} =: \lambda_t^2$$

holds for both $t = 0$ and $t = 1$. By the union bound and the definition of $g(\cdot)$, it holds that $\mathbb{P}[\mathcal{E}_\delta] \geq 1 - \delta$. Under our assumptions $\lambda_{\max} = \lambda_0 \geq \lambda_1$. Assuming $\mathcal{E}_\delta$ occur, Lemma 10 implies that if $\|\tilde{\theta}_0 - \tilde{\theta}_1\|^2 \geq 10\lambda_0^2$ then $Q_1^2 + \lambda_1^2 \geq \nu \cdot \lambda_0^2$ and so by the choice in (8) $\hat{\theta} = \tilde{\theta}_0$ and

$$\|\hat{\theta} - \theta_0\|^2 = \|\tilde{\theta}_0 - \theta_0\|^2 \leq \lambda_0^2 = \lambda_0^2 \wedge \frac{Q_1^2 + \lambda_1^2}{\nu} \leq \frac{1}{\nu}\left[\lambda_0^2 \wedge (Q_1^2 + \lambda_1^2)\right].$$

Otherwise, if $\|\tilde{\theta}_0 - \tilde{\theta}_1\|^2 \leq 10\lambda_0^2$ then Lemma 10 implies that $\lambda_0^2 \geq \nu Q_1^2$ and so $2\lambda_0^2 \geq \nu(Q_1^2 + \lambda_1^2)$. So by the choice in (8) $\hat{\theta} = \tilde{\theta}_1$ and so

$$\begin{aligned}
\|\hat{\theta} - \theta_0\|^2 &= \|\tilde{\theta}_1 - \theta_0\|^2 \\
&= \|\theta_1 - \theta_0 + \tilde{\theta}_1 - \theta_1\|^2 \\
&\leq 2\|\theta_1 - \theta_0\|^2 + 2\|\tilde{\theta}_1 - \theta_1\|^2 \\
&\leq 2Q_1^2 + 2\lambda_1^2 \\
&\leq \frac{4}{\nu}\left[\lambda_0^2 \wedge (Q_1^2 + \lambda_1^2)\right].
\end{aligned}$$

Combining both cases shows the claimed bound in (9).

We next prove the bound on the risk (expected loss). It holds by Lemma 7 that

$$\|\tilde{\theta}_t - \theta_t\|^2 = \|\bar{\theta}_t(N/2) - \theta_t\|^2 \leq c\log\left(\frac{e}{\delta}\right) \cdot \frac{d\sigma_t^2}{N/2}$$

with probability at least $1 - \delta$, both for $t \in \{0, 1\}$. Thus, both are sub-exponential r.v.'s. Consequently, it holds that $\mathbb{E}[\|\tilde{\theta}_t - \theta_t\|^4] \lesssim [d\sigma_t^2/(N/2)]^2$. More accurately, note that for any $r \geq 0$

$$\mathbb{P}\left[\|\tilde{\theta}_t - \theta_t\|^4 \geq r\right] \leq e\exp\left[-\frac{\sqrt{r}}{c}\frac{N/2}{d\sigma_t^2}\right],$$

and so

$$\begin{aligned}
\mathbb{E}\left[\|\tilde{\theta}_t - \theta_t\|^4\right] &\overset{(a)}{=} \int_0^\infty \mathrm{d}r \cdot \mathbb{P}\left[\|\tilde{\theta}_t - \theta_t\|^4 \geq r\right] \\
&\leq e\int_0^\infty \mathrm{d}r \cdot \exp\left[-\frac{\sqrt{r}}{c}\frac{N/2}{d\sigma_t^2}\right] \\
&\overset{(b)}{=} e\left(\frac{cd\sigma_t^2}{N/2}\right)^2 2\int_0^\infty \mathrm{d}s \cdot s\exp\left[-s\right] \\
&= 2ec^2\left(\frac{d\sigma_t^2}{N/2}\right)^2, \qquad (32)
\end{aligned}$$

where (a) is by tail probability integration, and (b) is using the change of variables

$$s = \frac{\sqrt{r}}{c}\frac{N/2}{d\sigma_t^2},$$

for which

$$\frac{\mathrm{d}s}{\mathrm{d}r} = \frac{N/2}{cd\sigma_t^2}\frac{1}{2\sqrt{r}} = \left(\frac{N/2}{cd\sigma_t^2}\right)^2 \frac{1}{2s}.$$

Hence,

$$\begin{aligned}\mathbb{E}\left[\|\tilde{\theta}_1 - \theta_0\|^4\right] &= \mathbb{E}\left[\|\tilde{\theta}_1 - \theta_1 + \theta_1 - \theta_0\|^4\right] \\ &\overset{(a)}{\leq} 8\mathbb{E}\left[\|\tilde{\theta}_1 - \theta_1\|^4\right] + 8\mathbb{E}\left[\|\theta_1 - \theta_0\|^4\right] \\ &\overset{(b)}{\leq} 48c^2\left(\frac{d\sigma_1^2}{N/2}\right)^2 + 8Q_1^4,\end{aligned}$$

where (a) follows from $(a+b)^4 \leq 8a^4 + 8b^4$ (for $a, b \in \mathbb{R}$), and (b) follows from (32). Since $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ (for $a, b \geq 0$) we also get

$$\sqrt{\mathbb{E}\left[\|\tilde{\theta}_1 - \theta_0\|^4\right]} \leq \sqrt{48c^2\left(\frac{d\sigma_1^2}{N/2}\right)^2 + 8Q_1^2} \leq 7c\left[\frac{d\sigma_1^2}{N/2} + Q_1^2\right]. \tag{33}$$

Next we note that it holds with probability 1 that $\hat{\theta}$ is either $\tilde{\theta}_0$ or $\tilde{\theta}_1$, and so

$$\begin{aligned}\mathbb{E}\left[\|\hat{\theta} - \theta_0\|^4\right] &= \mathbb{E}\left[\|\tilde{\theta}_0 - \theta_0\|^4 \cdot \mathbb{1}\{\hat{\theta} = \tilde{\theta}_0\}\right] + \mathbb{E}\left[\|\tilde{\theta}_1 - \theta_0\|^4 \cdot \mathbb{1}\{\hat{\theta} = \tilde{\theta}_1\}\right] \\ &\leq \mathbb{E}\left[\|\tilde{\theta}_0 - \theta_0\|^4\right] + \mathbb{E}\left[\|\tilde{\theta}_1 - \theta_0\|^4\right].\end{aligned}$$

Thus,

$$\begin{aligned}\sqrt{\mathbb{E}\left[\|\hat{\theta} - \theta_0\|^4\right]} &\leq \sqrt{\mathbb{E}\left[\|\tilde{\theta}_0 - \theta_0\|^4\right]} + \sqrt{\mathbb{E}\left[\|\tilde{\theta}_1 - \theta_0\|^4\right]} \\ &\overset{(a)}{\leq} 2c\left(\frac{d\sigma_0^2}{N/2}\right) + 7c\left[\frac{d\sigma_1^2}{N/2} + Q_1^2\right] \\ &\leq 14c\left(\frac{d\sigma_0^2}{N} + \frac{d\sigma_1^2}{N} + Q_1^2\right),\end{aligned} \tag{34}$$

where (a) follows from (32) (with $t = 0$) and (33). Using the above, we decompose

$$\begin{aligned}\mathbb{E}\left[\|\hat{\theta} - \theta_0\|^2\right] &= \mathbb{E}\left[\|\hat{\theta} - \theta_0\|^2 \cdot \mathbb{1}[\mathcal{E}_\delta]\right] + \mathbb{E}\left[\|\hat{\theta} - \theta_0\|^2 \cdot \mathbb{1}[\mathcal{E}_\delta^c]\right] \\ &\overset{(a)}{\leq} \mathbb{E}\left[\|\hat{\theta} - \theta_0\|^2 \cdot \mathbb{1}[\mathcal{E}_\delta]\right] + \sqrt{\mathbb{E}\left[\|\hat{\theta} - \theta_0\|^4\right]\mathbb{P}[\mathcal{E}_\delta^c]} \\ &\overset{(b)}{\leq} \frac{14c}{\nu}\left[\log(\frac{2}{\delta}) \cdot \min\left\{Q_1^2 + \frac{d\sigma_1^2}{N}, \frac{d\sigma_0^2}{N}\right\} + \left(\frac{d\sigma_0^2}{N} + \frac{d\sigma_1^2}{N} + Q_1^2\right) \cdot \sqrt{\delta}\right],\end{aligned} \tag{35}$$

where (a) holds by Cauchy-Schwarz inequality, (b) holds by the high probability bound and (34). Choosing $\delta \leq \delta_0$ where

$$\delta_0 := \left[\frac{\min\left\{Q_1^2 + \frac{d\sigma_1^2}{N}, \frac{d\sigma_0^2}{N}\right\}}{\frac{d\sigma_0^2}{N} + \frac{d\sigma_1^2}{N} + Q_1^2}\right]^2$$

assures that

$$\mathbb{E}\left[\|\hat{\theta}-\theta_0\|^2\right] \leq \frac{14c}{\nu}\left(\log\left(\frac{2}{\delta}\right)+1\right)\cdot\min\left\{Q_1^2+\frac{d\sigma_1^2}{N},\frac{d\sigma_0^2}{N}\right\}. \tag{36}$$

It holds that

$$\begin{aligned}
\sqrt{\delta_0} &= \min\left\{\frac{Q_1^2+\frac{d\sigma_1^2}{N}}{\frac{d\sigma_0^2}{N}+\frac{d\sigma_1^2}{N}+Q_1^2},\frac{\frac{d\sigma_0^2}{N}}{\frac{d\sigma_0^2}{N}+\frac{d\sigma_1^2}{N}+Q_1^2}\right\}\\
&\geq \min\left\{\frac{\sigma_1^2}{\sigma_0^2+\sigma_1^2},\frac{1}{4},\frac{\frac{d\sigma_0^2}{N}}{2Q_1^2}\right\}\\
&\geq \min\left\{\frac{\sigma_1^2}{\sigma_0^2+\sigma_1^2},\frac{1}{4},\frac{d\sigma_0^2}{8C_\theta^2 N}\right\}\\
&= \sqrt{\delta_*},
\end{aligned}$$

where the first inequality holds since $(a_1+a_2)/(a_3+a_4) \geq (a_1/a_3)\wedge(a_2/a_4)$ for $a_1,a_2,a_3,a_4\geq 0$ and since $\sigma_1^2\leq\sigma_0^2$. Inserting δ_* to (36) and simplifying leads to the claim bound (10). ■

Appendix G. Proof of Theorem 4 (multiple source models $T>1$)

Proof [of Theorem 4] Algorithm 1 uses at most N samples since we set $\bar{N}=N/\bar{r}+2$. Assume that the event $\mathcal{E}_\delta$ occurs, in which the empirical averages $\bar{\theta}_0(\bar{N})$ and $\bar{\theta}_t(\bar{N}_{t,r})$ are close to θ_0 and θ_t, respectively, as

$$\|\bar{\theta}_0(\bar{N})-\theta_0\|^2 \leq g(\bar{\delta})\frac{d\sigma_0^2}{\bar{N}}$$
$$\|\bar{\theta}_t(\bar{N}_{t,r})-\theta_t\|^2 \leq g(\bar{\delta})\frac{d\sigma_t^2}{\bar{N}_{t,r}},$$

for all $\bar{r}$ rounds, and all $t\in[T]$, and similarly, that the last round estimate is close to θ_{t^*}, that is

$$\|\bar{\theta}_{t^*}(\bar{N})-\theta_{t^*}\|^2 \leq g(\bar{\delta})\frac{d\sigma_{t^*}^2}{\bar{N}}.$$

By the union bound, all these $T\bar{r}+2$ events hold with probability at least $1-\delta$. We continue the analysis assuming that $\mathcal{E}_\delta$ occurs. Under this event, Corollary 2 implies that after the first round only the source models $[T_1]$ with $T_1=T\beta_\delta(1)$ are retained, after the second round only $[T_2]$ with $T_2=T\beta_\delta(\beta_\delta(1))$, and so on. After $\bar{r}$ rounds, it holds that $\mathcal{T}_{\text{alg}}=[T_{\bar{r}}]$. The algorithm then chooses an arbitrary $t^*\in[T_{\bar{r}}]$ and the final high probability bound (15) stems from

$$\begin{aligned}
\|\hat{\theta}-\theta_0\|^2 &= \|\bar{\theta}_{t^*}(\bar{N})-\theta_0\|^2\\
&= \|\bar{\theta}_{t^*}(\bar{N})-\theta_0+\bar{\theta}_{t^*}(\bar{N})-\theta_{t^*}\|^2\\
&\overset{(a)}{\leq} 2\|\theta_{t^*}-\theta_0\|^2+2\|\bar{\theta}_{t^*}(\bar{N})-\theta_{t^*}\|^2\\
&\leq 2Q_{t^*}^2+2g(\bar{\delta})\cdot\frac{d\sigma_{t^*}^2}{\bar{N}}.
\end{aligned}$$

We next consider the bound in expectation, for which we set $g(\delta) = c\log(e/\delta)$. The proof of follows the same lines as the proof in the single source case (Theorem 1), and so we briefly outline it for the multiple source case. In general, we will be rather loose with the upper bounds, since they only affect a logarithmic factor. Note that the final empirical estimator uses $\bar{N}$ samples. So, if we let $\tilde{\theta}_t = \bar{\theta}_t(\bar{N})$ be the empirical estimator in the last round in which model t is retained, it holds that

$$\mathbb{E}\left[\|\tilde{\theta}_t - \theta_t\|^4\right] \leq 2ec^2\left(\frac{d\sigma_t^2}{\bar{N}}\right)^2.$$

Hence, similarly to the derivations leading to (33), it holds for any $t \in [T]$ that

$$\sqrt{\mathbb{E}\left[\|\tilde{\theta}_t - \theta_0\|^4\right]} \leq 48c^2\left(\frac{d\sigma_t^2}{\bar{N}}\right)^2 + 8Q_t^4.$$

Next we note that it holds with probability 1 that $\hat{\theta}$ is from the set $\{\tilde{\theta}_t\}_{t\in[\![T]\!]}$, and so

$$\mathbb{E}\left[\|\hat{\theta} - \theta_0\|^4\right] \leq \sum_{t\in[\![T]\!]} \mathbb{E}\left[\|\tilde{\theta}_t - \theta_0\|^4\right]$$

Since $\sqrt{\sum_i a_i} \leq \sum_i \sqrt{a_i}$ we get

$$\sqrt{\mathbb{E}\left[\|\hat{\theta} - \theta_0\|^4\right]} \leq \sum_{t\in[\![T]\!]} 7c\left[\frac{d\sigma_t^2}{\bar{N}} + Q_t^2\right]$$

