Exact Mean Square Linear Stability Analysis for SGD

Rotem Mulayoff ROTEM.MULAYOF@GMAIL.COM
Technion – Israel Institute of Technology

Tomer Michaeli TOMER.M@EE.TECHNION.AC.IL
Technion – Israel Institute of Technology

Editors: Shipra Agrawal and Aaron Roth

Abstract

The dynamical stability of optimization methods at the vicinity of minima of the loss has recently attracted significant attention. For gradient descent (GD), stable convergence is possible only to minima that are sufficiently flat w.r.t. the step size, and those have been linked with favorable properties of the trained model. However, while the stability threshold of GD is well-known, to date, no explicit expression has been derived for the exact threshold of stochastic GD (SGD). In this paper, we derive such a closed-form expression. Specifically, we provide an explicit condition on the step size that is both necessary and sufficient for the linear stability of SGD in the mean square sense. Our analysis sheds light on the precise role of the batch size B. In particular, we show that the stability threshold is monotonically non-decreasing in the batch size, which means that reducing the batch size can only decrease stability. Furthermore, we show that SGD's stability threshold is equivalent to that of a mixture process which takes in each iteration a full batch gradient step w.p. $1-p$, and a single sample gradient step w.p. p, where $p \approx 1/B$. This indicates that even with moderate batch sizes, SGD's stability threshold is very close to that of GD's. We also prove simple necessary conditions for linear stability, which depend on the batch size, and are easier to compute than the precise threshold. Finally, we derive the asymptotic covariance of the dynamics around the minimum, and discuss its dependence on the learning rate. We validate our theoretical findings through experiments on the MNIST dataset.

Keywords: SGD, Dynamical systems, Linear stability, Mean square analysis

1. Introduction

The dynamical stability of optimization methods has been shown to play a key role in shaping the properties of trained models. For instance, gradient descent (GD) can stably converge only to minima that are sufficiently flat with respect to the step size (Cohen et al., 2021), and in the context of neural networks, such minima were shown to correspond to models with favorable properties. These include smoothness of the predictor function (Ma and Ying, 2021; Nacson et al., 2023; Mulayoff et al., 2021), balancedness of the layers (Mulayoff and Michaeli, 2020), and arguably better generalization (Hochreiter and Schmidhuber, 1997; Keskar et al., 2016; Jastrzębski et al., 2017; Wu et al., 2017; Ma and Ying, 2021). While the stability threshold of GD is well-known, that of stochastic GD (SGD) has yet to be fully understood. Several empirical works studied SGD's stability (Jastrzębski et al., 2019, 2020; Cohen et al., 2021; Gilmer et al., 2022), yet they did not determine a definitive stability condition. Various theoretical works studied SGD's dynamics using linear stability, *i.e.,* via second-order Taylor expansion at the vicinity of minima, focusing on stability in the mean square sense (Wu et al., 2018; Granziol et al., 2022; Velikanov et al., 2023), higher moments (Ma and Ying, 2021), and in probability (Ziyin et al., 2023). However, these works either do not provide explicit stability conditions or rely on strong assumptions. For example, Wu et al. (2018); Ma and Ying (2021) present the condition as a complex optimization problem. Similarly, Granziol et al. (2022) consider infinite network widths and make strong assumptions on the nature of the batching noise. Likewise, Velikanov et al. (2023) analyze SGD with momentum, assuming momentum parameter close to 1 and "spectrally expressible" dynamics, and present their result in terms of a moment generating function. Overall, the exact stability threshold of SGD in the general case remains unknown.

In this paper, we analyze the linear stability of SGD in the mean square sense. We start by considering interpolating minima, which are common in training of overparametrized models. In this case, we provide an explicit threshold on the step size η that is both necessary and sufficient for stability. Our analysis sheds light on the precise role of the batch size B. In particular, we show that the maximal step size allowing stable convergence is monotonically non-decreasing in the batch size. Namely, decreasing the batch size can only decrease the stability threshold of SGD. Moreover, we show that this threshold is equivalent to that of a mixture process that takes in each iteration a full batch gradient step w.p. $1 - p$, and a single sample gradient step w.p. p, where $p \approx 1/B$. This suggests that even with moderate batch sizes, SGD's stability threshold is very close to that of GD's. Although our result gives an explicit condition on the step size for stability, its computation may still be challenging in practical applications. Thus, we also prove simple necessary criteria for stability, which depend on the batch size and are easier to compute.

Next, we turn to study a broader class of minima which we call *regular*. Specifically, in interpolating minima, the loss of each individual sample has zero gradient and a positive semi-definite (PSD) Hessian. In regular minima, the individual Hessians are still required to be PSD, but the gradients can be arbitrary. Only the average of the gradients over all samples has to vanish (as in any minimum). In this setting, the dynamics can wander within the null-space of the Hessian, if the gradients have nonzero components in that subspace. However, the interesting question is whether the process is stable within the orthogonal complement of the null space. Here we again provide an explicit condition on the step size that is both necessary and sufficient for linear stability. We further derive the theoretical limit of the covariance matrix of the dynamics, as well as the limit values of the expected squared distance to the minimum, the expected loss, and the expected squared norm of the gradient, and show how they all decrease when reducing the learning rate. This provides a theoretical explanation of the behavior encountered in common learning rate scheduling strategies.

Finally, we validate our theoretical results through experiments on the MNIST dataset (LeCun, 1998). These confirm that our theory correctly predicts the stability threshold of SGD, and its dependence on the batch size. Furthermore, the experiments suggest that SGD converges at the edge of its (mean-square) stability region at least in certain training regimes, which is an interesting subject for future research.

Contributions. To summarize, our main contributions are as follows.

- We derive a closed-form expression for the maximal step size with which mini-batch SGD is linearly stable in the mean square sense. The threshold depends on the batch size (Thm. 5).
- We prove that the stability threshold is monotonically non-decreasing with the batch size, so that smaller batches can only compromise stability (Prop. 6).
- We show that the stability threshold of mini-batch SGD is the same as the stability threshold of an algorithm that randomly chooses in each iteration whether to perform a GD step or a single-sample SGD step, where the probability is roughly one over the batch size (Prop. 7).
- We determine a lower bound on the batch size such that the stability threshold of SGD is close to that of GD (Prop. 8).

- We provide simpler necessary conditions for the linear stability of mini-batch SGD (Prop. 9).
- Apart from the common setting of interpolating minima, we also study a large family of non-interpolating minima. For interpolating minima, SGD converges below the stability threshold. In contrast, for non-interpolating minima, SGD randomly wanders around the minimum. Here we again derive a closed-form for the stability threshold, as well as an expression for the covariance matrix of the dynamics (Thm. 11 and Thm. 12).
- Key to our derivations is a fundamental algebraic result, which we prove for sums of Kroneker products of symmetric matrices (Thm. 14).

2. Background: Linearized dynamics

Let $\ell_i : \mathbb{R}^d \to \mathbb{R}$ be differentiable almost everywhere for all $i \in [n]$. We consider the minimization of a loss function

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(\boldsymbol{\theta}) \tag{1}$$

using the SGD iterations

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \nabla \hat{\mathcal{L}}_t(\boldsymbol{\theta}_t). \tag{2}$$

Here, η is the step size and $\hat{\mathcal{L}}_t$ is a stochastic approximation of $\mathcal{L}$ obtained as

$$\hat{\mathcal{L}}_t(\boldsymbol{\theta}) = \frac{1}{B}\sum_{i\in\mathfrak{B}_t} \ell_i(\boldsymbol{\theta}), \tag{3}$$

where $\mathfrak{B}_t$ is a batch of size B sampled at iteration t. We assume that the batches $\{\mathfrak{B}_t\}$ are drawn uniformly at random from the $\binom{n}{B}$ possible options, independently across iterations. Namely, there are distinct samples within each batch and possible repetitions between different batches.

Analyzing the full dynamics of this process is intractable in most cases. Yet near minima, accurate characterization of the stability of the iterates can be obtained via linearization (Wu et al., 2018; Ma and Ying, 2021; Mulayoff et al., 2021), as is common in the analysis of nonlinear systems.

Definition 1 (Linearized dynamics) *Let $\boldsymbol{\theta}^*$ be a twice differentiable minimum of $\mathcal{L}$, and denote*

$$\boldsymbol{g}_i \triangleq \nabla \ell_i(\boldsymbol{\theta}^*), \qquad \boldsymbol{H}_i \triangleq \nabla^2 \ell_i(\boldsymbol{\theta}^*). \tag{4}$$

Then the linearized dynamics of SGD near $\boldsymbol{\theta}^*$ is given by

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \frac{\eta}{B}\sum_{i\in\mathfrak{B}_t} \boldsymbol{H}_i(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*) - \frac{\eta}{B}\sum_{i\in\mathfrak{B}_t} \boldsymbol{g}_i. \tag{5}$$

Note that since $\boldsymbol{\theta}^*$ is a minimum point of $\mathcal{L}$ we have that

$$\nabla \mathcal{L}(\boldsymbol{\theta}^*) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}_i = \boldsymbol{0}. \tag{6}$$

Furthermore, the Hessian of the loss, which we denote by $\boldsymbol{H}$, is given by

$$\boldsymbol{H} \triangleq \nabla^2 \mathcal{L}(\boldsymbol{\theta}^*) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{H}_i. \tag{7}$$

Thus, the linearized dynamics are in fact SGD iterates on the second-order Taylor expansion of $\mathcal{L}$ at $\boldsymbol{\theta}^*$,

$$\tilde{\mathcal{L}}(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}^*) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^{\mathrm{T}} \boldsymbol{H} (\boldsymbol{\theta} - \boldsymbol{\theta}^*). \tag{8}$$

3. Stability of first and second moments

Our focus is on the stability of SGD's dynamics at the vicinity of minima. We specifically examine the dynamics within two subspaces: the null space of the Hessian $\boldsymbol{H}$ at the minimum, and its orthogonal complement. We denote the projection of any vector $\boldsymbol{v} \in \mathbb{R}^d$ onto the null space of $\boldsymbol{H}$ by $\boldsymbol{v}^{\parallel}$, and its projection onto the orthogonal complement of the null space by $\boldsymbol{v}^{\perp}$.

Multiple works studied the stability of SGD's dynamics. Commonly, this was done by analyzing the evolution of the moments of the linearized dynamics (see Sec. 2) over time, with a specific emphasis on the second moment, which is the approach we take here. However, before discussing the evolution of the second moment, let us summarize the behavior of the first moment. Specifically, it is easy to demonstrate that the first moment of SGD's linearized trajectory $\{\mathbb{E}[\boldsymbol{\theta}_t]\}$ is the same as GD's. Since GD is stable if and only if $\eta \leq 2/\lambda_{\max}(\boldsymbol{H})$, we have the following (see proof in App. B.2).

Theorem 2 (Stability of the mean) *Assume that $\boldsymbol{\theta}^*$ is a twice differentiable minimum. Consider the linear dynamics of $\{\boldsymbol{\theta}_t\}$ from Def. 1 and let*

$$\eta^*_{\text{mean}} \triangleq \frac{2}{\lambda_{\max}(\boldsymbol{H})}. \tag{9}$$

Then

1. $\mathbb{E}\left[\boldsymbol{\theta}_t^{\parallel}\right] = \mathbb{E}\left[\boldsymbol{\theta}_0^{\parallel}\right]$ *for all* $t \geq 0$*;*
2. $\limsup_{t\to\infty} \left\|\mathbb{E}\left[\boldsymbol{\theta}_t\right] - \boldsymbol{\theta}^*\right\|$ *is finite if and only if* $\eta \leq \eta^*_{\text{mean}}$*;*
3. $\lim_{t\to\infty} \left\|\mathbb{E}\left[\boldsymbol{\theta}_t^{\perp}\right] - \boldsymbol{\theta}^{*\perp}\right\| = 0$ *if* $\eta < \eta^*_{\text{mean}}$*.*

We next proceed to analyze the dynamics of the second moment, which determine stability in the mean square sense. Note that boundedness of the first moment is a necessary condition for boundedness of the second moment. Therefore, the condition $\eta \leq \eta^*_{\text{mean}}$ is a prerequisite for stability in the mean square sense. However, how much smaller than η^*_{mean} is SGD's mean square stability threshold, is not currently known in closed-form. Here, we determine the precise threshold for the mean square stability of SGD's linearized dynamics. To achieve this, we leverage the approach taken by Ma and Ying (2021), who investigated the stability of SGD in the context of interpolating minima.