(with $Q_0 = 0$). Using the above, we decompose as in (35) and rearrange to obtain

$$\begin{aligned}\mathbb{E}\left[\|\hat{\theta} - \theta_0\|^2\right] &= \mathbb{E}\left[\|\hat{\theta} - \theta_0\|^2 \cdot \mathbb{1}[\mathcal{E}_\delta]\right] + \mathbb{E}\left[\|\hat{\theta} - \theta_0\|^2 \cdot \mathbb{1}[\mathcal{E}_\delta^c]\right] \\ &\leq 8c(\bar{r}+2)\cdot\log\left(\frac{(T\bar{r}+2)}{\delta}\right)\cdot\left[\frac{d\sigma^2_{\bar{t}(\mathcal{T}_{\text{alg}})}}{N} + Q^2_{\bar{t}(\mathcal{T}_{\text{alg}})} + \left(\sum_{t\in[\![T]\!]}\left[\frac{d\sigma_t^2}{\bar{N}} + Q_t^2\right]\right)\sqrt{\delta}\right],\end{aligned}$$

Choosing $\delta \leq \delta_0$ where now

$$\delta_0 := \left[\frac{\frac{d\sigma^2_{\bar{t}(\mathcal{T}_{\text{alg}})}}{N} + Q^2_{\bar{t}(\mathcal{T}_{\text{alg}})}}{\sum_{t\in[\![T]\!]}\left[\frac{d\sigma_t^2}{\bar{N}} + Q_t^2\right]}\right]^2$$

assures that

$$\mathbb{E}\left[\|\hat{\theta} - \theta_0\|^2\right] \leq 16c(\bar{r}+2)\cdot\log\left(\frac{(T\bar{r}+2)}{\delta}\right)\cdot\left[\frac{d\sigma^2_{\bar{t}(\mathcal{T}_{\text{alg}})}}{N} + Q^2_{\bar{t}(\mathcal{T}_{\text{alg}})}\right]. \tag{37}$$

It holds that

$$\sqrt{\delta_0} = \frac{\frac{d\sigma^2_{\bar{t}(\mathcal{T}_{\text{alg}})}}{N} + Q^2_{\bar{t}(\mathcal{T}_{\text{alg}})}}{\sum_{t\in[\![T]\!]}\left[\frac{d\sigma_t^2}{\bar{N}} + Q_t^2\right]}$$

$$
\begin{aligned}
&\geq \frac{\min_{t\in[\![T]\!]} \frac{d\sigma_t^2}{N} + Q_t^2}{\sum_{t\in[\![T]\!]} \left[\frac{d\sigma_t^2}{N} + Q_t^2\right]} \\
&\geq \frac{1}{(T+1)} \cdot \frac{\min_{t\in[\![T]\!]} \frac{d\sigma_t^2}{N} + Q_t^2}{\max_{t\in[\![T]\!]} \frac{d\sigma_t^2}{N} + Q_t^2} \\
&\geq \frac{1}{(T+1)} \cdot \frac{\min_{t\in[\![T]\!]} \frac{d\sigma_t^2}{N} + Q_t^2}{\max_{t\in[\![T]\!]} \frac{d\sigma_t^2}{N} + 4C_\theta^2} \\
&\geq \frac{1}{(T+1)} \cdot \frac{\min_{t\in[\![T]\!]} \frac{d\sigma_t^2}{N}}{\max_{t\in[\![T]\!]} \frac{d\sigma_t^2}{N} + 4C_\theta^2} \\
&\geq \frac{1}{2(T+1)} \cdot \left[\min_{t,t'\in[\![T]\!]} \frac{\sigma_t^2}{\sigma_{t'}^2} \vee \min_{t\in[\![T]\!]} \frac{d\sigma_t^2}{8NC_\theta^2}\right] \\
&= \sqrt{\delta_*}.
\end{aligned}
$$

Inserting δ_* to (37) and simplifying leads to the claim bound (16). ■

Appendix H. An extension of Algorithm 1 to unknown noise covariance matrices

The covariance matrix of the noise in the mean estimation model $\mathcal{M}_t$ is $\Sigma_t = \sigma_t^2 \cdot \overline{\Sigma}_t$ where $\overline{\Sigma}_t \in \overline{\mathbb{S}}_d^{++}$. We have assumed thus far that $\{\Sigma_t\}_{t\in[\![T]\!]}$ are known to the learner. In this appendix, we consider the case in which the learner knows that $\overline{\Sigma}_t = I_d$ for all $t \in [\![T]\!]$, but not $\{\sigma_t^2\}_{t\in[\![T]\!]}$, and then the case that $\{\Sigma_t\}_{t\in[\![T]\!]}$ are completely unknown. We show that under rather mild conditions, Algorithm 1 can be further extended to these settings too. We focus on the mean estimation setting, though a similar extension can be made in the linear regression setting too (Appendix D).

The idea is straightforward and is based on a preliminary step of estimating either $\{\sigma_t^2\}_{t\in[\![T]\!]}$ (for the first setting) or $\{\Sigma_t\}_{t\in[\![T]\!]}$ (for the second setting), and then plugging in the estimated values into Algorithm 1, instead of the true values. Additional minor modifications then guarantee an upper bound on the loss which is of the same order as in the case the noise covariance is known to the learner. The aforementioned conditions assure that the additional error due to the use of inaccurate values of $\{\sigma_t^2\}_{t\in[\![T]\!]}$ or $\{\Sigma_t\}_{t\in[\![T]\!]}$ is negligible compared to the loss of the algorithm.

For the case of unknown $\{\sigma_t^2\}_{t\in[\![T]\!]}$, the condition is provided in Proposition 11, and requires that $dN \gtrsim \bar{r}T$. In the regime discussed above in Proposition 3 it is required that $\bar{r} = \Theta(\log T)$, and then the qualifying condition is $dN = \Omega(T \log T)$. This is a fairly mild condition since $N \geq T$ is assumed, and typically one would like to have in practice $N \gg T$ to allow exploration of all source models. The proof of Proposition 11 is based on Lemma 12, which provides a *multiplicative* confidence interval for an estimator $\hat{\sigma^2}(K)$ that uses K samples to estimate the variance σ^2. It is shown that $\frac{1}{2}\sigma^2 \leq \hat{\sigma^2}(K) \leq \frac{3}{2}\sigma^2$ with probability at least $1 - 2e^{-d(K-1)/24}$, and this result is used with $K = N/T$.

For the case of unknown $\{\Sigma_t\}_{t\in[\![T]\!]}$, the idea is similar, and compared to the case of unknown $\{\sigma_t^2\}_{t\in[\![T]\!]}$, only Lemma 12 is adapted. As shown in Lemma 7 (Appendix C), the noise covariance matrix Σ_t affects the estimation error of the empirical mean estimators $\overline{\theta}_t(N)$ error via $\varsigma_t^2 := \frac{1}{d}\operatorname{Tr}[\Sigma_t]$. Lemma 13 then replaces the confidence interval on the estimator of σ_t^2 with a

similar confidence interval on ς_t^2. It is shown that $\frac{1}{2}\varsigma^2 \le \hat{\varsigma^2}(K) \le \frac{3}{2}\varsigma^2$ with probability at least $1 - 2e^{-[(K-1)/24 - \log d]}$, and this result is used with $K = N/T$. Given Lemma 13, the required modifications to Proposition 11 are minor and thus not explicitly stated. The main modification is that the condition (38) is replaced by the more stringent condition $dN \gtrsim d \log d \cdot \bar{r}T$. That is, N in the case of unknown covariance matrix should be larger by a factor of $d \log d$ compared to the case of unknown variance.

H.1. Unknown noise variances

We employ a preliminary step to Algorithm 1 in which $\{\sigma_t^2\}_{t \in [\![T]\!]}$ are estimated. Then, the estimated values are plugged into Algorithm 1 instead of the true values. Since the performance guarantees of Algorithm 1 are based on the order $\Theta(d\sigma_t^2/N)$, it suffices that the estimated value of σ_t^2 should match the true value up to multiplicative constants. As we next show, under mild conditions, the additional error of Algorithm 1 due to this preliminary estimation step is negligible. The necessary conditions and modifications are stated in the following proposition:

Proposition 11 *Let $\delta \in (0,1)$ be given. Assume that $N \ge 2(\bar{r}+3)(T+1)$ and that*

$$dN \ge 48(\bar{r}+3)(T+1) \cdot \log \frac{2}{\bar{\delta}}. \tag{38}$$

Suppose that Algorithm 1 is run with the following changes:

1. *Line 1: Set*
$$\bar{N} \leftarrow \frac{N}{\bar{r}+3}$$
and
$$\bar{\delta} \leftarrow \frac{\delta}{T(\bar{r}+1)+3}.$$

2. *The values $\{\sigma_t^2\}_{t \in [\![T]\!]}$ are not input to the algorithm, and instead, there is an additional step after line 1: The noise variances are estimated as $\hat{\sigma_t^2}(\bar{N}/(T+1))$ for $t \in [\![T]\!]$.*

3. *Line 2: The number of samples allocated in the rth round to the $t \in \mathcal{T}_{r-1}$ is*
$$\hat{N}_{r,t} := \frac{\bar{N}}{T_{r-1}} \cdot \frac{\hat{\sigma_t^2}\left(\frac{\bar{N}}{T+1}\right)}{\frac{1}{T_{r-1}}\sum_{t=1}^{T_{r-1}} \hat{\sigma_t^2}\left(\frac{\bar{N}}{T+1}\right)}.$$

4. *Line 3: Model $t \in \mathcal{T}_{r-1}$ is eliminated at the rth round if*
$$\|\tilde{\theta}_0 - \tilde{\theta}_t\|^2 \ge 10 g(\bar{\delta}) \left[\frac{d \cdot \frac{1}{T_{r-1}} \sum_{t=1}^{T_{r-1}} \hat{\sigma_t^2}\left(\frac{\bar{N}}{T+1}\right)}{\bar{N}/T_{r-1}} \vee \frac{d\hat{\sigma_0^2}}{\bar{N}} \right].$$

Then, the performance guarantees of Theorem 4 hold with the pre-factor changed from $\bar{r}+2$ to $\bar{r}+3$ and in the definition of $\beta_\delta(\tau)$, the constant ν is replaced by $\nu/6$.

It should be stressed that items 3 and 4 can indeed be computed by the learner, since they do not depend on the unknown σ_t^2; only on their estimated values. As long as the mild condition (38) holds, the modified algorithm is efficient, and performs as well as Algorithm 1, up to numerical constants.

The proof of Proposition 11 is based on the following lemma:

Lemma 12 *Assume that $Y = \theta + \epsilon$ where $\theta \in \mathbb{R}^d$ and $\epsilon \sim \mathcal{N}(0, \sigma^2 \cdot I_d)$. Let $\delta \in (0,1)$ be given. Consider an estimator for σ^2 based on K i.i.d. samples $\{Y(i)\}_{i\in[K]}$ given by*

$$\hat{\sigma^2}(K) := \frac{1}{d(K-1)} \sum_{i=1}^{K} \left\| Y(i) - \frac{1}{K} \sum_{i'=1}^{K} Y(i') \right\|^2 . \tag{39}$$

Then, with probability at least $1 - 2e^{-d(K-1)/24}$

$$\frac{1}{2}\sigma^2 \leq \hat{\sigma^2}(K) \leq \frac{3}{2}\sigma^2.$$

Proof Let $Y(i) = (Y(i;1), \ldots, Y(i;d))$. As is well known, (*e.g.*, (Schervish and DeGroot, 2014, Theorem 8.3.1)) it holds for any $j \in [d]$ that

$$\tilde{\sigma^2}(j) = \frac{1}{K-1} \sum_{i=1}^{K} \left(Y(i;j) - \frac{1}{K} \sum_{i'=1}^{K} Y(i';j) \right)^2$$

is an unbiased estimator of σ^2 (that is, $\mathbb{E}[\tilde{\sigma^2}(j)] = \sigma^2$), and moreover, that

$$\frac{(K-1)}{\sigma^2} \cdot \tilde{\sigma^2}(j) \sim \chi^2_{K-1}.$$

Now, $\{\tilde{\sigma^2}(j)\}_{j\in[d]}$ are i.i.d., and so

$$\sum_{j=1}^{d} \frac{(K-1)}{\sigma^2} \cdot \tilde{\sigma^2}(j) \sim \chi^2_{d(K-1)}.$$

Hence,

$$\frac{d(K-1)}{\sigma^2} \cdot \hat{\sigma^2}(K) = \frac{1}{\sigma^2} \sum_{i=1}^{K} \sum_{j=1}^{d} \left(Y(i;j) - \frac{1}{K} \sum_{i'=1}^{K} Y(i';j) \right)^2 \sim \chi^2_{d(K-1)}.$$

By Laurent and Massart (2000, Lemma 1) or (Boucheron et al., 2013, Remark 2.11) (note that $\mathbb{E}[\chi^2_\ell] = \ell$) it holds that

$$\mathbb{P}\left[\left| \chi^2_\ell - \ell \right| \geq 2\sqrt{\ell t} + 2t \right] \leq 2e^{-t}.$$

and so

$$\mathbb{P}\left[\left| \hat{\sigma^2}(K) - \sigma^2 \right| \geq \left[2\sqrt{\frac{t}{d(K-1)}} + \frac{2t}{d(K-1)} \right] \cdot \sigma^2 \right] \leq 2e^{-t}.$$

Choosing $t = {}^{d(K-1)}/_{24}$ we get that the deviation pre-factor is

$$\frac{2\sqrt{\ell t} + 2t}{d(K-1)} \leq \frac{1}{2}$$

which, in turn, implies that

$$\frac{1}{2}\sigma^2 \leq \hat{\sigma^2}(K) \leq \frac{3}{2}\sigma^2$$

with probability at least $1 - 2\exp[-d(K-1)/24]$. ■

We may now prove Proposition 11.

Proof [of Proposition 11] In order to accommodate for the unknown variances, we need to consider their influence on a few steps of the algorithm. First, their influence on the additional samples required for variance estimation and the additional possible events that one of these estimated values is inaccurate. Second, their influence on the allocation of samples at each round, that is, the computation of $\bar{N}_{t,r}$. This, in turn, affects the condition used by the algorithm to eliminate models, and on the models which satisfy this condition.