3.1. Interpolating minima

We begin by studying interpolating minima, which are prevalent in the training of overparametrized models. In this case, the model fits the training set perfectly, which means that these global minima are also minima for each sample individually. This is expressed mathematically as follows.

Definition 3 (Interpolating minima) *A twice differentiable minimum $\boldsymbol{\theta}^*$ is said to be* interpolating *if for each sample $i \in [n]$ the gradient $\boldsymbol{g}_i = \boldsymbol{0}$ and the Hessian $\boldsymbol{H}_i$ is PSD.*

In this setting, Ma and Ying (2021) showed that the evolution over time of any moment of SGD's linearized dynamics is fully tractable. Specifically, for the second moment, they proved the following.

Theorem 4 (Ma and Ying (2021), Thm. 1 + Cor. 3) *Assume that $\boldsymbol{\theta}^*$ is a twice differentiable interpolating minimum. Consider the linear dynamics of $\{\boldsymbol{\theta}_t\}$ from Def. 1, and let*

$$\boldsymbol{Q}(\eta, B) \triangleq (\boldsymbol{I} - \eta\boldsymbol{H}) \otimes (\boldsymbol{I} - \eta\boldsymbol{H}) + \frac{n-B}{B(n-1)} \frac{\eta^2}{n} \sum_{i=1}^{n} (\boldsymbol{H}_i \otimes \boldsymbol{H}_i - \boldsymbol{H} \otimes \boldsymbol{H}), \tag{10}$$

where $\otimes$ denotes the Kronecker product. Then $\limsup_{t\to\infty} \mathbb{E}[\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|^2]$ is finite if and only if

$$\max_{\boldsymbol{\Sigma}\in\mathcal{S}_+(\mathbb{R}^{d\times d})} \frac{\|\boldsymbol{Q}(\eta, B)\ \mathrm{vec}\,(\boldsymbol{\Sigma})\|}{\|\boldsymbol{\Sigma}\|_{\mathrm{F}}} \leq 1, \tag{11}$$

where $\mathcal{S}_+(\mathbb{R}^{d\times d})$ denotes the set of all PSD matrices over $\mathbb{R}^{d\times d}$. Furthermore, if the spectral radius $\rho(\boldsymbol{Q}(\eta, B)) \leq 1$ then $\limsup_{t\to\infty} \mathbb{E}\left[\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|^2\right]$ is finite.

Below, we omit the dependence of $\boldsymbol{Q}$ on η and B whenever these are not essential for the discussion. In this theorem, $\boldsymbol{\Sigma}$ represents the second-moment matrix of $\boldsymbol{\theta}_t - \boldsymbol{\theta}^*$. Specifically, the matrix $\boldsymbol{\Sigma}_t = \mathbb{E}[(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*)(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*)^{\mathrm{T}}]$ evolves over time as $\mathrm{vec}(\boldsymbol{\Sigma}_{t+1}) = \boldsymbol{Q}\,\mathrm{vec}(\boldsymbol{\Sigma}_t)$. Therefore, the stability condition of (11) simply states that if the dynamics of the dominant initial state of the system (which is restricted to PSD matrices) is bounded, then $\boldsymbol{\Sigma}_t$ is bounded and vice versa. However, this characterization leaves us with a constrained optimization problem over a d^2-dimensional space, which is hard to solve numerically. Therefore, this approach does not reduce the problem into a condition from which we can gain any theoretical insight into SGD's stability.

Our first key result is that the constrained optimization problem in (11) can be reduced to an eigenvalue problem. Specifically, we establish (see Sec. 3.3) that when the eigenvectors of the $d^2 \times d^2$ matrix $\boldsymbol{Q}$ are reshaped into $d \times d$ matrices, they correspond to either symmetric or skew-symmetric matrices[1]. Moreover, the top eigenvalue of $\boldsymbol{Q}$ is a dominant eigenvalue, and always corresponds to a PSD matrix. Consequently, the maximizer of (11) is the top eigenvector of $\boldsymbol{Q}$, which we use, along with some algebraic manipulation, to derive the following result (see proof in App. B.8).

Theorem 5 (Mean square stability for interpolating minima) *Assume that $\boldsymbol{\theta}^*$ is a twice differentiable interpolating minimum. Consider the linear dynamics of $\{\boldsymbol{\theta}_t\}$ from Def. 1, and let*

$$\boldsymbol{C} \triangleq \frac{1}{2}\boldsymbol{H} \oplus \boldsymbol{H}, \qquad \boldsymbol{D} \triangleq (1-p)\,\boldsymbol{H} \otimes \boldsymbol{H} + p\,\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{H}_i \otimes \boldsymbol{H}_i, \tag{12}$$

where $\oplus$ denotes the Kronecker sum and $p \triangleq \frac{n-B}{B(n-1)} \in [0,1]$. Define

$$\eta^*_{\mathrm{var}} \triangleq \frac{2}{\lambda_{\max}\left(\boldsymbol{C}^\dagger \boldsymbol{D}\right)}, \tag{13}$$

where $\boldsymbol{C}^\dagger$ denotes the Moore-Penrose inverse of $\boldsymbol{C}$. Then

1. $\boldsymbol{\theta}_t^{\parallel} = \boldsymbol{\theta}_0^{\parallel}$ *(surely) for all* $t \geq 0$;
2. $\limsup_{t\to\infty} \mathbb{E}\left[\|\boldsymbol{\theta}_t^{\perp} - \boldsymbol{\theta}^{*\perp}\|^2\right]$ *is finite if and only if* $\eta \leq \eta^*_{\mathrm{var}}$;
3. $\lim_{t\to\infty} \mathbb{E}\left[\|\boldsymbol{\theta}_t^{\perp} - \boldsymbol{\theta}^{*\perp}\|^2\right] = 0$ *if* $\eta < \eta^*_{\mathrm{var}}$.

[1] Eigenbases corresponding to eigenvalues of multiplicity greater than one, always have a basis consisting of symmetric and skew-symmetric matrices.

This result provides an explicit characterization of the mean square stability of SGD. Here we see that the set of step sizes that are stable in the mean square sense, is an interval. This is in contrast to stability in probability, where the stable learning rates can comprise of several disjoint intervals (Ziyin et al., 2023). Moreover, SGD's threshold, η^*_{var}, has the same form as the threshold for GD, $2/\lambda_{\max}$, but with a different matrix. In App. J we show how Thm. 5 recovers GD's condition when $B = n$.

The dependence of η^*_{var} on the batch size B may not be immediate to see from the theorem. However, we can prove the following (see proof in App. D).

Proposition 6 (Monotonicity of the stability threshold) *Assume that $\boldsymbol{\theta}^*$ is a twice differentiable interpolating minimum. Then η^*_{var} is a non-decreasing function of B.*

This result implies that decreasing the batch size can only impair stability, which settles with empirical observations, *e.g.,* in (Wu et al., 2018, Fig. 1). Additionally, since η^*_{var} is non-decreasing with B, and for $B = n$ it equals η^*_{mean}, we have that the gap between $\lambda_{\max}(\boldsymbol{C}^\dagger\boldsymbol{D})$ and $\lambda_{\max}(\boldsymbol{H})$ is nonnegative for all $B \in [1, n]$ and non-increasing in B. For stable minima, $\lambda_{\max}(\boldsymbol{C}^\dagger\boldsymbol{D})$ is bounded from above by $2/\eta$. This suggests that training with smaller batches leads to lower $\lambda_{\max}(\boldsymbol{H})$, *i.e.,* flatter minima, which aligns with experimental results (Keskar et al., 2016; Jastrzębski et al., 2017).

At what rate does η^*_{var} increase with B towards η^*_{mean}? To understand this, note that $\boldsymbol{D}$ is a convex combination of two matrices, where p represents the combination weight. The first matrix, $\boldsymbol{H} \otimes \boldsymbol{H}$, is associated with full batch SGD ($B = n$), while the second matrix, $\frac{1}{n}\sum_{i=1}^n \boldsymbol{H}_i \otimes \boldsymbol{H}_i$, is related to single sample SGD ($B = 1$). We can use this fact to explain the effect of the batch size on dynamical stability by presenting an equivalent stochastic process that has the same stability threshold as SGD (see proof in App. E).

Proposition 7 (Equivalent mixture process) *Let $\texttt{ALG}(p)$ be a stochastic optimization algorithm in which*

$$\boldsymbol{\theta}_{t+1} = \begin{cases} \boldsymbol{\theta}_t - \eta\nabla\ell_{i_t}(\boldsymbol{\theta}_t) & \textit{w.p.} \quad p, \\ \boldsymbol{\theta}_t - \eta\nabla\mathcal{L}(\boldsymbol{\theta}_t) & \textit{w.p.} \quad 1-p, \end{cases} \tag{14}$$

where $\{i_t\}$ are i.i.d. random indices distributed uniformly over the training set. Assume that $\boldsymbol{\theta}^*$ is a twice differentiable interpolating minimum. Then when $p = \frac{n-B}{B(n-1)}$, $\texttt{ALG}(p)$ has the same stability threshold in the vicinity of $\boldsymbol{\theta}^*$ as SGD with batch size B.

In simpler terms, $\texttt{ALG}(p)$ is a mixture process that takes in each iteration a gradient step with a batch of one sample ($B = 1$) with probability p and with a full batch ($B = n$) with probability $1 - p$. This result shows that the stability conditions of SGD and of $\texttt{ALG}(p)$ are the same for $p = \frac{n-B}{B(n-1)}$. For $n \gg B$, we get $p \approx 1/B$. Thus, Prop. 7 implies that, in the context of stability, even moderate values of B make mini-batch SGD behave like GD. We next quantify how large B needs to be in order for the gap between η^*_{var} and η^*_{mean} to be small (see proof in App. F).

Proposition 8 (Stability gap) *Define $\boldsymbol{E} = \frac{1}{n}\sum_{i=1}^n(\boldsymbol{H}_i - \boldsymbol{H}) \otimes (\boldsymbol{H}_i - \boldsymbol{H})$ and let $\varepsilon \in (0, 1)$. If*

$$B \geq \frac{1-\varepsilon}{\varepsilon}\,\frac{\lambda_{\max}(\boldsymbol{C}^\dagger\boldsymbol{E})}{\lambda_{\max}(\boldsymbol{H})}, \tag{15}$$

then

$$(1-\varepsilon)\eta^*_{\text{mean}} \leq \eta^*_{\text{var}} \leq \eta^*_{\text{mean}}. \tag{16}$$

Here $\boldsymbol{E}$ captures the variance of the per-sample Hessians. Thus, this result suggests that when these Hessians are similar (*i.e.,* the entries of $\boldsymbol{E}$ are small), moderate batch sizes are sufficient to guarantee a small gap between η^*_{mean} and η^*_{var}. On the other hand, if the variance of the Hessians is large, then the batch size B is expected to be large for the stability thresholds of SGD and GD to be close. We note that while propositions 6-8 were presented in the context of interpolating minima, they also apply to regular minima (see Sec. 3.2).

Although Thm. 5 provides an explicit threshold for the step size, its computation may be challenging in practical applications, as it requires inverting, multiplying, and computing the spectral norm of large ($d^2 \times d^2$) matrices. Still, we can obtain necessary criteria for stability that are simple and easier to verify, and which also depend on the batch size. To do so, we compute quadratic forms over $\boldsymbol{C}^\dagger \boldsymbol{D}$ with non-optimal yet interesting vectors. In this way, we bound $\lambda_{\max}(\boldsymbol{C}^\dagger \boldsymbol{D})$ from below to get the following result (see detail and proof in App. G).

Proposition 9 (Necessary conditions for stability) *Let $\boldsymbol{v}_{\max}$ be a top eigenvector of $\boldsymbol{H}$. Then the step size η^*_{var} satisfies*

$$\eta^*_{\text{var}} \leq \frac{2\lambda_{\max}(\boldsymbol{H})}{\lambda^2_{\max}(\boldsymbol{H}) + \frac{p}{n}\sum_{i=1}^{n}(\boldsymbol{v}^{\mathrm{T}}_{\max}\boldsymbol{H}_i\boldsymbol{v}_{\max} - \lambda_{\max}(\boldsymbol{H}))^2}, \tag{17}$$

as well as

$$\eta^*_{\text{var}} \leq \frac{2\mathrm{Tr}(\boldsymbol{H})}{(1-p)\left\|\boldsymbol{H}\right\|^2_{\mathrm{F}} + \frac{p}{n}\sum_{i=1}^{n}\left\|\boldsymbol{H}_i\right\|^2_{\mathrm{F}}}. \tag{18}$$

From (17), we can deduce a lower bound on the gap between the stability thresholds of GD and SGD. Specifically, when the variance of $\boldsymbol{H}_i$ along the direction of the top eigenvector of $\boldsymbol{H}$ is large, η^*_{var} is far from η^*_{mean} for moderate p. In general, this condition is expected to be quite tight when there is a clear dominant direction in $\boldsymbol{H}$ caused by some $\boldsymbol{H}_i$. In contrast, condition (18) is expected to be tight if all $\{\boldsymbol{H}_i\}$ have roughly the same spectrum but with different bases, *i.e.,* when no sample is dominant and the samples are incoherent.

It is worthwhile mentioning that if the stability condition of Thm. 5 is not met, then the linearized dynamics diverge. However, in practice, the full (non-linearized) dynamics can just move to a different point on the loss landscape, where the generalized sharpness $\lambda_{\max}(\boldsymbol{C}^\dagger \boldsymbol{D})$ is lower. It was shown that GD possesses such a stabilizing mechanism (Damian et al., 2023). An interesting open question is whether a similar mechanism exists in SGD.

3.2. Non-interpolating minima

While for interpolating minima, we saw that $\boldsymbol{\theta}^\perp_t$ can converge to $\boldsymbol{\theta}^{*\perp}$, this is generally not the case for non-interpolating minima. In this section, we explore the dynamics of SGD at the vicinity of a broader class of minima. Specifically, we consider the following definition.

Definition 10 (Regular minima) *A twice differentiable minimum $\boldsymbol{\theta}^*$ is said to be* regular *if for each sample $i \in [n]$ the Hessian $\boldsymbol{H}_i$ is PSD.*

This definition encompasses a broader class of minima than Def. 3, as it allows for arbitrary (nonzero) gradients $\boldsymbol{g}_i$. Only the mean of the gradients has to vanish (as in any minimum). Intuitively speaking, although a regular minimum does not necessarily fit all the training points, it does not involve a major disagreement among them. This can be understood through the second-order Taylor expansion of the loss per sample, which can have descent directions in the parameter space, yet it can only decrease (on behalf of raising the loss to other samples) linearly with the parameters, and not quadratically.

Clearly, having gradients with nonzero components in the null space of the Hessian pushes the dynamics to diverge. Interestingly, for regular minima, the dynamics of SGD in the null space and in its orthogonal complement are separable. Thus, despite having a random walk in the null space, we can give a condition for stability within its orthogonal complement (see proof in App. B.9).

Theorem 11 (Mean square stability for regular minima) *Assume that $\boldsymbol{\theta}^*$ is a twice differentiable regular minimum. Consider the linear dynamics of $\{\boldsymbol{\theta}_t\}$ from Def. 1. Then*

1. $\lim\limits_{t\to\infty} \mathbb{E}\left[\|\boldsymbol{\theta}_t^{\parallel} - \boldsymbol{\theta}^{*\parallel}\|^2\right] = \infty$ *if and only if* $\sum_{i=1}^n \|\boldsymbol{g}_i^{\parallel}\|^2 > 0$*;*
2. *If* $\eta < \eta^*_{\text{var}}$ *then* $\limsup\limits_{t\to\infty} \mathbb{E}\left[\|\boldsymbol{\theta}_t^{\perp} - \boldsymbol{\theta}^{*\perp}\|^2\right]$ *is finite;*
3. *If* $\limsup\limits_{t\to\infty} \mathbb{E}\left[\|\boldsymbol{\theta}_t^{\perp} - \boldsymbol{\theta}^{*\perp}\|^2\right]$ *is finite then* $\eta \leq \eta^*_{\text{var}}$.

We see that η^*_{var} is the stability threshold also for regular minima. Recall that when $\eta < \eta^*_{\text{var}}$, we also have stability of the first moment, and thus $\mathbb{E}[\boldsymbol{\theta}_t^{\parallel}] = \mathbb{E}[\boldsymbol{\theta}_0^{\parallel}]$ for any $t \geq 0$. Namely, SGD's dynamics in the null space is a random walk without drift. Note that moving in the null space does not increase the loss, however it might change the trained model. Furthermore, in the proof, we show that under a mild assumption, $\limsup\limits_{t\to\infty}\mathbb{E}[\|\boldsymbol{\theta}_t^{\perp} - \boldsymbol{\theta}^{*\perp}\|^2]$ is finite if and only if $0 \leq \eta < \eta^*_{\text{var}}$.

Next, we turn to compute the limit of the second moment of the dynamics (see proof in App. H).

Theorem 12 (Covariance limit) *Assume that $\boldsymbol{\theta}^*$ is a twice differentiable regular minimum. Consider the linear dynamics of $\{\boldsymbol{\theta}_t\}$ from Def. 1. If $0 < \eta < \eta^*_{\text{var}}$ then*

$$\lim_{t\to\infty} \operatorname{vec}\left(\boldsymbol{\Sigma}_t^{\perp}\right) = \eta p \left(2\boldsymbol{C} - \eta \boldsymbol{D}\right)^{\dagger} \operatorname{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{g}}^{\perp}\right), \tag{19}$$

where

$$\boldsymbol{\Sigma}_{\boldsymbol{g}}^{\perp} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}_i^{\perp} \left(\boldsymbol{g}_i^{\perp}\right)^{\mathrm{T}}. \tag{20}$$

Using this result we can obtain the mean squared distance to the minimum, the mean of the second-order Taylor expansion of the loss, and the mean of the squared norm of the expansion's gradient at large times (see proof in App. I).

Corollary 13 *Let $\boldsymbol{\theta}^*$ be a twice differentiable regular minimum. Consider the second-order Taylor expansion of the loss, $\tilde{\mathcal{L}}$ of (8) and the linear dynamics of $\{\boldsymbol{\theta}_t\}$ from Def. 1. If $\eta < \eta^*_{\text{var}}$ then*

1. $\lim\limits_{t\to\infty} \mathbb{E}\left[\|\boldsymbol{\theta}_t^{\perp} - \boldsymbol{\theta}^{*\perp}\|^2\right] = \eta p (\operatorname{vec}(\boldsymbol{I}))^{\mathrm{T}} \left(2\boldsymbol{C} - \eta\boldsymbol{D}\right)^{\dagger} \operatorname{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{g}}^{\perp}\right)$*;*
2. $\lim\limits_{t\to\infty} \mathbb{E}\left[\tilde{\mathcal{L}}(\boldsymbol{\theta}_t)\right] - \tilde{\mathcal{L}}(\boldsymbol{\theta}^*) = \frac{1}{2}\eta p (\operatorname{vec}(\boldsymbol{H}))^{\mathrm{T}} \left(2\boldsymbol{C} - \eta\boldsymbol{D}\right)^{\dagger} \operatorname{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{g}}^{\perp}\right)$*;*
3. $\lim\limits_{t\to\infty} \mathbb{E}\left[\|\nabla\tilde{\mathcal{L}}(\boldsymbol{\theta}_t)\|^2\right] = \eta p \left(\operatorname{vec}\left(\boldsymbol{H}^2\right)\right)^{\mathrm{T}} \left(2\boldsymbol{C} - \eta\boldsymbol{D}\right)^{\dagger} \operatorname{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{g}}^{\perp}\right)$.

We see that these values depend linearly on the covariance matrix of the gradients. Specifically, if $\boldsymbol{\Sigma}_{\boldsymbol{g}} = \boldsymbol{0}$ then we recover the results of interpolating minima. Moreover, note that for $\eta \ll \eta^*_{\text{var}}$, we have that $2\boldsymbol{C} - \eta\boldsymbol{D} \approx 2\boldsymbol{C}$. Therefore, the main dependence on η comes from the factor of η preceding these expressions. We thus get that when decreasing the learning rate, the loss level drops, and the parameters $\boldsymbol{\theta}_t$ get closer to the minimum. This explains the empirical behavior observed when decreasing the learning rate in neural network training, which causes the loss level to drop.

3.3. Derivation of the stability threshold η^*_{var}

In this section, we give a sketch of the derivation of the stability threshold η^*_{var} in the simple case of interpolating minima (*i.e.,* the second statement of Thm. 5). For the full proof, please see App. B.8. In interpolating minima, the gradient vanishes for each sample, *i.e.,* $\boldsymbol{g}_i = \mathbf{0}$ for all $i \in [n]$. Therefore, from the linearized dynamics (5) we get

$$\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^* = \Big(\boldsymbol{I} - \frac{\eta}{B}\sum_{i\in\mathfrak{B}_t}\boldsymbol{H}_i\Big)(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*) = \boldsymbol{A}_t(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*), \tag{21}$$

where $\boldsymbol{A}_t \triangleq \boldsymbol{I} - \frac{\eta}{B}\sum_{i\in\mathfrak{B}_t}\boldsymbol{H}_i$. Note that $\{\boldsymbol{A}_t\}$ are i.i.d. and that $\boldsymbol{\theta}_t$ is constructed from $\boldsymbol{A}_0, \dots, \boldsymbol{A}_{t-1}$, so that $\boldsymbol{\theta}_t$ and $\boldsymbol{A}_t$ are statistically independent. Therefore, the covariance of the dynamics evolves as

$$\begin{aligned}
\boldsymbol{\Sigma}_{t+1} = \mathbb{E}\left[(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^*)(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^*)^{\mathrm{T}}\right] &= \mathbb{E}\left[\boldsymbol{A}_t(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*)(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*)^{\mathrm{T}}\boldsymbol{A}_t^{\mathrm{T}}\right] \\
&= \mathbb{E}\left[\boldsymbol{A}_t\mathbb{E}\left[(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*)(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*)^{\mathrm{T}}\middle|\boldsymbol{A}_t\right]\boldsymbol{A}_t^{\mathrm{T}}\right] \\
&= \mathbb{E}\left[\boldsymbol{A}_t\mathbb{E}\left[(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*)(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*)^{\mathrm{T}}\right]\boldsymbol{A}_t^{\mathrm{T}}\right] = \mathbb{E}\left[\boldsymbol{A}_t\boldsymbol{\Sigma}_t\boldsymbol{A}_t^{\mathrm{T}}\right],
\end{aligned} \tag{22}$$

where in the first line we used (21), in the second we used the law of total expectation, and in the third we used the fact that $\boldsymbol{\theta}_t$ is statistically independent of $\boldsymbol{A}_t$. Using vectorization we get

$$\operatorname{vec}(\boldsymbol{\Sigma}_{t+1}) = \mathbb{E}[\boldsymbol{A}_t \otimes \boldsymbol{A}_t]\operatorname{vec}(\boldsymbol{\Sigma}_t). \tag{23}$$

Ma and Ying (2021) showed that $\mathbb{E}[\boldsymbol{A}_t \otimes \boldsymbol{A}_t] = \boldsymbol{Q}$, where $\boldsymbol{Q}$ is given in (10). Since $\boldsymbol{\Sigma}_t$ is PSD by definition, we only care about the effect of $\boldsymbol{Q}$ on vectorizations of PSD matrices. Thus, $\{\boldsymbol{\Sigma}_t\}$ are bounded if and only if (see proof in (Ma and Ying, 2021))

$$\max_{\boldsymbol{\Sigma}\in\mathcal{S}_+(\mathbb{R}^{d\times d})} \frac{\|\boldsymbol{Q}(\eta, B)\operatorname{vec}(\boldsymbol{\Sigma})\|}{\|\boldsymbol{\Sigma}\|_{\mathrm{F}}} \le 1. \tag{24}$$

This constrained optimization problem is hard to solve. Yet, if we ignore the constraint then the solution becomes simple – it is the spectral radius of $\boldsymbol{Q}$ (since $\boldsymbol{Q}$ is symmetric). Surprisingly, it turns out that removing the constraint does not affect the solution because the matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ that maximizes the objective in (24) without constraints is guaranteed to be PSD, so that it also maximizes the objective under the constraint $\boldsymbol{\Sigma} \in \mathcal{S}_+(\mathbb{R}^{d\times d})$. Proving this fundamental algebraic property is a main challenge and a key contribution of our work (see proof in App. C).