The number of samples and the reliability: Instead of $\bar{r}+2$ batches of equal size samples (combined from all the models) we now have $\bar{r}+3$ including the one needed for variance estimation. So we redefine $\bar{N} = N/(\bar{r}+3)$. We add to the union bound over $T\bar{r}+2$ events, additional $T+1$ events, which assure that $\overline{\sigma_t^2}$ are on the same scale as σ_t^2 for all $t \in [\![T]\!]$. To handle this, we redefine $\bar{\delta} = \delta/T(\bar{r}+1)+3$, and require that

$$\frac{1}{2}\sigma_t^2 \leq \hat{\sigma_t^2}\left(\frac{\bar{N}}{T+1}\right) \leq \frac{3}{2}\sigma_t^2 \tag{40}$$

with probability at least $1-\bar{\delta}$. By Lemma 12, this holds if

$$2e^{-d(\bar{N}/(T+1)-1)/24} \leq \bar{\delta},$$

or, equivalently

$$d\left(\frac{N}{(\bar{r}+3)(T+1)} - 1\right) \geq 24 \cdot \log\frac{2}{\bar{\delta}}.$$

Under the assumption of the lemma, a simple (looser) sufficient condition for this to hold is (38).

The sample allocation at each round and the elimination condition: Algorithm 1 uses the allocation

$$\bar{N}_{t,r} = \frac{\bar{N}}{T_{r-1}} \cdot \frac{\sigma_t^2}{\overline{\sigma^2}(\mathcal{T}_{r-1})}$$

and thus depends on the noise variances. One may replace the exact variances with their estimates as

$$\hat{N}_{r,t} := \frac{\bar{N}}{T_{r-1}} \cdot \frac{\hat{\sigma_t^2}\left(\frac{\bar{N}}{T+1}\right)}{\frac{1}{T_{r-1}}\sum_{t=1}^{T_{r-1}}\hat{\sigma_t^2}\left(\frac{\bar{N}}{T+1}\right)}. \tag{41}$$

Corollary 2 guarantees that if

$$Q_t^2 \geq 2 \cdot \frac{10g(\bar{\delta})}{\nu}\left[\frac{d\cdot\sigma_t^2}{\hat{N}_{r,t}} \vee \frac{d\sigma_0^2}{\bar{N}}\right] \tag{42}$$

then

$$\|\tilde{\theta}_0 - \tilde{\theta}_t\|^2 \geq 20g(\bar{\delta})\left[\frac{d\cdot\sigma_t^2}{\hat{N}_{r,t}} \vee \frac{d\sigma_0^2}{\bar{N}}\right] \tag{43}$$

will occur w.h.p., and then $\mathcal{M}_t$ will be eliminated. Note that the above statement uses the actual number of samples $\hat{N}_{r,t}$ taken by the algorithm, yet the exact values of the noise variances σ_0^2, σ_t^2 affecting the empirical estimates $\tilde{\theta}_0$ and $\tilde{\theta}_t$. Also note the additional factor of 2 to be used. Now, under the high probability event (40), which is assumed to hold, the event in (43) also implies that

$$\begin{aligned}\|\tilde{\theta}_0 - \tilde{\theta}_t\|^2 &\geq 10g(\bar{\delta}) \left[\frac{d \cdot \hat{\sigma_t^2}\left(\frac{\bar{N}}{T+1}\right)}{\hat{N}_{r,t}} \vee \frac{d\hat{\sigma_0^2}}{\bar{N}}\right] \\ &= 10g(\bar{\delta}) \left[\frac{d \cdot \frac{1}{T_{r-1}}\sum_{t=1}^{T_{r-1}} \hat{\sigma_t^2}\left(\frac{\bar{N}}{T+1}\right)}{\bar{N}/T_{r-1}} \vee \frac{d\hat{\sigma_0^2}}{\bar{N}}\right],\end{aligned}$$

which is exactly the condition for elimination in the modified algorithm (see item 4 in the statement of the proposition). Thus, if (42) holds then $\mathcal{M}_t$ is eliminated. Now, using the expression for $\hat{N}_{r,t}$ in (41) we note that (42) is equivalent to

$$Q_t^2 \geq 2 \cdot \frac{10g(\bar{\delta})}{\nu} \left[\frac{d \cdot \frac{1}{T_{r-1}}\sum_{t=1}^{T_{r-1}} \hat{\sigma_t^2}\left(\frac{\bar{N}}{T+1}\right)}{\bar{N}/T_{r-1}} \cdot \frac{\sigma_t^2}{\hat{\sigma_t^2}\left(\frac{\bar{N}}{T+1}\right)} \vee \frac{d\sigma_0^2}{\bar{N}}\right]. \tag{44}$$

In turn, the high probability event (40) also implies that if

$$Q_t^2 \geq 3 \cdot 2 \cdot \frac{10g(\bar{\delta})}{\nu} \left[\frac{d \cdot \frac{1}{T_{r-1}}\sum_{t=1}^{T_{r-1}} \sigma_t^2}{\bar{N}/T_{r-1}} \vee \frac{d\sigma_0^2}{\bar{N}}\right]$$

then (44), and so also (42), hold. Evidently, this is the same condition as for Algorithm 1, where the numerical constant ν is replaced by $\frac{\nu}{6}$. ■

H.2. Unknown noise covariance matrices

Next assume that the covariance matrices $\{\Sigma_t\}_{t\in[\![T]\!]}$ are unknown to the learner, where $\Sigma_t := \sigma_t^2 \cdot \overline{\Sigma}_t$. An efficient estimator is still the empirical mean $\hat{\theta}_t(N)$, and this estimator does not depend on the unknown covariance matrix Σ_t (Lemma 7). For Algorithm 1, the required unknown parameter is then $\varsigma_t^2 = \frac{1}{d}\operatorname{Tr}[\Sigma_t]$ (replacing σ_t^2). Since $\operatorname{Tr}[\Sigma_t]$ equals to the sum of its diagonal elements, the estimator (39) proposed for the case $\Sigma_t = \sigma_t^2 \cdot I_d$ is still intact. Nonetheless, in that case, the d coordinates are independent and reduce the variance in estimating σ^2, whereas here they are not independent. The following lemma modifies Lemma 12 to an estimator of $\varsigma^2 := \frac{1}{d}\operatorname{Tr}[\Sigma]$.

Lemma 13 *Assume that $Y = \theta + \epsilon$ where $\theta \in \mathbb{R}^d$ and $\epsilon \sim \mathcal{N}(0, \Sigma)$ with $\Sigma \in \mathbb{S}_d^{++}$. Let $\delta \in (0,1)$ be given. Consider an estimator for $\varsigma^2 := \frac{1}{d}\operatorname{Tr}[\Sigma]$ based on K i.i.d. samples $\{Y(i)\}_{i\in[K]}$ given by*

$$\hat{\varsigma^2}(K) := \frac{1}{d(K-1)} \sum_{i=1}^{K} \left\| Y(i) - \frac{1}{K}\sum_{i'=1}^{K} Y(i') \right\|^2. \tag{45}$$

Then, with probability at least $1 - 2e^{-[(K-1)/24 - \log d]}$

$$\frac{1}{2}\varsigma^2 \leq \hat{\varsigma^2}(K) \leq \frac{3}{2}\varsigma^2. \tag{46}$$

If we make the mild assumption $(K-1) \geq 48 \log d$ then the probability guaranteed by Lemma 13 is $1 - e^{-\Theta(K)}$ compared to the larger probability $1 - e^{-\Theta(dK)}$ guaranteed by Lemma 13.

Proof We follow the proof of Lemma 12. Now, $\sigma^2(j) = \Sigma(j,j)$ is the jth diagonal coordinate of Σ, and $\tilde{\sigma^2}(j)$ defined therein (the sample variance of $\{Y(i;j)\}_{i\in[K]}$) is an unbiased estimator of $\sigma^2(j)$ and moreover, it holds that

$$\frac{(K-1)}{\sigma^2(j)} \cdot \tilde{\sigma^2}(j) \sim \chi^2_{K-1}.$$

However, unlike the proof of Lemma 12, now $\{\tilde{\sigma^2}\}_{j\in[d]}$ are not necessarily independent, and thus their sum is not necessarily distributed as a chi-squared. To overcome this, we require the estimate of $\tilde{\sigma^2}(j)$ will be accurate for all d coordinates. Specifically, using the same arguments used in the proof of Lemma 12 (setting $d = 1$ therein), it holds that

$$\frac{1}{2}\sigma^2(j) \leq \tilde{\sigma^2}(j) \leq \frac{3}{2}\sigma^2(j) \tag{47}$$

with probability at least $1 - 2e^{-(K-1)/24}$. By the union bound, this holds for all $j \in [d]$ with probability at least

$$1 - 2de^{-(K-1)/24} = 1 - 2e^{-[(K-1)/24 - \log d]}.$$

The event that (47) holds for all $j \in [d]$ implies that (46) holds too. ■

Appendix I. Empirical experiments

In this appendix, we run an implementation of Algorithm 1 on simple simulated estimation settings, and empirically demonstrate the various properties of this elimination algorithm we have discussed during the theoretical analysis. Python code for these simulations can be found in https://anonymous.4open.science/r/CurriculumLearning-CAD5/. In what follows we will use $\tilde{Q}_t^2 := {}^{Q_t^2}/_{(d\sigma_0^2/N)}$ to denote normalized square distances.

In the first experiment, a two task setting $T = 2$ is considered. We fix the parameter of the first source model $t = 1$ at distance $Q_1 = 0$, and sweep over Q_2, the distance of parameter of the second source to the target parameter, where we normalize Q_2^2 by the typical MSE of the target task ${}^{d\sigma_0^2}/_N$. Figure 2 shows the loss of the algorithm on a logarithmic scale, and the empirical probability that the algorithm chooses task 2. The outcomes, as depicted in Figure 2, reveal two discernible regimes. In the initial regime, approximately when $\tilde{Q}_2^2 \in (0, 100)$, the set $\mathcal{T}_{\text{alg}}$ is consistently non-empty and the probability to choose task $t = 2$ is roughly 0.5. This implies that, within this regime, Algorithm 1 correctly identifies that both source models are better than the target model for estimating the target parameter, however, the algorithm is unable to consistently eliminate task 2. The observed increase in error aligns with the characteristics of a weak oracle error, as $Q_2^2 \lesssim {}^{d\sigma_0^2}/_N$. In the subsequent regime, when $\tilde{Q}_2^2 \in (100, 300)$, the probability of selecting task $t = 2$ gradually diminishes until reaching 0, concurrently with a reduction in the error as $\tilde{Q}_2^2$ increases. This stems from the fact that $Q_2^2 \gtrsim {}^{d\sigma_0^2}/_N$ and is the result of the improved ability of Algorithm 1 to eliminate the source model $t = 2$, and then use the better source model $t = 1$ for estimation of the target parameter.

In the second experiment, we consider settings with large number of models. We define three types of parameter locations, relative to the target parameter, designated as "close", "medium", and

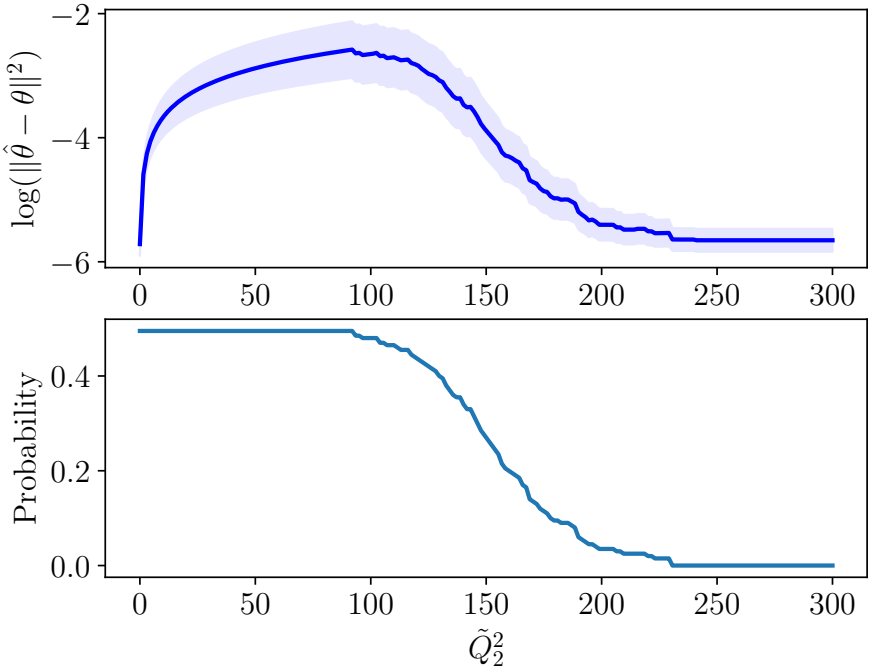

Figure 2: The first experiment: Runs of Algorithm 1 over 200 repetitions. Parameters are $T = 2,\ N = 1000,\ d = 2,\ \sigma^2 = 1,\ \sigma_0^2 = 10$.

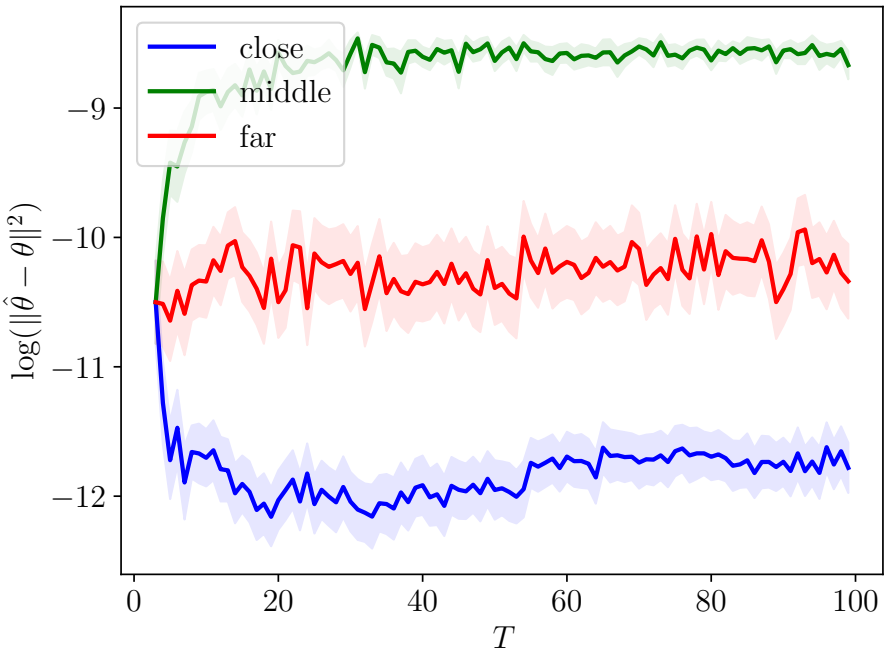

Figure 3: The second experiment: Runs of Algorithm 1 over 200 repetitions. Parameters are $N = 10^5,\ d = 2,\ \sigma^2 = 0.1,\ \sigma_0^2 = 1,\ \tilde{Q}^2_{\text{close}} = 0,\ \tilde{Q}^2_{\text{medium}} = 10,\ \tilde{Q}^2_{\text{far}} = 2 \cdot 10^4$.

"far". We consider a sequence of problem settings, in which we first set one source parameter at each of the location types. Then, we add more models from one of the types. As can be seen in Figure 3, Algorithm 1 behaves differently in response to the addition of models in different location types. The inclusion of models in proximity to the target task facilitates better estimation error, a phenomenon that aligns with intuitive expectations. Conversely, in scenarios with a substantial number of models categorized as "medium," the algorithm struggles to eliminate those models and thus achieve high MSE. Notably, when models are introduced in the "far" location the algorithm eliminates them effectively and the addition of models from this category does not affect the algorithm's loss. The experiment appearing in Section 4 generalizes this experiment to other mixtures of types of source models.

In the third experiment, we set $T = 100$ source models, and choose

$$\tilde{Q}_t^2 = \begin{cases} 0, & t \in [10] \\ (t-10)^\gamma, & t \in \{11, 12, \ldots, 100\} \end{cases},$$

for which various values of γ were examined. For each value we generated the corresponding $\beta_\delta(\tau)$, in order to demonstrate the connection between $\beta_\delta(\tau)$ and the algorithm's efficacy in model elimination. To this end, we record for each run of Algorithm 1 which models were eliminated and which were not. We evaluate the algorithm performance using the *precision* score, defined as

$$\text{precision} := \frac{\left|\mathcal{T}_{\text{alg}} \cap \mathcal{T}_{\text{w.o.}}\right|}{\left|\mathcal{T}_{\text{alg}}\right|},$$

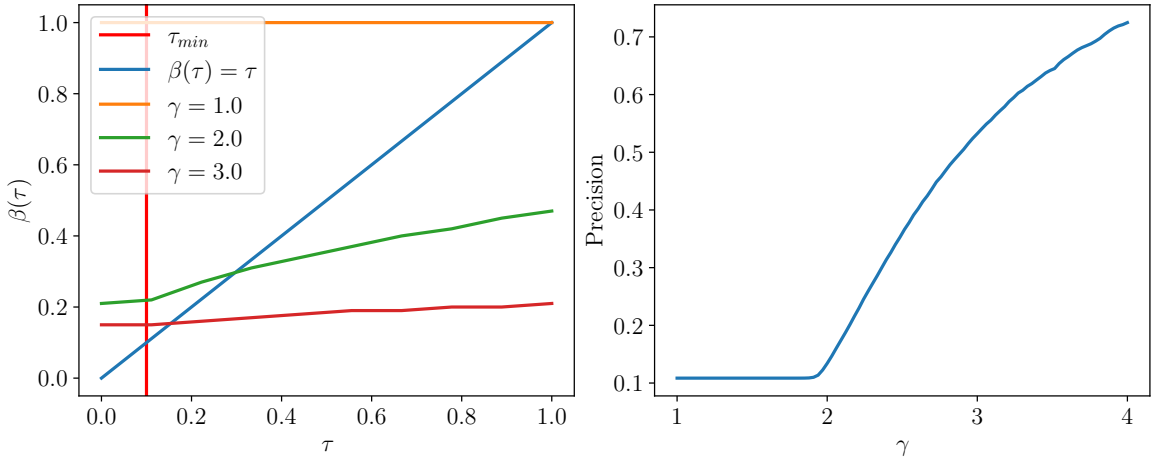

Figure 4: The third experiment: Parameters are $N = 10^5,\ d = 2,\ \sigma^2 = 0.1,\ \sigma_0^2 = 1,\ \delta = 0.05$.

and the *recall* score, defined as

$$\text{recall} := \frac{\left|\mathcal{T}_{\text{alg}} \cap \mathcal{T}_{\text{w.o.}}\right|}{\left|\mathcal{T}_{\text{w.o.}}\right|}.$$

As seen in Figure 4, as γ is increases, the location of the fixed point $\beta_\delta(\tau) = \tau$ decreases, which should result a better precision in eliminating models. This is simply because increasing γ also increases the number of tasks at distant locations, which are easier for the algorithm to eliminate in the first round, thus allowing for further elimination in second round, and so on. In all our experiments, Algorithm 1 has achieved perfect recall, that is, all models in the weak oracle set are identified as such by the algorithm, and as seen in Figure 4, the precision improves with increasing γ.

Appendix J. Minimax lower bounds: Extended discussion and proofs for Section 5

J.1. A general reduction to hypothesis testing

For a CL algorithm, the N observed samples are not i.i.d., because the learner may select the model from which the next sample is collected based on previous samples. In accordance, the reduction to hypothesis testing should be made to a similarly adaptive tester. For completeness, we next formally define the hypothesis testing problem and then the reduction (Proposition 14). The problem formulation is similar to the general learning CL setting in Section 2, except that the set of possible models is of finite cardinality K, and one of the models is chosen by Nature with a uniform probability. The learner should then collect samples in order to identify the model, and its performance is judged by the error probability. Let K be a positive integer. Let $\mathbf{\Theta}_{\text{test}} := \{\theta^{(j)}\}_{j\in[K]} \subset \mathbf{\Psi} \subset (\mathbb{R}^d)^{T+1}$ be a *testing set* of parameters, each one of them $\theta^{(j)} = (\theta_0^{(j)}, \theta_1^{(j)}, \ldots, \theta_T^{(j)})$ is comprised of parameters for the $T+1$ models, where the index j represents the hypothesis. Let $J \sim \text{Uniform}[K]$ be an r.v. that chooses the hypothesis. Given that $J = j$, the samples of the $T+1$ models are obtained from the model $\mathcal{M}_t$ with parameter $\theta_t^{(j)}$, for all $t \in [\![T]\!]$. A CL hypothesis tester collects samples as a general CL learner, described in Section 2. Given the collected samples $(A_i, S_i)_{i\in[N]}$, the tester

produces a guess of J, given by a testing function $\hat{J}\colon (([\![T]\!])\times\mathcal{Z})^N \to [K]$. A CL testing algorithm is the pair $\mathcal{A} := (\pi, \hat{J})$, comprised of a sampling policy and a testing function. The expected error probability of the tester when the testing set is $\mathbf{\Theta}_{\text{test}}$ and it uses an algorithm $\mathcal{A}$ is

$$e_N(\mathbf{\Phi}_{\text{test}}, \mathcal{A}) := \mathbb{P}_{J,\mathcal{A}}\left[\hat{J}\left((A_i, S_i)_{i\in[N]}\right) \neq J\right],$$

where $\mathbf{\Phi}_{\text{test}} := \{\phi^{(j)}\}_{j\in[K]}$, $\phi^{(j)} = (\phi_0^{(j)}, \phi_1^{(j)}, \ldots, \phi_T^{(j)})$ and $\phi_t^{(j)} := (\theta_t^{(j)}, \sigma_t^2, I_d)$.

As is well known, the estimation problem is more difficult than hypothesis testing, and this reduction is formulated below:

Proposition 14 (A variation of Wainwright (2019, Proposition 15.1)) *Let $\mathbf{\Phi}_{test}$ be given such that $\mathbf{\Theta}_{test} \subset \mathbf{\Psi}$, and let*

$$\eta \equiv \eta(\mathbf{\Theta}_{test}) := \min_{j_1,j_2\in[K]\colon\, j_1\neq j_2} \frac{1}{2}\|\theta_0^{(j_1)} - \theta_0^{(j_2)}\|. \tag{48}$$

Then, for any CL testing algorithm $\mathcal{A} = (\pi, \hat{\phi})$

$$L_{N,d}(\mathbf{\Psi}, \mathbf{\Gamma}) \geq \eta^2 \cdot \min_{\mathcal{A}} e_N(\mathbf{\Phi}_{test}, \mathcal{A}). \tag{49}$$

Proof Let $\mathcal{A}_* = (\pi_*, \hat{\theta}_*)$ be the optimal policy and estimator in the CL setting, and let $\hat{J}$ be an arbitrary testing function. Then,

$$\begin{aligned} L_{N,d}(\mathbf{\Psi}, \mathbf{\Gamma}) &= \sup_{\boldsymbol{\theta}\in\mathbf{\Psi}} \mathbb{E}_{\boldsymbol{\phi},\mathcal{A}_*}\left[\left\|\hat{\theta}_*\left((A_i, S_i)_{i\in[N]}\right) - \theta_0\right\|^2\right] \\ &\overset{(*)}{\geq} \eta^2 \cdot e_N(\mathbf{\Phi}_{\text{test}}, (\pi_*, \hat{J})) \\ &\geq \eta^2 \cdot \min_{\mathcal{A}} e_N(\mathbf{\Phi}_{\text{test}}, \mathcal{A}), \end{aligned}$$

where $(*)$ holds since when the policy is fixed to π_*, the CL learner estimates the parameter, or tests the hypothesis from a standard statistical model, for which (49) was originally derived in (Wainwright, 2019, Proposition 15.1). ■

Utilizing Proposition 14, the goal is now to derive a lower bound on the error probability of CL tests. This is, however, challenging, since lower bounds for adaptive tests are not as widely available, compared to Le-Cam's and Fano's bounds for non-adaptive testers. We next propose two simple ways to circumvent the need for bounds on the error probability of adaptive testers. One idea was used in the proof of Xu and Tewari (2022, Theorem 1), where in Appendix E.1 we have discussed in detail its proof, and highlighted a few subtleties related to it.

First, the CL problem is easier if the learner can collect N samples from the target task and from each of the source models, and thus a total of $N \cdot (T+1)$ samples. We state this simple idea as a general claim.

Claim 1 *Let $\boldsymbol{Y} := \{Y_{i,t}\}_{i\in[N],t\in[\![T]\!]}$ be a collection of $N\cdot(T+1)$ independent r.v.'s so that $Y_{i,t}$ is drawn from the model $\mathcal{M}_t$ in (2). Let $\hat{\vartheta}$ be an estimator for θ_0 based on $\boldsymbol{Y}$. Then*

$$L_{N,d}(\mathbf{\Psi}, \mathbf{\Gamma}) \geq \inf_{\hat{\vartheta}} \sup_{\boldsymbol{\theta}\in\mathbf{\Psi}} \mathbb{E}_{\boldsymbol{Y}\sim\boldsymbol{\theta}}\left[\left\|\hat{\vartheta}\left(\boldsymbol{Y}\right) - \theta_0\right\|^2\right] := \tilde{L}_{N,d}(\mathbf{\Psi}, \mathbf{\Gamma}).$$

Proof The proof is trivial data-processing argument since any adaptive CL algorithm $\mathcal{A} = (\pi, \hat{\theta})$ that collects a total of N samples may first collect the $N(T+1)$ samples of $\boldsymbol{Y}$, run the sampling policy π, and only use the N samples collected by the policy to compute the estimator estimate $\hat{\theta}((A_i, S_i)_{i\in[N]})$. ■

This method does not capture the correct dependence of the risk on T, and so is effective whenever the number of source models is fixed.

Second, under proper assumptions, it can be argued that an optimal learner does not sample from some of the models. For example, source models whose parameter is known to be at large distance from the target parameter, or that are known to have large noise, may be ignored by the learner. This is the origin of the claim by Xu and Tewari (2022, Theorem 1), that a strong-oracle learner, which is fully informed of $\{Q_t^2\}_{t\in[T]}$, may only sample from the model with minimal Q_t^2 (assuming, for simplicity, that the noise variance of all source models is the same). In general, it appears to be difficult to rigorously prove this claim in an estimation context, however, it can be proved for a test, that is, after the reduction to a finite set of hypotheses in $\boldsymbol{\Theta}_{\text{test}}$. Concretely, will use the following two arguments:

Claim 2 *Let $\boldsymbol{\Theta}_{test}$ be a given testing set.*

- *If $\theta_t^{(j)}$ is the same for all $j \in [K]$ then an optimal policy in the CL hypothesis testing problem does not sample from $\mathcal{M}_t$.*
- *If there exists $\alpha \in \mathbb{R}$ so that $\theta_{t_1}^{(j)} = \alpha\theta_{t_2}^{(j)}$ for some $t_1, t_2 \in [T]$ and all $j \in [K]$ and $\sigma_{t_2}^2/\alpha^2 \geq \sigma_{t_1}^2$ then an optimal tester does not sample from $\mathcal{M}_{t_2}$.*

Proof In the first case, the observations of $\mathcal{M}_t$ are independent of J. In the second case, the learner can simulate samples from $\mathcal{M}_{t_2}$ by sampling from $\mathcal{M}_{t_1}$, scaling, and adding independent Gaussian noise with proper variance. ■

J.2. The choice of the localization set and associated challenges

We have discussed in Section 5 the choice of the set $\boldsymbol{\Psi}$, and showed that choosing it to be too small may allow the learner to bypass the actual difficulty of the CL problem, because given that $\boldsymbol{\theta} \in \boldsymbol{\Psi}$ may already reveal too much information to the learner, and allow it to obtain unrealistically low risk. Thus, the choice of the set $\boldsymbol{\Psi}$ should be judicious in order to obtain a non-trivial localized bound. We have provided one example, in which the choice of the set $\boldsymbol{\Psi}_{=}(\boldsymbol{q})$ with a specific choice of $\boldsymbol{q}$ allowed the learner to use a majority rule, which obtains an unrealistic zero risk. This shows that no trivial lower bound can be derived. Here we continue presenting such examples, and show that in some cases the correct dependency of the lower bound on d cannot be obtained, simply because the learner can effectively reduce the dimension of the problem to a lower effective dimension.