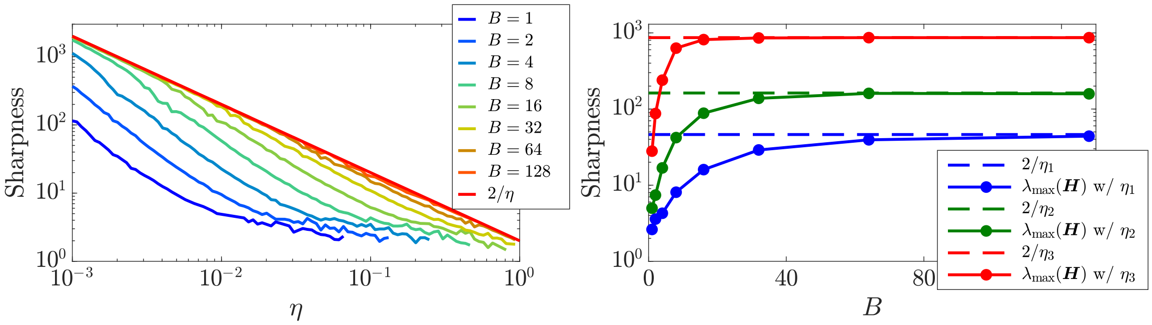
Theorem 14 (Symmetric Kronecker systems) *Let $\{\boldsymbol{Y}_i\}$ be symmetric matrices in $\mathbb{R}^{d\times d}$. Define*

$$\boldsymbol{Q} = \sum_{i=1}^{M}\boldsymbol{Y}_i \otimes \boldsymbol{Y}_i, \tag{25}$$

and let $\boldsymbol{z}_{\max}$ be a top eigenvector of $\boldsymbol{Q}$. Then

1. *there always exists a complete set of eigenvectors $\{\boldsymbol{z}_j\}$ for $\boldsymbol{Q}$ such that each $\boldsymbol{Z}_j = \operatorname{vec}^{-1}(\boldsymbol{z}_j)$ is either a symmetric or a skew-symmetric matrix;*
2. *the top eigenvalue is a dominant eigenvalue[2], i.e., the spectral radius $\rho(\boldsymbol{Q}) = \lambda_{\max}(\boldsymbol{Q})$;*
3. *there exists a top eigenvector corresponding to a PSD matrix, i.e., $\operatorname{vec}^{-1}(\boldsymbol{z}_{\max}) \in \mathcal{S}_+(\mathbb{R}^{d\times d})$.*

[2] By "top" we refer to the largest eigenvalue, and by "dominant" we refer to the largest eigenvalues in absolute value.



(a) Sharpness vs. step size

(b) Sharpness vs. batch size

Figure 1: **Sharpness vs. step size and batch size.** We trained single hidden-layer ReLU networks using varying step sizes and batch sizes on a subset of MNIST. Panel (a) visualizes the sharpness of the converged minima versus learning rate for different batch sizes. For small batch sizes, $\lambda_{\max}(\boldsymbol{H})$ deviates significantly from $2/\eta$. Yet as the batch size increases to a moderate value, these curves coincide, indicating that in terms of stability, SGD behaves similarly to GD. Panel (b) plots the sharpness against the batch size for three different learning rates $\eta_1 = 0.043, \eta_2 = 0.012, \eta_3 = 0.002$. Here we see a similar trend where SGD behaves like GD for $B \geq 32$.

Taking $\{\boldsymbol{Y}_i\}$ to be the $\binom{n}{B}$ realizations that $\boldsymbol{A}_t$ can take, we obtain that $\boldsymbol{Q} = \mathbb{E}[\boldsymbol{A}_t \otimes \boldsymbol{A}_t]$ is of the form (25). Note that the realizations of $\boldsymbol{A}_t$ are symmetric, so that Thm. 14 applies. Therefore, the matrix that attains the maximum in the optimization problem (24) without the constraint, whose vectorization is generally a dominant eigenvector of $\boldsymbol{Q}$, is guaranteed to be a PSD matrix. Furthermore, the corresponding objective value is $\lambda_{\max}(\boldsymbol{Q})$. This implies that the linear system in (23) is stable if and only if $\lambda_{\max}(\boldsymbol{Q}) \leq 1$. Since $\boldsymbol{Q}$ is symmetric, $\lambda_{\max}(\boldsymbol{Q}) \leq 1$ is equivalent to $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{u} \leq 1$ for all $\boldsymbol{u} \in \mathbb{S}^{d^2-1}$. It is easy to show that $\boldsymbol{Q} = \boldsymbol{I} - 2\eta\boldsymbol{C} + \eta^2\boldsymbol{D}$ (see (67)). In App. B.8 we prove that $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{u} = 1 - 2\eta\boldsymbol{u}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{u} + \eta^2\boldsymbol{u}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{u} \leq 1$ holds for all $\boldsymbol{u} \in \mathbb{S}^{d^2-1}$ if and only if

$$\eta \leq \frac{2}{\lambda_{\max}\left(\boldsymbol{C}^{\dagger}\boldsymbol{D}\right)} = \eta^*_{\mathrm{var}}. \tag{26}$$

4. Experiments

In this section, we experimentally validate our theoretical results in a setting with nonlinear dynamics. We trained single hidden-layer ReLU networks with varying step sizes and batch sizes on a subset of the MNIST dataset (see App. K). Since training with cross-entropy in overparametrized networks results in infima rather than minima, we used the quadratic loss. Specifically, each class was labeled with a one-hot vector, and the network was trained to predict the label without softmax. Our primary goal in this experiment is to test the stability threshold of SGD; hence, we initialized the training with large weights to ensure that the minimum closest to the starting point is unstable (large weights imply large Hessians, and are thus more likely to violate the stability criterion). We used the same initial point for all the training runs to eliminate initialization effects. To avoid divergence, we started with a very small step size and gradually increased it until it reached its designated value (*i.e.,* learning rate warm-up). Together, large initialization and warm-up force SGD out of the unstable region until it finds a stable minimum and converges as closely as possible to the stability threshold. Convergence was determined when the loss remained below 10^{-6} for 200 consecutive epochs.

Figure 1(a) visualizes the sharpness of the converged minima versus the learning rate for several values of B. Here we observe that for small batch sizes, $\lambda_{\max}(\boldsymbol{H})$ is far from $2/\eta$. Yet for moderate batch sizes and above (*e.g.,* $B \geq 32$), these curves virtually coincide, indicating that, in the context of stability, SGD behaves like GD. Figure 1(b) shows the sharpness versus the batch size for three step sizes. Here the stability threshold of SGD rapidly converges to that of GD as the batch size increases.

Apart for the sharpness $\lambda_{\max}(\boldsymbol{H})$, we also want to compare the generalized sharpness $\lambda_{\max}(\boldsymbol{C}^{\dagger}\boldsymbol{D})$ to $2/\eta$. Since computing the generalized sharpness is impractical in this task, we underestimate it via a lower bound, which results in a tighter necessary condition than (17). The bound corresponds to restricting the optimization problem in (11) to rank one PSD matrices, and is given by (see App. G.1)

$$\frac{2}{\eta^*_{\text{var}}} = \lambda_{\max}\left(\boldsymbol{C}^{\dagger}\boldsymbol{D}\right) \geq \max_{\boldsymbol{v}:\|\boldsymbol{v}\|=1}\left\{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{v} + p\frac{\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}_i\boldsymbol{v} - \boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{v})^2}{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{v}}\right\}. \tag{27}$$

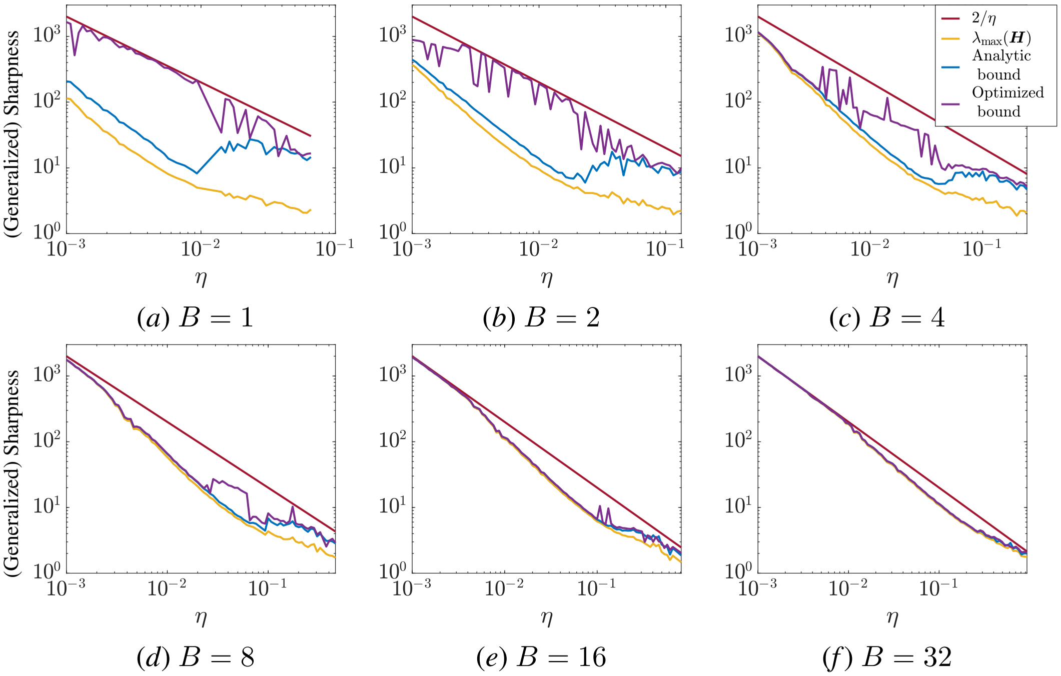
We solve this optimization problem numerically, by using GD on the unit sphere with predetermined scheduled geodesic step size. In the following, we present graphs of the sharpness $\lambda_{\max}(\boldsymbol{H})$ at the minima to which we converged, as well as the bounds (27) and (17) on the generalized sharpness $\lambda_{\max}(\boldsymbol{C}^{\dagger}\boldsymbol{D})$. Using the color coding of Fig. 2, these correspond to

$$\begin{aligned}\frac{2}{\eta} &\geq \lambda_{\max}\left(\boldsymbol{C}^{\dagger}\boldsymbol{D}\right) \quad \left(= \frac{2}{\eta^*_{\text{var}}}\right)\\ &\geq \max_{\boldsymbol{v}:\|\boldsymbol{v}\|=1}\left\{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{v} + p\frac{\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}_i\boldsymbol{v} - \boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{v})^2}{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{v}}\right\}\\ &\geq \lambda_{\max}(\boldsymbol{H}) + p\frac{\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{v}_{\max}^{\mathrm{T}}\boldsymbol{H}_i\boldsymbol{v}_{\max} - \lambda_{\max}(\boldsymbol{H}))^2}{\lambda_{\max}(\boldsymbol{H})}\\ &\geq \lambda_{\max}(\boldsymbol{H}),\end{aligned} \tag{28}$$

where $\boldsymbol{v}_{\max}$ denotes the top eigenvector of $\boldsymbol{H}$.

Figure 2 depicts the expressions in (28) versus the step size for six batch sizes. We see that for $B = 1$ and $B = 2$, the gap between $2/\eta$ (red) and the optimized bound (27) (purple) upon convergence is small. Particularly, they coincide over a wide range of step sizes η. Since the generalized sharpness $\lambda_{\max}(\boldsymbol{C}^{\dagger}\boldsymbol{D})$ must reside between those two curves, we can deduce two things: (a) Our theory correctly predicts the stability threshold, while SGD converged at the edge of stability (as designed in our experiment); (b) For small batches, the second order moment matrix that maximizes (11) is rank one. As the batch size increases, the two curves draw apart, indicating that the rank of the dominant second-order moment matrix becomes larger. Furthermore, the gap between our simple necessary condition (17) (blue) and the trivial bound of $2/\lambda_{\max}(\boldsymbol{H})$ (yellow) is large for high learning rates and small for small step sizes. This gap represents the variance of the widths of the minima of the per-sample losses (corresponding to the widths of the quadratic functions $\{(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^{\mathrm{T}}\boldsymbol{H}_i(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\}$) in the direction of $\boldsymbol{v}_{\max}$, the top eigenvector of $\boldsymbol{H}$. Thus we find that for small learning rates, this variance is small and the model is aligned in this direction, and for large learning rates, this variance is high.