Example 2 *Assume noiseless source models similarly to the example in Section 5, but that $T = 2$ and that $q_0^2 = q_1^2 = 0 < q_2^2 \lesssim d\sigma_0^2/N$. The loss of the weak oracle in (13) is given by $\|\hat{\theta}_{w.o.} - \theta_0\|^2 \lesssim q_2^2$. Contrary to example in Section 5, a learner which collects one sample from each source model cannot decide which one of them is located at the target parameter and which one of them is at distance q_2 from the target parameter. However, let $\tilde{\theta}$ be such that $\|\tilde{\theta} - \theta_0\| = q_2$. Then, given the two samples, say Y_1, Y_2, the learner can sample directly from the target model, and project the*

samples in the direction of $Y_1 - Y_2 = \pm(\theta_0 - \tilde{\theta})$. *This reduces the noise variance to* $\frac{\sigma_0^2}{N}$. *So, if* $q_2^2 \gtrsim \sigma_0^2/N$ *then the learner can identify whether* $Y_1 = \theta_0$ *or* $Y_2 = \theta_0$. *The resulting lower bound is smaller by a factor of* d *compared to the desired lower bound.*

The example in Section 5 and Example 2 appear to be specialized to cases in which there exists a source parameter that exactly equals the target parameter, and the noise variance of the source models is zero. However, this is not the case, as we next explain. To this end, recall that the testing set is chosen as a packing (or separated) set of target parameters, for which the distance between each pair of parameters is larger than some q. Then, if it can be shown that the error probability in this hypothesis testing problem is bounded away from zero, this implies a lower bound of $\Theta(q^2)$ on the minimax risk. If we let $t_{\text{w.o.}}$ be the maximal source model index in the weak-oracle set, that is, the largest t so that $Q_t^2 \lesssim d\sigma_0^2/N$, then ideally we would like to set $q = q_{\text{w.o.}}$ in order to establish the weak-oracle learner risk as the lower bound on the minimax risk. In this case, due to the triangle inequality, source parameters which are at distance less than $q/2$ from the target parameter are effectively the same as ones at zero distance, in the sense that if the learner knows that a source parameter is at distance $Q_t \leq q/2$, it can identify the target parameter in the packing set, and thus identify the hypothesis with zero error probability. This leads to a zero lower bound on the estimation loss. This will continue to hold if the noise variance is non-zero, yet sufficiently small. Thus, the aspects illuminated in Section 5 and Example 2 depict the actual challenges of deriving minimax lower bounds by a reduction to a finite set of hypotheses for the CL problem.

For the simple case of a one-dimensional parameter $d = 1$, the problem is even more severe, since if the learner knows the distance between θ_t and θ_0 this already reduces the ambiguity of the learner to just two possibilities. This is demonstrated in the following example.

Example 3 *Assume that* $d = 1$, *and* $T = 1$. *In this case, the source parameter may be just one of two possibilities* $\theta_1 = \theta_0 + sq_1$, *where* $s \in \{\pm 1\}$ *is an unknown sign. The learner can actually achieve an error on the order of* σ_1^2/N *(which is better than the weak oracle) even if* $q_1^2 \gtrsim \sigma_0^2/N \gtrsim \sigma_1^2/N$. *To achieve this, the CL learner will estimate* θ_0 *(resp.* θ_1*) via an empirical mean estimator* $\overline{\theta}_0(N)$ *(resp.* $\overline{\theta}_1(N)$*), to obtain an error of* $\Theta(\sigma_0/\sqrt{N})$ *(resp.* $\Theta(\sigma_1/\sqrt{N})$*); both error bounds hold w.h.p. Under the assumption on* q_1, *and under the high probability event, the learner can guess the sign* $\hat{s} := \operatorname{sgn}(\overline{\theta}_1(N) - \overline{\theta}_0(N))$, *and succeed w.h.p. Then, it will estimate the target parameter as* $\overline{\theta}_1(N) - \hat{s} \cdot q_1$. *Such a learner will achieve a squared error loss of* $\Theta(\sigma_1^2/N)$, *which is better than the weak oracle rate given by* $\Theta(\sigma_0^2/N)$.

Evidently, the initial uncertainty of a CL algorithm goes beyond an unknown sign. One may claim that this is a result of the learner knowing exactly that $|\theta_1 - \theta_0| = q_1$, but the same method can be used even if q_1 is only known approximately.

J.3. A comparison of Theorem 5 with the lower bound of Xu and Tewari (2022) and its proof

Comparison to Xu and Tewari (2022, Theorem 2) The lower bound of Xu and Tewari (2022, Theorem 2) is developed under the simplifying assumption that $q_1 = 0$ and $q_t = \overline{q}$ for $t \in [T]\backslash\{1\}$ and is given by

$$L_{N,d}\left(\mathbf{\Psi}_{\leq}(\boldsymbol{q}), \mathbf{\Gamma}\right) \gtrsim \frac{\sigma_0^2 \log T}{N} \wedge \overline{q}^2 + \min_{t\in[T]} \frac{d\sigma_t^2}{N}. \tag{50}$$

The term $\min_{t\in[T]} d\sigma_t^2/N$ comes from the strong-oracle lower bound of Xu and Tewari (2022, Theorem 1), and so we next focus on the case in which the first term dominates. Thus we focus the

comparison on $\sigma_t^2 = 0$ for all $t \in [T]$, for which the lower bound is $\Omega(\sigma_0^2 \log T/N \wedge \overline{q}^2)$. For the case at hand, and ignoring logarithmic terms, if $\sigma_0^2/N \lesssim \overline{q}^2 \lesssim d\sigma_0^2/N$ then the lower bound of Theorem 5 is $\Omega(\overline{q}^2)$ which matches the weak-oracle learner risk, and is better than the risk lower bound $\tilde{\Omega}(\sigma_0^2/N)$ of Xu and Tewari (2022, Theorem 2). If $\overline{q}^2 \lesssim \sigma_0^2/N$ then both bounds are of the same order, and match the weak-oracle learner risk. Finally, if $\overline{q}^2 \gtrsim d\sigma_0^2/N$ then the construction of Xu and Tewari (2022, Appendix B) is such that the distance between the distant models parameter and the target parameter is $q_t \asymp \sigma_0/\sqrt{N} \lesssim \overline{q}$ for $t \in [T]\backslash\{1\}$. Thus, all models are in $\mathcal{T}_{\text{w.o.}}$, and in turn, all bounds are on the order $\tilde{\Omega}(\sigma_0^2/N)$.[3]

Proof outline Following Claim 1, we focus on bounding $\tilde{L}_{N,d}(\mathbf{\Psi}, \mathbf{\Gamma})$ rather than its adaptive-sampling counterpart. We then prove Theorem 5 in a few gradual steps. For a fixed dimension, there is no need to invoke Fano's method (as Mousavi Kalan et al. (2020) did), and the simpler Le-Cam's two-point method suffices to obtain lower bounds. As a warm-up, we recall this method for the standard mean estimation case. We then consider the single source task case $T = 1$ (considered by Mousavi Kalan et al. (2020)), and provide a self-contained proof, in a form that will allow us to generalize it to $T > 1$. We then extend this proof to $T = 2$ source models, which provides the clearest intuition on why it is the performance of the weak-oracle learner that is achieved rather than that of the strong-oracle learner. Then, in order to illuminate the challenge of extending such an argument to a general $T > 2$, we explicitly consider the case $T = 3$. We aim to provide a construction of an hypothesis testing problem which naturally generalizes the $T = 2$ case, but then explicitly point out the problematic aspect of such an extension. We then prove Theorem 5 by reducing $T > 2$ to the $T = 2$ case, which is already solved.

We will repeatedly use the following expression for the KL divergence between Gaussian r.v.'s:

Fact 1 ((Wainwright, 2019, Exercise 15.13)) *For $\mu_1, \mu_2 \in \mathbb{R}^d$ and $\Sigma_1, \Sigma_2 \in \mathbb{S}_{++}^d$ it holds that*

$$D_{KL}\left(\mathcal{N}(\mu_1, \Sigma_1) \,||\, \mathcal{N}(\mu_2, \Sigma_2)\right) = \frac{1}{2}\log\frac{|\Sigma_2|}{|\Sigma_1|} - \frac{d}{2} + \frac{1}{2}\operatorname{Tr}\left[\Sigma_1\Sigma_2^{-1}\right] + \frac{1}{2}\left(\mu_2 - \mu_1\right)^\top \Sigma_2^{-1}\left(\mu_2 - \mu_1\right).$$

Warm-up: Le-Cam's method for the Gaussian location model Recall Le-Cam's two-point method: A family of probability measures $\{P_\theta\}_{\theta\in\Theta}$ is given, where θ is a parameter in the class Θ. Assume for simplicity that $\Theta \subset \mathbb{R}$ and the squared loss function. Le-Cam's method is based on reducing the estimation problem of θ based on a sample from P_θ, to a binary hypothesis problem between a pair $\theta^{(0)}, \theta^{(1)} \in \Theta$. If the optimal test between $P_{\theta^{(0)}}$ and $P_{\theta^{(1)}}$ with equal prior has large average error probability (bounded away from zero), then the estimation error must be at least on the order of $(\theta^{(0)} - \theta^{(1)})^2$. In turn, the error probability is characterized using the total variation between $P_{\theta^{(0)}}$ and $P_{\theta^{(1)}}$ (which can later be bounded using the Hellinger or the KL divergence, which are simpler to analyze). The resulting bound is given, *e.g.*, in (Wainwright, 2019, Eq. (15.14)),

$$\inf_{\hat{\theta}} \sup_{\theta\in\Theta} \mathbb{E}\left[\|\theta - \hat{\theta}\|^2\right] \geq \frac{(\theta^{(0)} - \theta^{(1)})^2}{2} \cdot \left[1 - \mathrm{d_{TV}}(P_{\theta^{(0)}}, P_{\theta^{(1)}})\right],$$

3. The bound (50) of Xu and Tewari (2022) in case the first term dominates is interpreted as "*a potential improvement of a factor of $d/\log T$ when $\overline{q}$ and σ_t^2 are small*" (Xu and Tewari, 2022, p. 4, after Theorem 2); this is due to the optimal risk of $\Theta(d\sigma_0^2/N)$ obtained by sampling only from the target model (when no source models are available), and then taking the ratio $(\sigma_0^2 \log T/N)/(d\sigma_0^2/N)$. However, more carefully stated, this is the gain only when $\overline{q}^2 \gtrsim \sigma_0^2 \log T/N \gtrsim \min_{t\in[T]} d\sigma_t^2/N$. Otherwise, the ratio between the r.h.s. of (50) and $d\sigma_0^2/N$ is different from $d/\log T$.

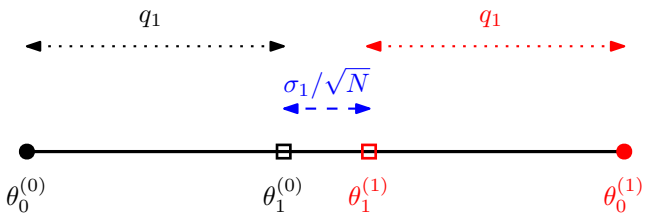

Figure 5: The parameter hypotheses for $T = 1$. One of the hypotheses is in black and the other in red. The target parameter is designated by a disc and the source parameter by a square.

which can be optimized over the choice of the pair $\theta^{(0)}$ and $\theta^{(1)}$. Using Pinsker's inequality (Wainwright, 2019, Eq. (3.59)), it holds for a Gaussian location model $\mathcal{N}(\theta, \sigma^2/N)$ (say P_θ represents N independent measurements from the same model) that (Fact 1)

$$\mathrm{d_{TV}}(P_{\theta^{(0)}}, P_{\theta^{(1)}}) \le \sqrt{\frac{1}{2}\mathrm{D_{KL}}(P_{\theta^{(0)}}, P_{\theta^{(1)}})} = \frac{|\theta^{(0)} - \theta^{(1)}|}{2\sigma/\sqrt{N}}.$$

Thus, choosing $\theta^{(0)}$ and $\theta^{(1)}$ to be any pair satisfying $|\theta^{(0)} - \theta^{(1)}| = \sigma/N$, implies that

$$\inf_{\hat{\theta}} \sup_{\theta_0 \in \Theta} \mathbb{E}\left[\|\theta - \hat{\theta}\|^2\right] \ge \frac{1}{8}\frac{\sigma^2}{N}.$$

A single source model ($T = 1$) We next consider the single source model setting, that is $T = 1$. While this model was considered by Mousavi Kalan et al. (2020), we provide here a self-contained proof, which is tailored to the CL setting, and which will be generalized later on for $T > 1$. In the case of $T = 1$, the learner knows that the parameter of the model $\mathcal{M}_1$ is at squared distance at most q_1^2 from the true parameter.

Proposition 15 *If* $\sigma_0^2 \ge \sigma_1^2$ *then*

$$L_{N,1}\left(\mathbf{\Psi}_{\le}(q_1), \mathbf{\Gamma}\right) \ge \frac{1}{20} \cdot \min\left\{\frac{\sigma_0^2}{N}, q_1^2 + \frac{\sigma_1^2}{N}\right\}.$$

The proof of the lower bound in the most interesting regime $q_1^2 \lesssim \sigma_0^2/N$ (the source is useful) is based on the construction of a binary hypothesis testing between the parameters shown in Figure 5. In this case, the distances $|\theta_t^{(0)} - \theta_t^{(1)}|$ are below the noise standard error $\sigma_t/\sqrt{N}$ for both $t = 0$ (target) and $t = 1$ (source). As a result, the error in the binary hypothesis testing is large, which leads to squared loss of $|\theta_0^{(0)} - \theta_0^{(1)}| \asymp q_1^2 + \sigma_1^2/N$. The proof in the regime in which the source is useless since $q_1^2 \gtrsim \sigma_0^2/N$ is based on the monotonicity of $\tilde{L}_{N,d}(\mathbf{\Psi}_{\le}(q_1), \mathbf{\Gamma})$ in $\boldsymbol{Q}$.