A comment is in place regarding the fluctuation of the optimized bound. As described above, the value of this bound is obtained by an optimization problem which we solved using GD for each pair of step size and batch size. It may be that we have not found the global optimum for every step size, and got stuck at local maxima for some set of hyperparameters. This can explain why the curve



(*a*) $B = 1$ (*b*) $B = 2$ (*c*) $B = 4$

(*d*) $B = 8$ (*e*) $B = 16$ (*f*) $B = 32$

Figure 2: **(Generalized) Sharpness vs. step size.** We trained single hidden-layer ReLU networks using varying step sizes and batch sizes on MNIST dataset. For each pair of hyper-parameters (η, B), we measured the sharpness of the minimum (yellow), our necessary condition for stability (blue), and the optimized bound (purple), which their relations are given in (28). We see that for small batch sizes $B = 1$ and $B = 2$, the optimized bound (27) coincides with $2/\eta$, confirming that SGD converged at the edge of stability ($\eta = \eta^*_{\text{var}}$). For additional insights and detail, see Sec. 4.

falls down and then comes up again at some of the learning rates. Additionally, as we mentioned, the optimized bound is equivalent to restricting the optimization problem from (11) to rank one symmetric matrices. It is possible that for some minima of the loss, the optimal matrix of (11) is rank one, and therefore the bound is tight, while for others the rank is higher and thus the bound is not tight. Further study of SGD is needed to determine the cause of this behavior, which we leave for future work. For more details and experimental results, please see App. K.

5. Related work

The stability of SGD in the vicinity of minima has been previously studied in multiple works. On the theoretical side, Wu et al. (2018) examined stability in the mean square sense and gave an implicit sufficient condition. Granziol et al. (2022) used random matrix theory to find the maximal stable learning rate as a function of the batch size. Their work assumes some conditions on the Hessian's noise caused by batching, and the result holds in the limit of an infinite number of samples and batch size. Velikanov et al. (2023) examined SGD with momentum and derived a bound on the maximal learning rate. Their derivation uses "spectrally expressible" approximations and the result is given implicitly through a moment-generating function. Ma and Ying (2021) studied the dynamics of higher moments of SGD and gave an implicit necessary and sufficient condition for stability (see

Thm. 4 and the discussion following it). Wu et al. (2022) gave a necessary condition for stability via the alignment property. However, the result assumes and uses a lower bound on a property they coin "alignment" but an analytic bound for this alignment property is lacking for the general case. Ziyin et al. (2023) studied the stability of SGD in probability, rather than in mean square. Since convergence in probability is a weaker requirement, theoretically, SGD can converge with high probability to minima which are unstable in the mean square sense. Indeed, their theory predicts that SGD can converge far beyond GD's threshold. Yet this did not happen in extensive experiments done in Cohen et al. (2021, App. G) and Gilmer et al. (2022). Finally, Mulayoff et al. (2021) analyzed the stability in non-differentiable minima, and gave a necessary condition for a minimum to be "strongly stable", *i.e.,* such that SGD does not escape a ball with a given radius from the minimum.

Liu et al. (2021) studied the covariance matrix of the stationary distribution of the iterates in the vicinity of minima. Ziyin et al. (2022) improved their results while deriving an implicit equation that relates this covariance to the covariance of the gradient noise. However, these works do not discuss the conditions under which the dynamics converge to the stationary state. Lee and Jang (2023) studied the stability of SGD along its trajectory and gave an explicit exact condition. Yet their result does not apply to minima, since the denominator in their condition vanishes at minima.

On the empirical side, Cohen et al. (2021) examined the behavior of GD, and showed that it typically converges at the edge of stability. Additionally, for SGD (see their App. G) they found that with large batches, the sharpness behaves similarly to full-batch gradient descent. Moreover, they found that the smaller the batch size, the lower the sharpness at the converged minimum. Gilmer et al. (2022) studied how the curvature of the loss affects the training dynamics in multiple settings. They observed that SGD *with momentum* is stable only when the optimization trajectory primarily resides in a region of parameter space where $\eta \lesssim 2/\lambda_{\max}(\boldsymbol{H})$. Further experimental results in Jastrzębski et al. (2020, 2019) show that the sharpness along the trajectory of SGD is implicitly regularized.

6. Conclusion

We presented an explicit threshold on SGD's step size, which is both necessary and sufficient for guaranteeing mean-square stability. We showed that this threshold is a monotonically non-decreasing function of the batch size, which implies that decreasing the batch size can only make the process less stable. Additionally, we interpreted the role of the batch size B through an equivalent process that takes in each iteration either a full batch gradient step or a single sample gradient step. Our interpretation highlights that even with moderate batch sizes, SGD's stability threshold is very close to that of GD. We also proved simpler necessary conditions for stability, which depend on the batch size, and are easier to compute. Finally, we verified our theory through experiments on MNIST.

Acknowledgments

The research of Rotem Mulayoff was partially supported by the Planning and Budgeting Committee of the Israeli Council for Higher Education, and by the Andrew and Erna Finci Viterbi Graduate Fellowship. The research of Tomer Michaeli was partially supported by the Israel Science Foundation (grant no. 2318/22) and by the Ollendorff Minerva Center, ECE faculty, Technion.

References

Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021. 1, 13, 19

Alex Damian, Eshaan Nichani, and Jason D. Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. In *The Eleventh International Conference on Learning Representations*, 2023. 7

James Allen Fill and Donniell E Fishkind. The moore–penrose generalized inverse for sums of matrices. *SIAM Journal on Matrix Analysis and Applications*, 21(2):629–635, 2000. 48

Justin Gilmer, Behrooz Ghorbani, Ankush Garg, Sneha Kudugunta, Behnam Neyshabur, David Cardoze, George Edward Dahl, Zachary Nado, and Orhan Firat. A loss curvature perspective on training instabilities of deep learning models. In *International Conference on Learning Representations*, 2022. 1, 13

Diego Granziol, Stefan Zohren, and Stephen Roberts. Learning rates as a function of batch size: A random matrix theory approach to neural network training. *J. Mach. Learn. Res*, 23:1–65, 2022. 1, 2, 12

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997. 1

Stanisław Jastrzębski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in SGD. *arXiv preprint arXiv:1711.04623*, 2017. 1, 6

Stanisław Jastrzębski, Zachary Kenton, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. On the relation between the sharpest directions of DNN loss and the SGD step length. In *International Conference on Learning Representations*, 2019. 1, 13

Stanisław Jastrzębski, Maciej Szymczak, Stanislav Fort, Devansh Arpit, Jacek Tabor, Kyunghyun Cho*, and Krzysztof Geras*. The break-even point on optimization trajectories of deep neural networks. In *International Conference on Learning Representations*, 2020. 1, 13

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016. 1, 6

Alan J Laub. *Matrix analysis for scientists and engineers*. SIAM, 2004. 26

Yann LeCun. The MNIST database of handwritten digits. 1998. 2

Sungyoon Lee and Cheongjae Jang. A new characterization of the edge of stability based on a sharpness measure aware of batch gradient distribution. In *The Eleventh International Conference on Learning Representations*, 2023. 13

Kangqiao Liu, Liu Ziyin, and Masahito Ueda. Noise and fluctuation of finite learning rate stochastic gradient descent. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International*

Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, pages 7045–7056. PMLR, 18–24 Jul 2021. 13

Chao Ma and Lexing Ying. On linear stability of SGD and input-smoothness of neural networks. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 1, 3, 4, 5, 9, 12, 27, 43

Rotem Mulayoff and Tomer Michaeli. Unique properties of flat minima in deep networks. In *International Conference on Machine Learning*, pages 7108–7118. PMLR, 2020. 1

Rotem Mulayoff, Tomer Michaeli, and Daniel Soudry. The implicit bias of minima stability: A view from function space. *Advances in Neural Information Processing Systems*, 34:17749–17761, 2021. 1, 3, 13, 19

Mor Shpigel Nacson, Rotem Mulayoff, Greg Ongie, Tomer Michaeli, and Daniel Soudry. The implicit bias of minima stability in multivariate shallow ReLU networks. In *The Eleventh International Conference on Learning Representations*, 2023. 1

Maksim Velikanov, Denis Kuznedelev, and Dmitry Yarotsky. A view of mini-batch SGD via generating functions: conditions of convergence, phase transitions, benefit from negative momenta. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2, 12

Lei Wu, Zhanxing Zhu, et al. Towards understanding generalization of deep learning: Perspective of loss landscapes. *arXiv preprint arXiv:1706.10239*, 2017. 1

Lei Wu, Chao Ma, and E Weinan. How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective. In *Advances in Neural Information Processing Systems*, pages 8279–8288, 2018. 1, 3, 6, 12

Lei Wu, Mingze Wang, and Weijie Su. The alignment property of sgd noise and how it helps select flat minima: A stability analysis. *Advances in Neural Information Processing Systems*, 35: 4680–4693, 2022. 13

Liu Ziyin, Kangqiao Liu, Takashi Mori, and Masahito Ueda. Strength of minibatch noise in SGD. In *International Conference on Learning Representations*, 2022. 13

Liu Ziyin, Botao Li, Tomer Galanti, and Masahito Ueda. The probabilistic stability of stochastic gradient descent. *arXiv preprint arXiv:2303.13093*, 2023. 1, 6, 13

Bruno Zumino. Normal forms of complex matrices. *Journal of Mathematical Physics*, 3(5):1055–1057, 1962. 37

Appendix A. Notations and the Kronecker product

Throughout our derivations, we use the following notations.

a	Lower case non-bold letters for scalars
$\boldsymbol{a}$	Lower case bold letters for vectors
$\boldsymbol{A}$	Upper case bold for matrices
$\boldsymbol{a}_{[p]},\ \boldsymbol{A}_{[\ell,p]}$	p'th element of $\boldsymbol{a}$, (ℓ,p) element of $\boldsymbol{A}$
$\boldsymbol{A}^{\mathrm{T}}$	Transpose of $\boldsymbol{A}$
$\boldsymbol{A}^{\dagger}$	Moore–Penrose inverse of $\boldsymbol{A}$
$\boldsymbol{P}_{\mathcal{V}}$	Orthogonal projection matrix onto the subspace $\mathcal{V}$
$\mathcal{N}(\boldsymbol{A})$	Null space of $\boldsymbol{A}$
$\mathcal{N}^{\perp}(\boldsymbol{A})$	Orthogonal complement of the null space of $\boldsymbol{A}$
$\mathcal{R}(\boldsymbol{A})$	Range of $\boldsymbol{A}$
$\mathcal{R}^{\perp}(\boldsymbol{A})$	Orthogonal complement of the range of $\boldsymbol{A}$
$\otimes$	Kronecker product
$\oplus$	Kronecker sum
$^{\odot k}$	k'th Hadamard power
$\mathbb{E}$	Expectation
$\mathbb{P}$	Probability
$\\|\boldsymbol{a}\\|$	Euclidean norm of $\boldsymbol{a}$
$\\|\boldsymbol{A}\\|$	Top singular value of $\boldsymbol{A}$
$\\|\boldsymbol{A}\\|_{\mathrm{F}}$	Frobenius norm of $\boldsymbol{A}$
$\rho(\boldsymbol{A})$	Spectral radius of $\boldsymbol{A}$
$\mathrm{vec}(\boldsymbol{A})$	Vectorization of $\boldsymbol{A}$ (column stack)
$\mathrm{vec}^{-1}(\boldsymbol{a})$	Reshaping $\boldsymbol{a}$ back to $d \times d$ matrix
$\mathcal{L}$	Loss function
$\boldsymbol{\theta}$	Parameters vector of the loss
$\boldsymbol{\theta}^*$	Minimum point of the loss
d	Dimension of $\boldsymbol{\theta}$
n	Number of training samples
η	Step size
B	Batch size
p	Defined to be $(n-B)/\big(B(n-1)\big)$
$\boldsymbol{I}_d$	The $d \times d$ identity matrix (when the dimensions are clear, the subscript is omitted)
$\boldsymbol{\Sigma}$	Second moment matrix
$\boldsymbol{H}$	Hessian of the full loss at $\boldsymbol{\theta}^*$
$\boldsymbol{H}_i$	Hessian of the loss of the sample i at $\boldsymbol{\theta}^*$
$\boldsymbol{g}_i$	Gradient of the loss of the sample i at $\boldsymbol{\theta}^*$
$\boldsymbol{a}^{\parallel}$	Projection of $\boldsymbol{a}$ onto the null space of $\boldsymbol{H}$
$\boldsymbol{a}^{\perp}$	Projection of $\boldsymbol{a}$ onto the orthogonal complement of the null space of $\boldsymbol{H}$
$\mathcal{S}_+(\mathbb{R}^{d\times d})$	The set of all positive semi-definite (PSD) matrices over $\mathbb{R}^{d\times d}$
$\mathbb{S}^{d-1}$	Unit sphere in $\mathbb{R}^d$

Table 1: Table of notations

Further notations that we use are given below.

$$\begin{aligned}
\boldsymbol{\mu}_t &\triangleq \mathbb{E}\left[\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\right], & \boldsymbol{\Sigma}_t &\triangleq \mathbb{E}\left[\left(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\right)\left(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\right)^{\mathrm{T}}\right], \\
\boldsymbol{\mu}_t^{\perp} &\triangleq \mathbb{E}\left[\boldsymbol{\theta}_t^{\perp} - \boldsymbol{\theta}^{*\perp}\right], & \boldsymbol{\Sigma}_t^{\perp} &\triangleq \mathbb{E}\left[\left(\boldsymbol{\theta}_t^{\perp} - \boldsymbol{\theta}^{*\perp}\right)\left(\boldsymbol{\theta}_t^{\perp} - \boldsymbol{\theta}^{*\perp}\right)^{\mathrm{T}}\right], \\
\boldsymbol{\mu}_t^{\parallel} &\triangleq \mathbb{E}\left[\boldsymbol{\theta}_t^{\parallel} - \boldsymbol{\theta}^{*\parallel}\right], & \boldsymbol{\Sigma}_t^{\parallel} &\triangleq \mathbb{E}\left[\left(\boldsymbol{\theta}_t^{\parallel} - \boldsymbol{\theta}^{*\parallel}\right)\left(\boldsymbol{\theta}_t^{\parallel} - \boldsymbol{\theta}^{*\parallel}\right)^{\mathrm{T}}\right].
\end{aligned} \tag{29}$$

Additionally, we make extensive use of the following properties of the Kronecker product throughout the derivations. For any matrices $\boldsymbol{M}_1, \boldsymbol{M}_2, \boldsymbol{M}_3, \boldsymbol{M}_4$,

$$\mathrm{vec}\left(\boldsymbol{M}_1\boldsymbol{M}_2\boldsymbol{M}_3\right) = \left(\boldsymbol{M}_3^{\mathrm{T}} \otimes \boldsymbol{M}_1\right)\mathrm{vec}\left(\boldsymbol{M}_2\right), \tag{P1}$$

$$\left(\boldsymbol{M}_1 \otimes \boldsymbol{M}_2\right)^{\mathrm{T}} = \boldsymbol{M}_1^{\mathrm{T}} \otimes \boldsymbol{M}_2^{\mathrm{T}}, \tag{P2}$$

$$\left(\boldsymbol{M}_1 \otimes \boldsymbol{M}_2\right)\left(\boldsymbol{M}_3 \otimes \boldsymbol{M}_4\right) = \left(\boldsymbol{M}_1\boldsymbol{M}_3\right) \otimes \left(\boldsymbol{M}_2\boldsymbol{M}_4\right), \tag{P3}$$

$$\left[\mathrm{vec}\left(\boldsymbol{M}_1\right)\right]^{\mathrm{T}}\left(\boldsymbol{M}_2 \otimes \boldsymbol{M}_3\right)\mathrm{vec}\left(\boldsymbol{M}_4\right) = \mathrm{Tr}\left(\boldsymbol{M}_1^{\mathrm{T}}\boldsymbol{M}_3\boldsymbol{M}_4\boldsymbol{M}_2^{\mathrm{T}}\right). \tag{P4}$$

Finally, we give here the definition of Kronecker sum. If $\boldsymbol{M}_1$ is $d_1 \times d_1$, $\boldsymbol{M}_2$ is $d_2 \times d_2$ and $\boldsymbol{I}_d$ denotes the $d \times d$ identity matrix then

$$\boldsymbol{M}_1 \oplus \boldsymbol{M}_2 = \boldsymbol{M}_1 \otimes \boldsymbol{I}_{d_2} + \boldsymbol{I}_{d_1} \otimes \boldsymbol{M}_2. \tag{30}$$

Appendix B. Stability of the first and second moments

Using our notation (see App. A), for all $\boldsymbol{v} \in \mathbb{R}^d$ we have $\boldsymbol{v}^{\perp} = \boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\boldsymbol{v}$ and $\boldsymbol{v}^{\parallel} = \boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})}\boldsymbol{v}$. Since $\boldsymbol{H}$ is symmetric,

$$\boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})}\boldsymbol{H} = \boldsymbol{H}\boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})} = \boldsymbol{0}, \qquad \text{and} \qquad \boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\boldsymbol{H} = \boldsymbol{H}\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})} = \boldsymbol{H}. \tag{31}$$

If $\boldsymbol{H}_i \in \mathcal{S}_+(\mathbb{R}^{d\times d})$ for all $i \in [n]$, then the null space of $\boldsymbol{H}$ is contained in the null space of each $\boldsymbol{H}_i$, and therefore we also have that

$$\boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})}\boldsymbol{H}_i = \boldsymbol{H}_i\boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})} = \boldsymbol{0}, \qquad \text{and} \qquad \boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\boldsymbol{H}_i = \boldsymbol{H}_i\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})} = \boldsymbol{H}_i. \tag{32}$$

B.1. Linearized dynamics

The linearized dynamics near $\boldsymbol{\theta}^*$ is

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \frac{\eta}{B}\sum_{i\in\mathfrak{B}_t}\boldsymbol{H}_i(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*) - \frac{\eta}{B}\sum_{i\in\mathfrak{B}_t}\boldsymbol{g}_i. \tag{33}$$

Therefore,

$$\begin{aligned}
\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^* &= \boldsymbol{\theta}_t - \boldsymbol{\theta}^* - \frac{\eta}{B}\sum_{i\in\mathfrak{B}_t}\boldsymbol{H}_i(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*) - \frac{\eta}{B}\sum_{i\in\mathfrak{B}_t}\boldsymbol{g}_i \\
&= \left(\boldsymbol{I} - \frac{\eta}{B}\sum_{i\in\mathfrak{B}_t}\boldsymbol{H}_i\right)(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*) - \frac{\eta}{B}\sum_{i\in\mathfrak{B}_t}\boldsymbol{g}_i.
\end{aligned} \tag{34}$$

Here we assume that the batches are chosen uniformly at random, independently across iterations.

Linearized dynamics in the orthogonal complement. Under the assumption that $\boldsymbol{H}_i \in \mathcal{S}_+(\mathbb{R}^{d\times d})$ for all $i \in [n]$, the linearized dynamics in the orthogonal complement is given by

$$\begin{aligned}
\boldsymbol{\theta}_{t+1}^{\perp} - \boldsymbol{\theta}^{*\perp} &= \boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})} \left(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^*\right) \\
&= \boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})} \left(\boldsymbol{I} - \frac{\eta}{B}\sum_{i\in\mathfrak{B}_t} \boldsymbol{H}_i\right)(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*) - \frac{\eta}{B}\sum_{i\in\mathfrak{B}_t} \boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})}\boldsymbol{g}_i \\
&= \left(\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})} - \frac{\eta}{B}\sum_{i\in\mathfrak{B}_t} \boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})}\boldsymbol{H}_i\right)(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*) - \frac{\eta}{B}\sum_{i\in\mathfrak{B}_t} \boldsymbol{g}_i^{\perp} \\
&= \left(\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})} - \frac{\eta}{B}\sum_{i\in\mathfrak{B}_t} \boldsymbol{H}_i\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})}\right)(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*) - \frac{\eta}{B}\sum_{i\in\mathfrak{B}_t} \boldsymbol{g}_i^{\perp} \\
&= \left(\boldsymbol{I} - \frac{\eta}{B}\sum_{i\in\mathfrak{B}_t} \boldsymbol{H}_i\right)\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})}(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*) - \frac{\eta}{B}\sum_{i\in\mathfrak{B}_t} \boldsymbol{g}_i^{\perp} \\
&= \left(\boldsymbol{I} - \frac{\eta}{B}\sum_{i\in\mathfrak{B}_t} \boldsymbol{H}_i\right)\left(\boldsymbol{\theta}_{t+1}^{\perp} - \boldsymbol{\theta}^{*\perp}\right) - \frac{\eta}{B}\sum_{i\in\mathfrak{B}_t} \boldsymbol{g}_i^{\perp}. \qquad (35)
\end{aligned}$$

Here, in the second step, we used (34), and in the fourth we used (32).

Linearized dynamics in the null space. Under the assumption that $\boldsymbol{H}_i \in \mathcal{S}_+(\mathbb{R}^{d\times d})$ for all $i \in [n]$, the linearized dynamics in the null space is given by

$$\begin{aligned}
\boldsymbol{\theta}_{t+1}^{\parallel} - \boldsymbol{\theta}^{*\parallel} &= \boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})} \left(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^*\right) \\
&= \boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})} \left(\boldsymbol{I} - \frac{\eta}{B}\sum_{i\in\mathfrak{B}_t} \boldsymbol{H}_i\right)(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*) - \frac{\eta}{B}\sum_{i\in\mathfrak{B}_t} \boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})}\boldsymbol{g}_i \\
&= \left(\boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})} - \frac{\eta}{B}\sum_{i\in\mathfrak{B}_t} \boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})}\boldsymbol{H}_i\right)(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*) - \frac{\eta}{B}\sum_{i\in\mathfrak{B}_t} \boldsymbol{g}_i^{\parallel} \\
&= \left(\boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})} - \boldsymbol{0}\right)(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*) - \frac{\eta}{B}\sum_{i\in\mathfrak{B}_t} \boldsymbol{g}_i^{\parallel} \\
&= \boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})}(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*) - \frac{\eta}{B}\sum_{i\in\mathfrak{B}_t} \boldsymbol{g}_i^{\parallel} \\
&= \boldsymbol{\theta}_t^{\parallel} - \boldsymbol{\theta}^{*\parallel} - \frac{\eta}{B}\sum_{i\in\mathfrak{B}_t} \boldsymbol{g}_i^{\parallel}. \qquad (36)
\end{aligned}$$

Again, in the second step, we used (34), and in the fourth we used (32). Overall,

$$\boldsymbol{\theta}_{t+1}^{\parallel} = \boldsymbol{\theta}_t^{\parallel} - \frac{\eta}{B}\sum_{i\in\mathfrak{B}_t} \boldsymbol{g}_i^{\parallel}. \qquad (37)$$

Note that if $\boldsymbol{g}_i^{\parallel} = \boldsymbol{0}$ for all $i \in [n]$ then

$$\boldsymbol{\theta}_{t+1}^{\parallel} = \boldsymbol{\theta}_t^{\parallel}. \qquad (38)$$

B.2. Mean dynamics (proof of Theorem 2)

First, we compute the mean of the linearized dynamics.

$$\begin{aligned}\boldsymbol{\mu}_{t+1} = \mathbb{E}\left[\boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}^*\right] &= \mathbb{E}\left[\left(\boldsymbol{I}-\frac{\eta}{B}\sum_{i\in\mathfrak{B}_t}\boldsymbol{H}_i\right)\left(\boldsymbol{\theta}_t-\boldsymbol{\theta}^*\right)\right]-\mathbb{E}\left[\frac{\eta}{B}\sum_{i\in\mathfrak{B}_t}\boldsymbol{g}_i\right]\\ &= \mathbb{E}\left[\left(\boldsymbol{I}-\frac{\eta}{B}\sum_{i\in\mathfrak{B}_t}\boldsymbol{H}_i\right)\mathbb{E}\left[(\boldsymbol{\theta}_t-\boldsymbol{\theta}^*)|\mathfrak{B}_t\right]\right]-\frac{\eta}{n}\sum_{i=1}^{n}\boldsymbol{g}_i\\ &= \left(\boldsymbol{I}-\eta\boldsymbol{H}\right)\mathbb{E}\left[(\boldsymbol{\theta}_t-\boldsymbol{\theta}^*)\right]\\ &= \left(\boldsymbol{I}-\eta\boldsymbol{H}\right)\boldsymbol{\mu}_t,\end{aligned} \tag{39}$$

where in the second step we used the law of total expectation, and in the third step we used (6). This system is stable if and only if the spectral radius $\rho(\boldsymbol{I}-\eta\boldsymbol{H})\leq 1$. This condition is equivalent to $\lambda_{\max}(\boldsymbol{H})\leq 2/\eta$ (see proof in, *e.g.*, Cohen et al. (2021); Mulayoff et al. (2021)), thus proving point 2 of Thm. 2.