Proof Let $V := q_1 + \sigma_1/(4\sqrt{N})$ for brevity, and consider the case $V^2 \le \sigma_0^2/(4N)$. We construct the following binary hypothesis testing problem between

$$\boldsymbol{\theta}^{(0)} := (\theta_0^{(0)}, \theta_1^{(0)}) = (-V, -V + q_1)$$

and

$$\boldsymbol{\theta}^{(1)} := (\theta_0^{(1)}, \theta_1^{(1)}) = (V, V - q_1)$$

using N samples from each of the models. Using Fact 1, it holds that

$$\begin{aligned}\mathrm{D}_{\mathrm{KL}}(P_{\boldsymbol{\theta}^{(0)}}, P_{\boldsymbol{\theta}^{(1)}}) &= N \cdot \mathrm{D}_{\mathrm{KL}}(P_{\theta_0^{(0)}}, P_{\theta_0^{(1)}}) + N \cdot \mathrm{D}_{\mathrm{KL}}(P_{\theta_1^{(0)}}, P_{\theta_1^{(1)}}) \\ &= N \cdot \frac{2V^2}{\sigma_0^2} + N\frac{2\left(\frac{\sigma_1}{4\sqrt{N}}\right)^2}{\sigma_1^2} \\ &\leq \frac{1}{2} + \frac{1}{8} = \frac{5}{8},\end{aligned}$$

where the inequality follows from the assumptions. By Pinsker's inequality (Wainwright, 2019, Eq. (3.59))

$$\mathrm{d}_{\mathrm{TV}}(P_{\boldsymbol{\theta}^{(0)}}, P_{\boldsymbol{\theta}^{(1)}}) \leq \sqrt{\frac{1}{2}\mathrm{D}_{\mathrm{KL}}(P_{\boldsymbol{\theta}^{(0)}}, P_{\boldsymbol{\theta}^{(1)}})} \leq \frac{\sqrt{5}}{4}.$$

Then, Le-Cam's method implies that

$$\begin{aligned}\inf_{\hat{\vartheta}} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Psi}_{\leq}(q_1,\boldsymbol{\Gamma})} \mathbb{E}_{\boldsymbol{Y}\sim\boldsymbol{\phi}}\left[\left|\hat{\vartheta}\left(\boldsymbol{Y}\right) - \theta_0\right|^2\right] &\geq \frac{(\theta_0^{(0)} - \theta_0^{(1)})^2}{2} \cdot \left[1 - \mathrm{d}_{\mathrm{TV}}(P_{\boldsymbol{\theta}^{(0)}}, P_{\boldsymbol{\theta}^{(1)}})\right] \\ &\geq \frac{(\theta^{(0)} - \theta^{(1)})^2}{5} \\ &= \frac{4\left(q_1 + \frac{\sigma_1}{4\sqrt{N}}\right)^2}{5}. \end{aligned} \tag{51}$$

The bound (51) implies that for $V^2 = \sigma_0^2/(4N)$ it holds that

$$\inf_{\hat{\vartheta}} \sup_{\boldsymbol{\theta}\in\boldsymbol{\Psi}_{\leq}(q_1,\boldsymbol{\Gamma})} \mathbb{E}_{\boldsymbol{Y}\sim\boldsymbol{\phi}}\left[\left|\hat{\vartheta}\left(\boldsymbol{Y}\right) - \theta_0\right|^2\right] \geq \frac{1}{5}\cdot\frac{\sigma_0^2}{N},$$

which together with the monotonicity of $\tilde{L}_{N,d}(\boldsymbol{\Psi}_{\leq}(q_1), \boldsymbol{\Gamma})$ in q_1 and (51) imply the lower bound

$$\tilde{L}_{N,d}(\boldsymbol{\Psi}_{\leq}(q_1), \boldsymbol{\Gamma}) \geq \min\left\{\frac{4\left(q_1 + \frac{\sigma_1}{4\sqrt{N}}\right)^2}{5}, \frac{1}{5}\cdot\frac{\sigma_0^2}{N}\right\}.$$

This is then further weakened to obtain the stated lower bound. ■

Two source models ($T = 2$) We next consider the setting of two source models, $T = 2$. In this setting, the learner has to decide if either of the two source models is useful for the estimation of the target parameter. We show that in this case, the lower bound on the risk matches the risk of the weak-oracle learner.

Proposition 16 *If $q_0^2 = 0 \leq q_1^2 \leq q_2^2$ and $\sigma_0^2 \geq \sigma_1^2 \geq \sigma_2^2$ then*

$$L_{N,1}\left(\boldsymbol{\Psi}_{\leq}(\boldsymbol{q}), \boldsymbol{\Gamma}\right) \geq \begin{cases} \frac{1}{24}\cdot\left(q_2^2 + \frac{\sigma_2^2}{N}\right), & q_1^2 \leq q_2^2 \leq \frac{1}{16}\frac{\sigma_0^2}{N} \\ \frac{1}{720}\cdot\left(q_1^2 + \frac{\sigma_1^2}{N}\right), & q_1^2 \leq \frac{1}{16}\frac{\sigma_0^2}{N} \leq q_2^2 \\ \frac{1}{45}\cdot\frac{\sigma_0^2}{N}, & q_2^2 \geq q_1^2 \geq \frac{1}{16}\frac{\sigma_0^2}{N} \end{cases}. \tag{52}$$

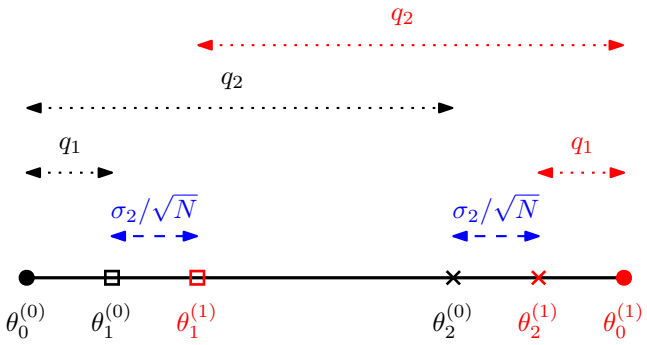

Figure 6: The parameter hypotheses for $T = 2$. One of the hypotheses is in black and the other in red. The target parameter is designated by a disc, the first source parameter by a square, and the second by a cross.

$$\geq \frac{1}{720} \cdot \left\{ \frac{d\sigma_{w.o.}^2}{N} + q_{w.o.}^2 \right\} \tag{53}$$

The final lower bound matches the risk of the weak-oracle learner.

The proof of the lower bound in the case $T = 2$ is more complicated than the proof in the single source case ($T = 1$), yet it is conceptually similar. In the most insightful regime, in which $q_2^2 \lesssim \sigma_0^2/N$, the proof is based on the construction of a binary hypothesis testing between the parameters shown in Figure 6. Again, the distances $|\theta_t^{(0)} - \theta_t^{(1)}|$ are below the noise standard error $\sigma_t/\sqrt{N}$ for $t \in \{0, 1, 2\}$ (target and two sources). As a result, the error in the binary hypothesis testing is large, which leads to squared loss of $|\theta_0^{(0)} - \theta_0^{(1)}| \asymp q_1^2 + q_2^2 + \sigma_2^2/N \asymp q_2^2 + \sigma_2^2/N$. The proof in the regime in which the second source is useless but the first one is useful since $q_1^2 \lesssim \sigma_0^2/N \lesssim q_2^2$ is based on a construction similar to the $T = 1$ case (*i.e.*, the second source can effectively be ignored), and the proof in the last regime, in which no source is useful because $q_2^2 \gtrsim q_1^2 \gtrsim \sigma_0^2/N$ is based again on the monotonicity of $L_{N,1}(\boldsymbol{\Psi}_{\leq}(\boldsymbol{q}), \boldsymbol{\Gamma})$ in $\boldsymbol{q}$.

Proof Denote for brevity $U := \frac{q_1}{2} + \frac{q_2}{2} + \sigma_2/4\sqrt{N}$. First consider the case $U^2 \leq \sigma_0^2/(4N)$. We construct the following binary hypothesis testing problem between

$$\boldsymbol{\theta}^{(0)} := (\theta_0^{(0)}, \theta_1^{(0)}, \theta_2^{(0)}) = (-U, -U + q_1, -U + q_2)$$

and

$$\boldsymbol{\theta}^{(1)} := (\theta_0^{(1)}, \theta_1^{(1)}, \theta_2^{(1)}) = (U, U - q_2, U - q_1)$$

using N samples from each of the models. This pair of hypotheses is constructed so that: (1) Under each of the hypotheses, there is one source parameter at distance q_1 from the target parameter θ_0 and one at distance q_2 (though the source model with the closer parameter is different for each hypotheses). (2) The parameters $\theta_t^{(0)}$ and $\theta_t^{(1)}$ are indistinguishable using observations from the model $\mathcal{M}_t$. Indeed, using Fact 1, it holds that

$$\mathrm{D}_{\mathrm{KL}}(P_{\boldsymbol{\theta}^{(0)}}, P_{\boldsymbol{\theta}^{(1)}}) = N \cdot \mathrm{D}_{\mathrm{KL}}(P_{\theta_0^{(0)}}, P_{\theta_0^{(1)}}) + N \cdot \mathrm{D}_{\mathrm{KL}}(P_{\theta_1^{(0)}}, P_{\theta_1^{(1)}}) + N \cdot \mathrm{D}_{\mathrm{KL}}(P_{\theta_2^{(0)}}, P_{\theta_2^{(1)}})$$

$$
\begin{aligned}
&= N \cdot \frac{2U^2}{\sigma_0^2} + N\frac{(2U - q_2 - q_1)^2}{2\sigma_1^2} + N\frac{(2U - q_2 - q_1)^2}{2\sigma_2^2} \\
&\leq N \cdot \frac{2U^2}{\sigma_0^2} + N\frac{(2U - q_2 - q_1)^2}{2\sigma_2^2} + N\frac{(2U - q_2 - q_1)^2}{2\sigma_2^2} \\
&\leq \frac{1}{2} + \frac{1}{8} + \frac{1}{8} = \frac{3}{4},
\end{aligned}
$$

where the inequality holds under the assumptions $U^2 \leq \sigma_0^2/(4N)$ and $\sigma_1^2 \geq \sigma_2^2$. By Pinsker's inequality (Wainwright, 2019, Eq. (3.59))

$$
\mathrm{d_{TV}}(P_{\boldsymbol{\theta}^{(0)}}, P_{\boldsymbol{\theta}^{(1)}}) \leq \sqrt{\frac{1}{2}\mathrm{D_{KL}}(P_{\boldsymbol{\theta}^{(0)}}, P_{\boldsymbol{\theta}^{(1)}})} \leq \sqrt{\frac{3}{8}}.
$$

Then, Le-Cam's method implies that

$$
\begin{aligned}
\tilde{L}_{N,1}\left(\boldsymbol{\Psi}_{\leq}(\boldsymbol{q}), \boldsymbol{\Gamma}\right) &\geq \frac{(\theta_0^{(0)} - \theta_0^{(1)})^2}{2} \cdot \left[1 - \mathrm{d_{TV}}(P_{\boldsymbol{\theta}^{(0)}}, P_{\boldsymbol{\theta}^{(1)}})\right] \\
&\geq \frac{4U^2}{2} \cdot \left[1 - \sqrt{\frac{3}{8}}\right] \\
&\geq \frac{30}{45}U^2. \qquad (54)
\end{aligned}
$$

Next, denote for brevity $V := q_1 + \sigma_1/(8\sqrt{N})$ and consider the case $U^2 \geq \sigma_0^2/(4N)$ yet $V^2 \leq \sigma_0^2/(4N)$. Note that these conditions imply that $V \leq U$ and consequently

$$
\frac{\sigma_1}{8\sqrt{N}} \leq q_2. \qquad (55)
$$

We construct a pair of hypotheses which are similar to the $T = 1$ setting for the target model and the first source model, and have *identical* parameter for the second source model, to wit

$$
\boldsymbol{\theta}^{(0)} := (\theta_0^{(0)}, \theta_1^{(0)}, \theta_2^{(0)}) = \frac{1}{3} \cdot (-V, -V + q_1, 0)
$$

and

$$
\boldsymbol{\theta}^{(1)} := (\theta_0^{(1)}, \theta_1^{(1)}, \theta_2^{(1)}) = \frac{1}{3} \cdot (V, V - q_1, 0).
$$

This set of hypotheses is in $\boldsymbol{\Psi}_{\leq}(\boldsymbol{q}, \boldsymbol{\Gamma})$ since

$$
\left|\theta_0^{(0)} - \theta_1^{(0)}\right|^2 = \left|\theta_0^{(1)} - \theta_1^{(1)}\right|^2 = \frac{q_1^2}{9}
$$

and

$$
\left|\theta_0^{(0)} - \theta_2^{(0)}\right|^2 = \left|\theta_0^{(1)} - \theta_2^{(1)}\right|^2 = \frac{1}{9}V^2 = \frac{1}{9}\left(q_1 + \frac{\sigma_1}{8\sqrt{N}}\right)^2 \leq q_2^2,
$$

where the inequality holds from the global assumptions $q_1 \leq q_2$ and (55). Since the second source model $t = 2$ clearly does not contribute to the KL divergence between the two hypotheses, this reduces the problem to the single source model. By scaling the bound in (51) by $1/3$, it holds that

$$
\tilde{L}_{N,1}\left(\boldsymbol{\Psi}_{\leq}(\boldsymbol{q}), \boldsymbol{\Gamma}\right) \geq \frac{1}{9} \cdot \frac{4\left(q_1 + \frac{\sigma_1}{8\sqrt{N}}\right)^2}{5}. \qquad (56)
$$

Next, similarly to the $T = 1$ setting, if $V^2 \geq \sigma_0^2/(4N)$ then the monotonicity of $\tilde{L}_{N,1}(\mathbf{\Psi}_{\leq}(\boldsymbol{q}), \mathbf{\Gamma})$ in $\boldsymbol{q}$ (which holds element-wise) and (56) imply the lower bound

$$\tilde{L}_{N,1}\left(\mathbf{\Psi}_{\leq}(\boldsymbol{q}), \mathbf{\Gamma}\right) \geq \frac{1}{45} \cdot \frac{\sigma_0^2}{N}.$$

From the above three cases we deduce that

$$\tilde{L}_{N,1}\left(\mathbf{\Psi}_{\leq}(\boldsymbol{q}), \mathbf{\Gamma}\right) \geq \begin{cases} \frac{30}{45} U^2, & U^2 \leq \frac{1}{4}\frac{\sigma_0^2}{N} \\ \frac{4}{45} \cdot V^2, & U^2 \geq \frac{1}{4}\frac{\sigma_0^2}{N} \geq V^2. \\ \frac{1}{45} \cdot \frac{\sigma_0^2}{N}, & V^2 \geq \frac{1}{4}\frac{\sigma_0^2}{N} \end{cases} \tag{57}$$

We next show that this lower bound matches the weak oracle rate. Consider the following cases:

1. $q_1^2 \leq q_2^2 \leq \sigma_0^2/(16N)$. It then holds that

$$U^2 = \left(\frac{q_1}{2} + \frac{q_2}{2} + \frac{\sigma_2}{4\sqrt{N}}\right)^2 \leq \left(q_2 + \frac{\sigma_0}{4\sqrt{N}}\right)^2 \leq 2q_2^2 + \frac{1}{8}\frac{\sigma_0^2}{N} \leq \frac{1}{4}\frac{\sigma_0^2}{N},$$

   and so (57) implies that

$$\begin{aligned} \tilde{L}_{N,1}\left(\mathbf{\Psi}_{\leq}(\boldsymbol{q}), \mathbf{\Gamma}\right) \geq \frac{30}{45} U^2 = \frac{30}{45}\left(\frac{q_1}{2} + \frac{q_2}{2} + \frac{\sigma_2}{4\sqrt{N}}\right)^2 \geq \frac{30}{45}\left(\frac{q_2^2}{4} + \frac{1}{16}\frac{\sigma_2^2}{N}\right) \\ \geq \frac{1}{24} \cdot \left(q_2^2 + \frac{\sigma_2^2}{N}\right). \end{aligned}$$

2. $q_1^2 \leq \sigma_0^2/(16N)$ and $\sigma_0^2/N \leq q_2^2$. It then holds that

$$U^2 = \left(\frac{q_1}{2} + \frac{q_2}{2} + \frac{\sigma_2}{4\sqrt{N}}\right)^2 \geq \frac{q_2^2}{4} \geq \frac{1}{4}\frac{\sigma_0^2}{N},$$

   and

$$V^2 = \left(q_1 + \frac{\sigma_1}{8\sqrt{N}}\right)^2 \leq 2q_1^2 + \frac{1}{32}\frac{\sigma_1^2}{N} \leq \frac{1}{4}\frac{\sigma_0^2}{N}$$

   and so (57) implies that

$$\tilde{L}_{N,1}\left(\mathbf{\Psi}_{\leq}(\boldsymbol{q}), \mathbf{\Gamma}\right) \geq \frac{4}{45} \cdot V^2 = \frac{4}{45} \cdot \left(q_1 + \frac{\sigma_1}{8\sqrt{N}}\right)^2 \geq \frac{1}{720} \cdot \left(q_1^2 + \frac{\sigma_1^2}{N}\right).$$

3. $q_2^2 \geq q_1^2 \geq \sigma_0^2/(4N)$. It then holds that

$$V^2 = \left(q_1 + \frac{\sigma_1}{8\sqrt{N}}\right)^2 \geq q_1^2 + \frac{1}{64}\frac{\sigma_1^2}{N} \geq \frac{1}{4}\frac{\sigma_0^2}{N}$$

   and so (57) implies that

$$\tilde{L}_{N,1}\left(\mathbf{\Psi}_{\leq}(\boldsymbol{q}), \mathbf{\Gamma}\right) \geq \frac{1}{45} \cdot \frac{\sigma_0^2}{N}.$$

Combining all the above, we obtain

$$\tilde{L}_{N,1}\left(\boldsymbol{\Psi}_{\leq}(\boldsymbol{q}),\boldsymbol{\Gamma}\right) \geq \begin{cases} \frac{1}{24}\cdot\left(q_2^2+\frac{\sigma_2^2}{N}\right), & q_1^2 \leq q_2^2 \leq \frac{1}{16}\frac{\sigma_0^2}{N} \\ \frac{1}{720}\cdot\left(q_1^2+\frac{\sigma_1^2}{N}\right), & q_1^2 \leq \frac{1}{16}\frac{\sigma_0^2}{N} \leq \frac{\sigma_0^2}{N} \leq q_2^2 \\ \frac{1}{45}\cdot\frac{\sigma_0^2}{N}, & q_2^2 \geq q_1^2 \geq \frac{1}{4}\frac{\sigma_0^2}{N} \end{cases}.$$

By the monotonicity property of $\tilde{L}_{N,1}(\boldsymbol{\Psi}_{\leq}(\boldsymbol{q}),\boldsymbol{\Gamma})$ in $\boldsymbol{q}$, the same bound holds for smaller values of $q_1$and q_2, that is, the bound in (52) holds. The bound in (53) then directly follows. ■

Multiple source models ($T \geq 3$) We next move to the proof of Theorem 5 for $T \geq 3$. The proof will exploit the monotonicity property of the semi-local minimax risk, but before providing it, we explain why the construction used in the $T = 2$ fails for $T = 3$ and above. To this end, let us focus on the interesting regime of $q_3^2 \lesssim \sigma_0^2/N$, which implies that all 3 source models are better than the target model. In the $T = 2$ case, each of the two constructed hypotheses satisfied that one of the source parameters at exactly q_1 distance from θ_0 and the other at exactly q_2 distance from θ_0. In Figure 7, we show a failed attempt for such a construction. In this construction, sufficiently large distances are first assured for the extremal source models $\mathcal{M}_1$ and $\mathcal{M}_3$, and any tester of the two hypotheses based on N samples from each model cannot decide between $\theta_1^{(0)}$ *vs.* $\theta_1^{(1)}$ or between $\theta_3^{(0)}$ *vs.* $\theta_3^{(1)}$ w.h.p. As in the $T = 1$ or $T = 2$ settings, this already assures suitable distance between the parameters of the target model $\mathcal{M}_0$. The problematic part of this construction is that after fixing these distances, there are not enough "degrees-of-freedom" to adjust the distances between the parameters of the "middle" source model $\mathcal{M}_2$. Indeed, referring to Figure 7, $\xi := |\theta_2^{(0)} - \theta_2^{(1)}| \lesssim \sigma_2/\sqrt{N}$ will hold only if

$$q_2 + \frac{\sigma_2}{\sqrt{N}} \gtrsim q_3 + \frac{\sigma_3}{\sqrt{N}}.$$

Since $q_2 \leq q_3$ and $\sigma_2^2 \geq \sigma_3^2$, this holds only if and only if $\sigma_2/\sqrt{N} \gtrsim q_3$, which does not necessarily hold. If $\theta_2^{(0)}$ and $\theta_2^{(1)}$ are too close to one another, then using the N samples from the models $\mathcal{M}_2$ may allow the tester to reliably decide on the hypothesis, and no non-trivial lower bound can be established by the reduction to this binary hypothesis test. We remark that considering multidimensional parameters ($d > 1$) will not allow to ameliorate this problem as it stems from constraint on Euclidean distances, which must hold for any dimension.

Proof [of Theorem 5 for $T \geq 2$] The proof is based on a reduction to the $T = 2$ case. We first assume that $d = 1$. Consider the following subset of $\boldsymbol{\Psi}_{\leq}(\boldsymbol{q})$ with the additional constraints that

$$\theta_1 = \theta_2 \cdots = \theta_{t_{\text{med}}-1} =: \tilde{\theta}_1$$

and

$$\theta_{t_{\text{med}}} = \theta_{t_{\text{med}}+1} = \cdots = \theta_T =: \tilde{\theta}_2,$$

where $\|\tilde{\theta}_1 - \theta_0\|^2 \leq q_1^2 := \tilde{q}_{\zeta(1)}^2$ and $\|\tilde{\theta}_2 - \theta_0\|^2 \leq q_{\text{med}}^2 := \tilde{q}_{\zeta(2)}^2$. Given this restricted subset, and since $\sigma_0^2 \geq \sigma_1^2 \geq \cdots \geq \sigma_T^2$, a simple data-processing argument (Claim 2) implies that an optimal CL tester may sample just from the models $\mathcal{M}_{t_{\text{med}}-1}$ and $\mathcal{M}_T$. Thus, the problem is reduced to a $\tilde{T} = 2$ source model problem with distances $\tilde{\boldsymbol{q}} = (\tilde{q}_1, \tilde{q}_2)$, where $\tilde{q}_1 = q_1$, and $\tilde{q}_2 = q_{\text{med}}$, for which it holds that $\tilde{q}_1^2 \leq \tilde{q}_2^2 \leq \sigma_0^2/(4N)$, and noise variances $\tilde{\boldsymbol{\Gamma}} := (\sigma_0^2, \tilde{\sigma}_1^2, \tilde{\sigma}_2^2)$, for which it holds

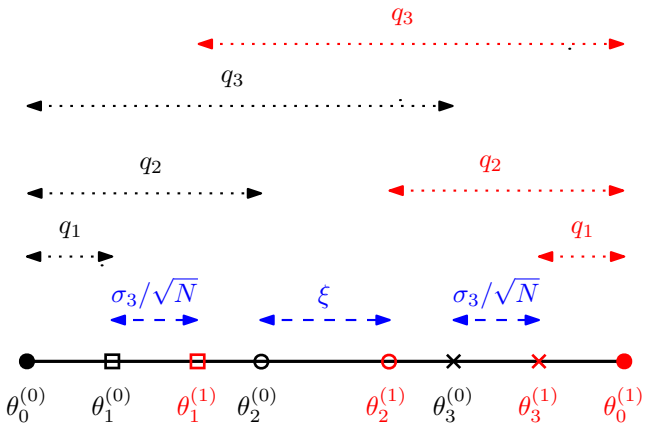

Figure 7: A failed attempt for a proper construction of parameter hypotheses for $T = 3$. One of the hypotheses is in black and the other in red. The target parameter is designated by a disc, the first source parameter by a square, the second by a circle, and the third by a cross.

that $\tilde{\sigma}_1^2 := \sigma_{t_{\text{med}}-1}^2$ and $\tilde{\sigma}_2^2 := \sigma_T^2$. Restricting the set $\mathbf{\Psi}_{\leq}(\boldsymbol{q})$ to such parameters only reduces the risk, and so

$$\tilde{L}_{N,1}(\mathbf{\Psi}_{\leq}(\boldsymbol{q}), \mathbf{\Gamma}) \geq \tilde{L}_{N,1}(\mathbf{\Psi}_{\leq}(\tilde{\boldsymbol{q}}), \tilde{\mathbf{\Gamma}}) \overset{(*)}{\geq} \frac{1}{720} \cdot \left(\tilde{q}_2^2 + \frac{\tilde{\sigma}_2^2}{N} \right) = \frac{1}{720} \cdot \left(q_{\text{med}}^2 + \frac{\sigma_T^2}{N} \right),$$

where $(*)$ follows from Proposition 16. This implies the statement of the theorem for $d = 1$. For any constant $d \geq 1$, we may assume that learner knows that all parameters belongs to a subset of $\mathbf{\Psi}_{\leq}(\boldsymbol{q})$, for which the last $d - 1$ coordinates are identically zero. An optimal learner may ignore these coordinates. The only modification required compared to the $d = 1$ case is since $\mathcal{T}_{\text{w.o.}}(\kappa)$ is defined as the set of source parameters for which $Q_t^2 \leq \kappa \cdot d\sigma_0^2/N$, the value of κ should be set so that $\kappa \cdot d = 1/4$ in order for the previous derivation to hold. ■

Remark 17 *As discussed, if $q_{t_{w.o.}+1} \geq \frac{1}{2}(q_1 + q_{med} + \sigma_{w.o.}/8\sqrt{N})$ then the dependency on the noise variance can be improved from σ_T^2 to $\sigma_{w.o.}^2$. This is obtained by modifying the proof as follows. First, we let*

$$\theta_{t_{med}} = \theta_{t_{med}+1} = \cdots = \theta_{t_{w.o.}} =: \tilde{\theta}_2,$$

and then add another possible point

$$\theta_{t_{w.o.}+1} = \cdots = \theta_T =: \tilde{\theta}_3.$$

We then consider a $\tilde{T} = 3$ setting, but construct an hypothesis testing problem between $\boldsymbol{\theta}^{(0)}$ and $\boldsymbol{\theta}^{(1)}$ as in $T = 2$, as in Figure 6, where $\tilde{\theta}_1$ and $\tilde{\theta}_2$ are in the role of θ_1 and θ_2. The third source parameter is placed at 0 for both hypotheses, and thus is useless for the distinction between them. We omit the details for brevity.

J.4. Proof of Theorem 6

We next prove Theorem 6, which is a tight lower bound for the general d-dimensional case and arbitrary T sources. This requires a packing set of more than a pair of points (exponential in d), and the use of Fano's method.

Proof outline The proof of Theorem 6 utilizes Fano's method, which is based on a packing set in $\mathbf{\Psi}$ whose number of points is exponential in d. The proof begins by a construction of such a packing set for a single parameter in $\mathbb{R}^d$ (Lemma 18). Points in this packing set belong to the unit sphere, and in addition, have a zero first coordinate. The proof then splits between the two different regimes $d\sigma^2_{\text{w.o.}}/N \leq q^2_{\text{w.o.}}$ and $q^2_{\text{w.o.}} \leq d\sigma^2_{\text{w.o.}}/N$. For the former, Lemma 20 uses the single-source packing set mentioned above to construct a specific testing set $\mathbf{\Theta}^*_{\text{test}}$ with appropriate distances between the parameters of the target and source models, as well as appropriate separation between the models (Lemma 20). This specific construction is tailored so that the optimal policy π is non-adaptive and uses just one source model, as shown in Lemma 21, which also invokes Fano's method for this construction. In fact, in this regime, the learner knows in advance that parameters of $\mathcal{M}_t$ are at distance q_t and have noise variance σ_t^2. Proposition 22 then completes the proof of the lower bound for this regime by standard bounding of the mutual information. The latter regime is then analyzed, for which it is shown that the nature of the set $\mathbf{\Psi}_*(q_{\text{w.o.}})$ allows to reduce the problem to the standard case, in which only the target model is sampled (using Claim 2). Fano's method is then similarly invoked to establish the lower bound.

Our construction of a testing set will be based on a proper construction of a set for a single task. Specifically, in standard single-task estimation the constructed set is a scaled version of a packing set of the Euclidean ball (Wainwright, 2019, Example 15.14). Here, we construct a slightly modified set, which is a packing set of the Euclidean *sphere,* and in addition, whose first coordinate is zero. The cardinality of this set is essentially similar to the standard packing set of the Euclidean ball, besides a worse numerical constant.

Lemma 18 *Assume that $d \geq 3$. Then, there exists a set $\{v^{(j)}\}_{j\in[K]} \subset \mathbb{R}^d$ such that: (1) Unit norm $\|v^{(j)}\|^2 = 1$. (2) The first coordinate of each $v^{(j)}$ is zero. (3) The points in the set are separated*

$$\|v^{(j_1)} - v^{(j_2)}\| \geq e^{-\frac{d+2}{d-2}} := c(d) \geq e^{-5}$$

for any $j_1 \neq j_2 \in [K]$. (4) Cardinality $K \geq 4 \cdot e^d$.