Mean dynamics in the orthogonal complement. In a similar manner, taking the expectation of both sides of (35) and using (6), we get

$$\boldsymbol{\mu}_{t+1}^{\perp} = \left(\boldsymbol{I}-\eta\boldsymbol{H}\right)\boldsymbol{\mu}_t^{\perp}, \tag{40}$$

Note that for all $t\geq 0$,

$$\boldsymbol{\mu}_{t+1}^{\perp} = \boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\boldsymbol{\mu}_{t+1}^{\perp} = \boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\left(\boldsymbol{I}-\eta\boldsymbol{H}\right)\boldsymbol{\mu}_t^{\perp} = \left(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}-\eta\boldsymbol{H}\right)\boldsymbol{\mu}_t^{\perp}. \tag{41}$$

Namely, $\boldsymbol{\mu}_t^{\perp} = \left(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}-\eta\boldsymbol{H}\right)^t\boldsymbol{\mu}_0^{\perp}$, and thus

$$\|\boldsymbol{\mu}_t^{\perp}\| = \left\|\left(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}-\eta\boldsymbol{H}\right)^t\boldsymbol{\mu}_0^{\perp}\right\| \leq \left\|\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}-\eta\boldsymbol{H}\right\|^t\|\boldsymbol{\mu}_0^{\perp}\|. \tag{42}$$

It is easy to show that

$$\left\|\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}-\eta\boldsymbol{H}\right\| = \max_{\lambda_i(\boldsymbol{H})\neq 0}\left\{|1-\eta\lambda_i(\boldsymbol{H})|\right\}. \tag{43}$$

Therefore, if $0<\eta<2/\lambda_{\max}$, we have that $\left\|\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}-\eta\boldsymbol{H}\right\|<1$ and thus

$$\lim_{t\to\infty}\|\boldsymbol{\mu}_t^{\perp}\| \leq \lim_{t\to\infty}\left\|\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}-\eta\boldsymbol{H}\right\|^t\|\boldsymbol{\mu}_0^{\perp}\| = 0. \tag{44}$$

This proves point 3 of Thm. 2.

Mean dynamics in the null space. Taking the expectation of both sides of (36) and using (6), we obtain

$$\boldsymbol{\mu}_{t+1}^{\parallel} = \boldsymbol{\mu}_t^{\parallel}. \tag{45}$$

This demonstrates that for all $t\geq 0$,

$$\mathbb{E}\left[\boldsymbol{\theta}_t^{\parallel}-\boldsymbol{\theta}^{*\parallel}\right] = \boldsymbol{\mu}_t^{\parallel} = \boldsymbol{\mu}_0^{\parallel} = \mathbb{E}\left[\boldsymbol{\theta}_0^{\parallel}-\boldsymbol{\theta}^{*\parallel}\right], \tag{46}$$

so that

$$\mathbb{E}\left[\boldsymbol{\theta}_t^{\parallel}\right] = \mathbb{E}\left[\boldsymbol{\theta}_0^{\parallel}\right]. \tag{47}$$

This proves Point 1 of Thm. 2.

B.3. Covariance dynamics for the orthogonal complement

Before providing a complete proof for Thm. 5 (see App. B.8) and Thm. 11 (see App. B.9), we next examine the evolution over time of the covariance of the parameter vector. We start by focusing on the orthogonal complement space. Define

$$\boldsymbol{A}_t = \boldsymbol{I} - \frac{\eta}{B}\sum_{i\in\mathfrak{B}_t}\boldsymbol{H}_i \qquad \text{and} \qquad \boldsymbol{v}_t = \frac{\eta}{B}\sum_{i\in\mathfrak{B}_t}\boldsymbol{g}_i, \tag{48}$$

so that (35) can be compactly written as

$$\boldsymbol{\theta}_{t+1}^{\perp} - \boldsymbol{\theta}^{*\perp} = \boldsymbol{A}_t\left(\boldsymbol{\theta}_t^{\perp} - \boldsymbol{\theta}^{*\perp}\right) - \boldsymbol{v}_t^{\perp}. \tag{49}$$

Recall that this holds under the assumption that $\boldsymbol{H}_i \in \mathcal{S}_+(\mathbb{R}^{d\times d})$ for all $i \in [n]$. Note that $\{\boldsymbol{A}_t\}$ are i.i.d. and that $\boldsymbol{\theta}_t^{\perp}$ is constructed from $\boldsymbol{A}_0, \ldots, \boldsymbol{A}_{t-1}$, so that $\boldsymbol{\theta}_t^{\perp}$ is independent of $\boldsymbol{A}_t$. We therefore have

$$\begin{aligned}
\boldsymbol{\Sigma}_{t+1}^{\perp} &= \mathbb{E}\left[\left(\boldsymbol{\theta}_{t+1}^{\perp} - \boldsymbol{\theta}^{*\perp}\right)\left(\boldsymbol{\theta}_{t+1}^{\perp} - \boldsymbol{\theta}^{*\perp}\right)^{\mathrm{T}}\right] \\
&= \mathbb{E}\left[\left(\boldsymbol{A}_t\left(\boldsymbol{\theta}_t^{\perp} - \boldsymbol{\theta}^{*\perp}\right) - \boldsymbol{v}_t^{\perp}\right)\left(\boldsymbol{A}_t\left(\boldsymbol{\theta}_t^{\perp} - \boldsymbol{\theta}^{*\perp}\right) - \boldsymbol{v}_t^{\perp}\right)^{\mathrm{T}}\right] \\
&= \mathbb{E}\left[\boldsymbol{A}_t\left(\boldsymbol{\theta}_t^{\perp} - \boldsymbol{\theta}^{*\perp}\right)\left(\boldsymbol{\theta}_t^{\perp} - \boldsymbol{\theta}^{*\perp}\right)^{\mathrm{T}}\boldsymbol{A}_t^{\mathrm{T}}\right] - \mathbb{E}\left[\boldsymbol{A}_t\left(\boldsymbol{\theta}_t^{\perp} - \boldsymbol{\theta}^{*\perp}\right)\left(\boldsymbol{v}_t^{\perp}\right)^{\mathrm{T}}\right] \\
&\quad - \mathbb{E}\left[\boldsymbol{v}_t^{\perp}\left(\boldsymbol{\theta}_t^{\perp} - \boldsymbol{\theta}^{*\perp}\right)^{\mathrm{T}}\boldsymbol{A}_t^{\mathrm{T}}\right] + \mathbb{E}\left[\boldsymbol{v}_t^{\perp}(\boldsymbol{v}_t^{\perp})^{\mathrm{T}}\right] \\
&= \mathbb{E}\left[\boldsymbol{A}_t\mathbb{E}\left[\left(\boldsymbol{\theta}_t^{\perp} - \boldsymbol{\theta}^{*\perp}\right)\left(\boldsymbol{\theta}_t^{\perp} - \boldsymbol{\theta}^{*\perp}\right)^{\mathrm{T}}\right]\boldsymbol{A}_t^{\mathrm{T}}\right] - \mathbb{E}\left[\boldsymbol{A}_t\mathbb{E}\left[\boldsymbol{\theta}_t^{\perp} - \boldsymbol{\theta}^{*\perp}\right]\left(\boldsymbol{v}_t^{\perp}\right)^{\mathrm{T}}\right] \\
&\quad - \mathbb{E}\left[\boldsymbol{v}_t^{\perp}\mathbb{E}\left[\boldsymbol{\theta}_t^{\perp} - \boldsymbol{\theta}^{*\perp}\right]^{\mathrm{T}}\boldsymbol{A}_t^{\mathrm{T}}\right] + \boldsymbol{\Sigma}_{\boldsymbol{v}}^{\perp} \\
&= \mathbb{E}\left[\boldsymbol{A}_t\boldsymbol{\Sigma}_t^{\perp}\boldsymbol{A}_t^{\mathrm{T}}\right] - \mathbb{E}\left[\boldsymbol{A}_t\boldsymbol{\mu}_t^{\perp}(\boldsymbol{v}_t^{\perp})^{\mathrm{T}}\right] - \mathbb{E}\left[\boldsymbol{v}_t^{\perp}(\boldsymbol{\mu}_t^{\perp})^{\mathrm{T}}\boldsymbol{A}_t^{\mathrm{T}}\right] + \boldsymbol{\Sigma}_{\boldsymbol{v}}^{\perp},
\end{aligned} \tag{50}$$

where in the second equality we used (49), and in the fourth the fact that $\boldsymbol{A}_t$ is independent of $\boldsymbol{\theta}_t^{\perp}$. Using vectorization, the above equation can be written as

$$\begin{aligned}
\operatorname{vec}\left(\boldsymbol{\Sigma}_{t+1}^{\perp}\right) &= \mathbb{E}\left[\operatorname{vec}\left(\boldsymbol{A}_t\boldsymbol{\Sigma}_t^{\perp}\boldsymbol{A}_t^{\mathrm{T}}\right)\right] - \mathbb{E}\left[\operatorname{vec}\left(\boldsymbol{A}_t\boldsymbol{\mu}_t^{\perp}(\boldsymbol{v}_t^{\perp})^{\mathrm{T}}\right)\right] - \mathbb{E}\left[\operatorname{vec}\left(\boldsymbol{v}_t^{\perp}(\boldsymbol{\mu}_t^{\perp})^{\mathrm{T}}\boldsymbol{A}_t^{\mathrm{T}}\right)\right] + \operatorname{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{v}}^{\perp}\right) \\
&= \mathbb{E}\left[\boldsymbol{A}_t \otimes \boldsymbol{A}_t\right]\operatorname{vec}\left(\boldsymbol{\Sigma}_t^{\perp}\right) - \mathbb{E}\left[\boldsymbol{v}_t^{\perp} \otimes \boldsymbol{A}_t\right]\boldsymbol{\mu}_t^{\perp} - \mathbb{E}\left[\boldsymbol{A}_t \otimes \boldsymbol{v}_t^{\perp}\right]\boldsymbol{\mu}_t^{\perp} + \operatorname{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{v}}^{\perp}\right) \\
&= \boldsymbol{Q}\operatorname{vec}\left(\boldsymbol{\Sigma}_t^{\perp}\right) - \left(\mathbb{E}\left[\boldsymbol{v}_t^{\perp} \otimes \boldsymbol{A}_t\right] + \mathbb{E}\left[\boldsymbol{A}_t \otimes \boldsymbol{v}_t^{\perp}\right]\right)\boldsymbol{\mu}_t^{\perp} + \operatorname{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{v}}^{\perp}\right),
\end{aligned} \tag{51}$$

where we denoted

$$\boldsymbol{Q} \triangleq \mathbb{E}\left[\boldsymbol{A}_t \otimes \boldsymbol{A}_t\right]. \tag{52}$$

Overall, the joint dynamics of $\boldsymbol{\Sigma}_t^{\perp}$ and $\boldsymbol{\mu}_t^{\perp}$ is given by

$$\begin{pmatrix} \boldsymbol{\mu}_{t+1}^{\perp} \\ \operatorname{vec}\left(\boldsymbol{\Sigma}_{t+1}^{\perp}\right) \end{pmatrix} = \begin{pmatrix} \boldsymbol{I} - \eta\boldsymbol{H} & \boldsymbol{0} \\ -\mathbb{E}\left[\boldsymbol{v}_t^{\perp} \otimes \boldsymbol{A}_t\right] - \mathbb{E}\left[\boldsymbol{A}_t \otimes \boldsymbol{v}_t^{\perp}\right] & \boldsymbol{Q} \end{pmatrix}\begin{pmatrix} \boldsymbol{\mu}_t^{\perp} \\ \operatorname{vec}\left(\boldsymbol{\Sigma}_t^{\perp}\right) \end{pmatrix} + \begin{pmatrix} \boldsymbol{0} \\ \operatorname{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{v}}^{\perp}\right) \end{pmatrix}. \tag{53}$$

In some cases, it is easier to look at a projected version of the transition matrix. In (41) we showed that

$$\boldsymbol{\mu}_{t+1}^{\perp} = \left(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})} - \eta\boldsymbol{H}\right)\boldsymbol{\mu}_t^{\perp}. \tag{54}$$