Proof Let $\mathbb{B}^d \equiv \mathbb{B}^d(1)$ be the unit ball in $\mathbb{R}^d$. From (Wainwright, 2019, Lemmas 5.5 and 5.7), for any $\epsilon \in (0,1)$ there exists a packing set $\{\tilde{v}^{(j)}\}_{j\in[K(\epsilon)]} \subset \mathbb{B}^{d-2}$ such that $\|\tilde{v}^{(j_1)} - \tilde{v}^{(j_2)}\|^2 \geq \epsilon$, and the cardinality satisfies $K(\epsilon) \geq 1/\epsilon^{d-2}$. Choosing $\epsilon = e^{-(d+2)/(d-2)}$, the set $\{\tilde{v}^{(j)}\}_{j\in[K]} \subset \mathbb{R}^d$ with $v^{(j)} = (0, \tilde{v}^{(j)}, \sqrt{1 - \|\tilde{v}^{(j)}\|})$ satisfies the first three required properties, and its cardinality satisfies $K \geq e^{((d+2)/(d-2))\cdot(d-2)} \geq 4 \cdot e^d$ for any $d \geq 3$. ■

Remark 19 *The constant $c(d) = e^{-5}$ for $d = 3$, however, improves with increasing dimension,* e.g., *$c(d) = e^{-3/2} \geq 0.22$ if we restrict to $d \geq 10$, whereas for the $d \to \infty$ constant is $e^{-1} \geq 0.36$.*

We next prove Theorem 6 separately in the two different regimes $d\sigma^2_{\text{w.o.}}/N \leq q^2_{\text{w.o.}}$ and $q^2_{\text{w.o.}} \leq d\sigma^2_{\text{w.o.}}/N$ in Propositions 22 and 23. We begin with the former regime. Consider the following construction of a testing set $\mathbf{\Theta}^*_{\text{test}} := \{\theta^{(*,j)}\}_{j\in[K]} \subset \mathbf{\Theta} = \mathbb{R}^d$ for $d \geq 3$. Let $\{v^{(j)}\}_{j\in[K]}$ be a set which satisfies the four properties stated in Lemma 18. Then $\mathbf{\Theta}^*_{\text{test}}$ is constructed as follows:

1. Parameters for the target and "close" models: For any $t \in \{0\} \cup [t_{\text{w.o.}}]$

$$\theta_t^{(*,j)} = \left(\sqrt{\frac{d\sigma^2_{\text{w.o.}}}{N}} - q_t \right) \cdot v^{(j)}.$$

2. Parameters for the "far" models: For any $t \in [T] \backslash [t_{\text{w.o.}}]$

$$\theta_t^{(*,j)} = \left(\sqrt{q_t^2 - \frac{d\sigma^2_{\text{w.o.}}}{N}}, 0, \ldots, 0 \right)$$

for any $j \in [K]$.

Lemma 20 *Assume that $d \geq 3$. Then, the set $\mathbf{\Theta}^*_{test} := \{\theta^{(*,j)}\}_{j\in[K]} \subset \mathbf{\Theta} = \mathbb{R}^d$ satisfies:*

1. *Distance between the task parameters:*

$$\|\theta_0^{(*,j)} - \theta_t^{(*,j)}\| = q_t^2 \tag{58}$$

for any $t \in [T]$ and any $j \in [K]$.

2. *Separation in the target task:*

$$\eta_* \equiv \eta(\mathbf{\Theta}^*_{test}) = \|\theta_0^{(*,j_1)} - \theta_0^{(*,j_2)}\| \geq c(d) \cdot \sqrt{\frac{d\sigma^2_{w.o.}}{N}} \tag{59}$$

for any $j_1, j_2 \in [K]$, $j_1 \neq j_2$.

Proof We first verify the distance between the model parameters. The key point is to assure that the magnitudes $\sqrt{d\sigma^2_{\text{w.o.}}/N} - q_t$ for close models and the magnitudes $\sqrt{q_t^2 - d\sigma^2_{\text{w.o.}}/N}$ used to the define the parameters are all (nonnegative) real numbers. Indeed, for the close models, $t \in [t_{\text{w.o.}}]$, it holds that $q_t^2 \leq q_{t_{\text{w.o.}}} \leq d\sigma^2_{\text{w.o.}}/N$ and then $\sqrt{d\sigma^2_{\text{w.o.}}/N} - q_t \geq 0$. For the far models, $t \in [T]\backslash[t_{\text{w.o.}}]$, it holds that $q_t^2 \geq d\sigma_0^2/N \geq d\sigma^2_{\text{w.o.}}/N$ and then $q_t^2 - d\sigma^2_{\text{w.o.}}/N \geq 0$. Thus, in both cases and for all models, the parameter set $\{\theta_t^{(*,j)}\}_{j\in[K]}$ is well defined. It is then easy to verify that (58) holds. The separation in the target task follows directly from Lemma 18 and $q_0 = 0$. ■

The next proposition is crucial, and is based on Claim 2 invoked for the constructed testing set $\mathbf{\Theta}^*_{\text{test}}$. In essence, it shows that the optimal policy π of a tester for a randomly chosen parameter in $\mathbf{\Theta}^*_{\text{test}}$ is to take all N samples from a single, optimally chosen, model. This reduces the CL setting to a standard single-task setting, and enables the utilization of Fano's method (Wainwright, 2019, Section 15.3) to lower bound the error probability, and thus also the risk. It should be noted that the learner knows in advance that parameters of $\mathcal{M}_t$ are at distance $Q_t = q_t$ and have noise variance σ_t^2.

Lemma 21 *An optimal CL tester for testing a uniform random hypothesis from $\mathbf{\Theta}^*_{test} := \{\theta^{(*,j)}\}_{j\in[K]}$ may collect all its N samples from a single model $\mathcal{M}_{t^*}$, where $t^* \in [t_{w.o.}]$. Given that $J = j$, the N resulting samples are equal in distribution to N i.i.d. samples from the model*

$$\mathcal{M}_* : Y_* = \theta_0^{(*,j)} + \epsilon_* \tag{60}$$

where $\epsilon_* \sim \mathcal{N}(0, \sigma_*^2 \cdot I_d)$*, and where* $\sigma_*^2 \geq \sigma_{w.o.}^2$ *holds. Furthermore, let* $Y_*^{\otimes N}$ *be* N *i.i.d. measurements from* (60)*. Then,*

$$\min_{\mathcal{A}} e_N(\mathbf{\Phi}_{test}, \mathcal{A}) \geq 1 - \frac{I(J; Y_*^{\otimes N}) + \log 2}{\log K}. \tag{61}$$

Proof For far models, $t \in [T] \backslash [t_{\text{w.o.}}]$, the parameters $\theta_t^{(*,j)} = (\sqrt{q_t^2 - \lambda_{\text{w.o.}}^2}, 0, \ldots, 0)$ are located at the same position for all $j \in [K]$ hypotheses. Thus, the first case of Claim 2 implies that an optimal learner will not sample from $\mathcal{M}_t$ for $t \in [T] \backslash [t_{\text{w.o.}}]$. For close models, we exploit the property that $\theta_{t_1}^{(*,j)} \propto \theta_{t_2}^{(*,j)}$ for any $t_1, t_2 \in \{0\} \cup [t_{\text{w.o.}}]$, and so the second case of Claim 2 implies that an optimal learner should only sample from one of them. Specifically, the one with minimal noise variance after scaling. Consider a model $t \in \{0\} \cup [t_{\text{w.o.}}]$. Since the learner knows the norm $\|\theta_t^{(*,j)}\| = \lambda_{\text{w.o.}} - q_t$ in advance, it can scale its samples by $\frac{\sqrt{d\sigma_{\text{w.o.}}^2/N}}{\sqrt{d\sigma_{\text{w.o.}}^2/N} - q_t} > 1$ to obtain an equivalent model

$$\tilde{\mathcal{M}}_t : \tilde{Y}_t = \theta_0^{(*,j)} + \tilde{\epsilon}_t$$

where $\tilde{\theta}_t^{(*,j)} = \theta_0^{(*,j)}$ and $\tilde{\epsilon}_t \sim \mathcal{N}(0, \tilde{\sigma}_t^2 \cdot I_d)$ with

$$\tilde{\sigma}_t^2 = \left(\frac{\sqrt{\frac{d\sigma_{\text{w.o.}}^2}{N}}}{\sqrt{\frac{d\sigma_{\text{w.o.}}^2}{N}} - q_t} \right)^2 \cdot \sigma_t^2.$$

This holds for all models $t \in \{0\} \cup [t_{\text{w.o.}}]$, and since the parameters of the different hypotheses are located at exactly the same positions $\{\theta_0^{(*,j)}\}_{j \in [K]}$, the learner can sample only from the model with the minimal (scaled) noise variance, that is, from

$$t^* := \operatorname*{argmin}_{t \in \{0\} \cup [t_{\text{w.o.}}]} \left(\frac{\sqrt{\frac{d\sigma_{\text{w.o.}}^2}{N}}}{\sqrt{\frac{d\sigma_{\text{w.o.}}^2}{N}} - q_t} \right)^2 \cdot \sigma_t^2.$$

We thus let $\mathcal{M}_* = \tilde{\mathcal{M}}_{t_*}$ and $\sigma_*^2 = \tilde{\sigma}_{t^*}^2$. It now holds that

$$\tilde{\sigma}_{t^*}^2 = \min_{t \in \{0\} \cup [t_{\text{w.o.}}]} \left(\frac{\sqrt{\frac{d\sigma_{\text{w.o.}}^2}{N}}}{\sqrt{\frac{d\sigma_{\text{w.o.}}^2}{N}} - q_t} \right)^2 \cdot \sigma_t^2 \geq \sigma_{\text{w.o.}}^2,$$

since $\sqrt{d\sigma_{\text{w.o.}}^2/N} - q_t > 0$ and due to the noise variance ordering assumption $\sigma_0^2 \geq \sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_T^2$. Given that an optimal CL tester may collect N samples from $\mathcal{M}_*$, the problem is now reduced to the standard hypothesis testing setting. By Fano's inequality (Wainwright, 2019, Eq. (15.31)), the error probability in testing J based on N i.i.d. samples from $\mathcal{M}_*$ is lower bounded by $1 - {}^{(I(J;Y_*^{\otimes N}) + \log 2)}/_{\log K}$, from which (61) directly follows. ■

The rest of the proof in this case is a technical standard bounding of the mutual information, leading to:

Proposition 22 *Consider the setting of Theorem 6 and further assume that* $d\sigma_{\text{w.o.}}^2/N \geq q_{\text{w.o.}}^2$. *Then,*

$$L_{N,d}(\mathbf{\Psi}_{=}(\boldsymbol{q}), \mathbf{\Gamma}) \geq \frac{c^2(d)}{2} \cdot \frac{d\sigma_{\text{w.o.}}^2}{N}.$$

Proof Let $P^{(*,j)}$ be distribution of a sample from the Gaussian location model $\mathcal{M}_*$ in (60), and let $\overline{P}$ be distribution of N i.i.d. samples from a pure noise model $\mathcal{N}(0, \sigma_*^2 \cdot I_d)$. The information-radius upper bound on the mutual information and the tensorization property of the KL divergence imply that

$$I(J; Y_*^{\otimes N}) \leq N \cdot \max_{j \in [K]} \mathrm{D}_{\mathrm{KL}}(P^{(*,j)} \mid\mid \overline{P}) = N \cdot \max_{j \in [K]} \frac{\|\theta^{(*,j)}\|^2}{2\sigma_*^2},$$

where the equality follows from Fact 1. The lower bound (61) then reads

$$\begin{aligned}
1 - \frac{I(J; Y_*^{\otimes N}) + \log 2}{\log K} &\geq 1 - \frac{N \frac{\|\theta^{(*,j)}\|^2}{2\sigma_*^2} + \log 2}{\log K} \\
&\overset{(a)}{\geq} 1 - \frac{\frac{d\sigma_{\text{w.o.}}^2}{2\sigma_*^2} + \log 2}{\log K} \\
&\overset{(b)}{\geq} 1 - \frac{\frac{d}{2} + \log 2}{\log K} \\
&\overset{(c)}{\geq} \frac{1}{2},
\end{aligned}$$

where (a) follows since $\|\theta^{(*,j)}\|^2 = d\sigma_{\text{w.o.}}^2/N$, (b) follows since Lemma 21 states that $\sigma_*^2 \geq \sigma_{\text{w.o.}}^2$, and (c) follows since Lemma 18 states that $K \geq 4 \cdot e^d$. Combining this with (61) (Lemma 21) and (49) (Proposition 14) implies that $L_{N,d}(\mathbf{\Psi}_{=}(\boldsymbol{q}), \mathbf{\Gamma}) \geq \eta_*^2/2$. The proof is completed by using separation in the target task bound (59) (Lemma 20). ■

Proposition 23 *Consider the setting of Theorem 6 and further assume that* $q_{\text{w.o.}}^2 \geq d\sigma_{\text{w.o.}}^2/N$. *Then,*

$$L_{N,d}(\mathbf{\Psi}_*(q_{\text{w.o.}}), \mathbf{\Gamma}) \geq \frac{c^2(d)}{8} \cdot q_{\text{w.o.}}^2.$$

Proof We construct the testing set $\mathbf{\Theta}_{\text{test}}^* := \{\theta^{(*,j)}\}_{j \in [K]}$ in which $\theta_0^{(*,j)} = q_{\text{w.o.}} \cdot v^{(j)}$ and $\theta_t^{(*,j)} = 0$ for all $t \in [T]$. Similarly to Lemma 20 it holds that $\mathbf{\Theta}_{\text{test}}^* \subset \mathbf{\Psi}_*(q_{\text{w.o.}})$ and that

$$\|\theta_0^{(*,j_1)} - \theta_0^{(*,j_2)}\| \geq q_{\text{w.o.}} \cdot c(d)$$

for all $j_1 \neq j_2 \in [K]$. By Claim 2, an optimal learner will collect N samples from $\mathcal{M}_0$. This reduces the problem to the standard estimation setting. Thus, Fano's inequality (Wainwright, 2019, Eq. (15.31)) implies that

$$\min_{\mathcal{A}} e_N(\mathbf{\Theta}_{\text{test}}^*, \mathcal{A}) \geq 1 - \frac{I(J; Y_0^{\otimes N}) + \log 2}{\log K},$$

where $Y_0^{\otimes N}$ are N i.i.d. samples from $\mathcal{M}_0$. Similarly to the proof of Proposition 22, this is further lower bounded as

$$\begin{aligned}
1 - \frac{I(J; Y_0^{\otimes N}) + \log 2}{\log K} &\geq 1 - \frac{N \frac{\|\theta^{(*,j)}\|^2}{2\sigma_0^2} + \log 2}{\log K} \\
&= 1 - \frac{N \frac{q_{\text{w.o.}}^2}{2\sigma_0^2} + \log 2}{\log K} \\
&\overset{(a)}{\geq} 1 - \frac{\frac{d}{2} + \log 2}{\log K} \\
&\overset{(b)}{\geq} 1 - \frac{\frac{d}{2} + \log 2}{\log K} \\
&\overset{(c)}{\geq} \frac{1}{2},
\end{aligned}$$

where (a) holds since $q_{\text{w.o.}}^2 \leq {}^{d\sigma_0^2}/_N$, and (b) since $K \geq 4 \cdot e^d$ (Lemma 18). The proof then follows from Proposition 14. ■