Moreover, from (51),

$$\begin{aligned}
\operatorname{vec}\left(\boldsymbol{\Sigma}_{t+1}^{\perp}\right) &= \operatorname{vec}\left(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\boldsymbol{\Sigma}_{t+1}^{\perp}\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\right)\\
&= \left(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\right)\operatorname{vec}\left(\boldsymbol{\Sigma}_{t+1}^{\perp}\right)\\
&= \left(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\right)\left(\boldsymbol{Q}\operatorname{vec}\left(\boldsymbol{\Sigma}_{t}^{\perp}\right)-\left(\mathbb{E}\left[\boldsymbol{v}_t^{\perp}\otimes\boldsymbol{A}_t\right]+\mathbb{E}\left[\boldsymbol{A}_t\otimes\boldsymbol{v}_t^{\perp}\right]\right)\boldsymbol{\mu}_t^{\perp}+\operatorname{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{v}}^{\perp}\right)\right)\\
&= \left(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\right)\boldsymbol{Q}\operatorname{vec}\left(\boldsymbol{\Sigma}_{t}^{\perp}\right)\\
&\quad-\left(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\right)\left(\mathbb{E}\left[\boldsymbol{v}_t^{\perp}\otimes\boldsymbol{A}_t\right]+\mathbb{E}\left[\boldsymbol{A}_t\otimes\boldsymbol{v}_t^{\perp}\right]\right)\boldsymbol{\mu}_t^{\perp}+\operatorname{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{v}}^{\perp}\right). \qquad (55)
\end{aligned}$$

Therefore, the linear system in (53) can be written as

$$\begin{aligned}
&\begin{pmatrix}\boldsymbol{\mu}_{t+1}^{\perp}\\ \operatorname{vec}\left(\boldsymbol{\Sigma}_{t+1}^{\perp}\right)\end{pmatrix} =\\
&\begin{pmatrix}\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}-\eta\boldsymbol{H} & \boldsymbol{0}\\ -\left(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\right)\left(\mathbb{E}\left[\boldsymbol{v}_t^{\perp}\otimes\boldsymbol{A}_t\right]+\mathbb{E}\left[\boldsymbol{A}_t\otimes\boldsymbol{v}_t^{\perp}\right]\right) & \left(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\right)\boldsymbol{Q}\end{pmatrix}\begin{pmatrix}\boldsymbol{\mu}_{t}^{\perp}\\ \operatorname{vec}\left(\boldsymbol{\Sigma}_{t}^{\perp}\right)\end{pmatrix}\\
&\quad+\begin{pmatrix}\boldsymbol{0}\\ \operatorname{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{v}}^{\perp}\right)\end{pmatrix}. \qquad (56)
\end{aligned}$$

B.4. The transition matrix of the covariance dynamics

We now proceed to develop an explicit expression for the covariance transition matrix $\boldsymbol{Q}$ of (52). We have

$$\begin{aligned}
\boldsymbol{Q}=\mathbb{E}\left[\boldsymbol{A}_t\otimes\boldsymbol{A}_t\right] &= \mathbb{E}\left[\left(\boldsymbol{I}-\frac{\eta}{B}\sum_{i\in\mathfrak{B}_t}\boldsymbol{H}_i\right)\otimes\left(\boldsymbol{I}-\frac{\eta}{B}\sum_{i\in\mathfrak{B}_t}\boldsymbol{H}_i\right)\right]\\
&= \mathbb{E}\left[\boldsymbol{I}-\frac{\eta}{B}\sum_{i\in\mathfrak{B}_t}\left(\boldsymbol{I}\otimes\boldsymbol{H}_i+\boldsymbol{H}_i\otimes\boldsymbol{I}\right)+\frac{\eta^2}{B^2}\sum_{i,j\in\mathfrak{B}_t}\boldsymbol{H}_i\otimes\boldsymbol{H}_j\right]\\
&= \boldsymbol{I}-\eta\left(\boldsymbol{I}\otimes\boldsymbol{H}+\boldsymbol{H}\otimes\boldsymbol{I}\right)+\eta^2\mathbb{E}\left[\frac{1}{B^2}\sum_{i,j\in\mathfrak{B}_t}\boldsymbol{H}_i\otimes\boldsymbol{H}_j\right]. \qquad (57)
\end{aligned}$$

Note that

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{B^2}\sum_{i,j\in\mathfrak{B}_t}\boldsymbol{H}_i\otimes\boldsymbol{H}_j\right] &= \mathbb{E}\left[\frac{1}{B^2}\sum_{i\neq j\in\mathfrak{B}_t}\boldsymbol{H}_i\otimes\boldsymbol{H}_j+\frac{1}{B^2}\sum_{i\in\mathfrak{B}_t}\boldsymbol{H}_i\otimes\boldsymbol{H}_i\right]\\
&= \frac{1}{B^2}\times B(B-1)\mathbb{E}\left[\boldsymbol{H}_i\otimes\boldsymbol{H}_j|i\neq j\in\mathfrak{B}_t\right]+\frac{1}{B^2}\mathbb{E}\left[\sum_{i\in\mathfrak{B}_t}\boldsymbol{H}_i\otimes\boldsymbol{H}_i\right]\\
&= \frac{B-1}{B}\mathbb{E}\left[\boldsymbol{H}_i\otimes\boldsymbol{H}_j|i\neq j\in\mathfrak{B}_t\right]+\frac{1}{nB}\sum_{i=1}^{n}\boldsymbol{H}_i\otimes\boldsymbol{H}_i. \qquad (58)
\end{aligned}$$

Specifically using symmetry and (7),

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{H}_i \otimes \boldsymbol{H}_j | i \neq j \in \mathfrak{B}_t\right] &= \sum_{i\neq j=1}^{n} \frac{1}{n(n-1)} \boldsymbol{H}_i \otimes \boldsymbol{H}_j \\ &= \frac{n}{(n-1)} \frac{1}{n^2} \sum_{i\neq j=1}^{n} \boldsymbol{H}_i \otimes \boldsymbol{H}_j \\ &= \frac{n}{(n-1)} \left(\boldsymbol{H} \otimes \boldsymbol{H} - \frac{1}{n^2} \sum_{i=1}^{n} \boldsymbol{H}_i \otimes \boldsymbol{H}_i \right). \end{aligned} \tag{59}$$

Hence,

$$\begin{aligned}\frac{B-1}{B}\mathbb{E}\left[\boldsymbol{H}_i \otimes \boldsymbol{H}_j | i \neq j \in \mathfrak{B}_t\right] &= \frac{n(B-1)}{B(n-1)} \left(\boldsymbol{H} \otimes \boldsymbol{H} - \frac{1}{n^2} \sum_{i=1}^{n} \boldsymbol{H}_i \otimes \boldsymbol{H}_i \right) \\ &= \frac{n(B-1)}{B(n-1)} \boldsymbol{H} \otimes \boldsymbol{H} - \frac{B-1}{Bn(n-1)} \sum_{i=1}^{n} \boldsymbol{H}_i \otimes \boldsymbol{H}_i \\ &= \boldsymbol{H} \otimes \boldsymbol{H} - \frac{n-B}{B(n-1)} \boldsymbol{H} \otimes \boldsymbol{H} - \frac{B-1}{Bn(n-1)} \sum_{i=1}^{n} \boldsymbol{H}_i \otimes \boldsymbol{H}_i. \end{aligned} \tag{60}$$

Overall,

$$\begin{aligned}\mathbb{E}\left[\frac{1}{B^2} \sum_{i,j\in\mathfrak{B}_t} \boldsymbol{H}_i \otimes \boldsymbol{H}_j \right] &= \boldsymbol{H} \otimes \boldsymbol{H} - \frac{n-B}{B(n-1)} \boldsymbol{H} \otimes \boldsymbol{H} - \frac{B-1}{Bn(n-1)} \sum_{i=1}^{n} \boldsymbol{H}_i \otimes \boldsymbol{H}_i \\ &\quad + \frac{1}{nB} \sum_{i=1}^{n} \boldsymbol{H}_i \otimes \boldsymbol{H}_i \\ &= \boldsymbol{H} \otimes \boldsymbol{H} - \frac{n-B}{B(n-1)} \boldsymbol{H} \otimes \boldsymbol{H} + \frac{n-B}{B(n-1)} \times \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{H}_i \otimes \boldsymbol{H}_i \\ &= \boldsymbol{H} \otimes \boldsymbol{H} + \frac{n-B}{B(n-1)} \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{H}_i \otimes \boldsymbol{H}_i - \boldsymbol{H} \otimes \boldsymbol{H} \right). \end{aligned} \tag{61}$$

Therefore, we have that $\boldsymbol{Q}$ is given by

$$\begin{aligned}\boldsymbol{Q}(B,\eta) &= \boldsymbol{I} - \eta\left(\boldsymbol{I} \otimes \boldsymbol{H} + \boldsymbol{H} \otimes \boldsymbol{I}\right) + \eta^2 \boldsymbol{H} \otimes \boldsymbol{H} + \eta^2 \frac{n-B}{B(n-1)} \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{H}_i \otimes \boldsymbol{H}_i - \boldsymbol{H} \otimes \boldsymbol{H} \right) \\ &= (\boldsymbol{I} - \eta\boldsymbol{H}) \otimes (\boldsymbol{I} - \eta\boldsymbol{H}) + \frac{n-B}{B(n-1)} \frac{\eta^2}{n} \sum_{i=1}^{n} (\boldsymbol{H}_i \otimes \boldsymbol{H}_i - \boldsymbol{H} \otimes \boldsymbol{H}), \end{aligned} \tag{62}$$

which reproduce the result of (10). Here we give an alternative form of $\boldsymbol{Q}$, which is useful in many derivations. We set

$$p = \frac{n-B}{B(n-1)}, \tag{63}$$

and continue from the first line in (62).

$$\begin{aligned}
\boldsymbol{Q}(B,\eta) &= \boldsymbol{I} - \eta\left(\boldsymbol{I}\otimes\boldsymbol{H} + \boldsymbol{H}\otimes\boldsymbol{I}\right) + \eta^2\boldsymbol{H}\otimes\boldsymbol{H} + \eta^2 p\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{H}_i\otimes\boldsymbol{H}_i - \boldsymbol{H}\otimes\boldsymbol{H}\right)\\
&= \boldsymbol{I} - \eta\left(\boldsymbol{I}\otimes\boldsymbol{H} + \boldsymbol{H}\otimes\boldsymbol{I}\right) + (1-p)\eta^2\boldsymbol{H}\otimes\boldsymbol{H} + p\times\frac{\eta^2}{n}\sum_{i=1}^{n}\boldsymbol{H}_i\otimes\boldsymbol{H}_i\\
&= (1-p)\boldsymbol{I} - (1-p)\eta\left(\boldsymbol{I}\otimes\boldsymbol{H} + \boldsymbol{H}\otimes\boldsymbol{I}\right) + (1-p)\eta^2\boldsymbol{H}\otimes\boldsymbol{H}\\
&\quad + p\boldsymbol{I} - p\eta\left(\boldsymbol{I}\otimes\boldsymbol{H} + \boldsymbol{H}\otimes\boldsymbol{I}\right) + p\times\frac{\eta^2}{n}\sum_{i=1}^{n}\boldsymbol{H}_i\otimes\boldsymbol{H}_i\\
&= (1-p)\left[\boldsymbol{I} - \eta\left(\boldsymbol{I}\otimes\boldsymbol{H} + \boldsymbol{H}\otimes\boldsymbol{I}\right) + \eta^2\boldsymbol{H}\otimes\boldsymbol{H}\right]\\
&\quad + p\left[\boldsymbol{I} - \eta\left(\boldsymbol{I}\otimes\boldsymbol{H} + \boldsymbol{H}\otimes\boldsymbol{I}\right) + \frac{\eta^2}{n}\sum_{i=1}^{n}\boldsymbol{H}_i\otimes\boldsymbol{H}_i\right]\\
&= (1-p)\left(\boldsymbol{I} - \eta\boldsymbol{H}\right)\otimes\left(\boldsymbol{I} - \eta\boldsymbol{H}\right)\\
&\quad + p\left[\boldsymbol{I} - \eta\left(\boldsymbol{I}\otimes\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{H}_i\right) + \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{H}_i\right)\otimes\boldsymbol{I}\right) + \frac{\eta^2}{n}\sum_{i=1}^{n}\boldsymbol{H}_i\otimes\boldsymbol{H}_i\right]\\
&= (1-p)\left(\boldsymbol{I} - \eta\boldsymbol{H}\right)\otimes\left(\boldsymbol{I} - \eta\boldsymbol{H}\right) + p\frac{1}{n}\sum_{i=1}^{n}\left[\boldsymbol{I} - \eta\left(\boldsymbol{I}\otimes\boldsymbol{H}_i + \boldsymbol{H}_i\otimes\boldsymbol{I}\right) + \eta^2\boldsymbol{H}_i\otimes\boldsymbol{H}_i\right]\\
&= (1-p)(\boldsymbol{I} - \eta\boldsymbol{H})\otimes(\boldsymbol{I} - \eta\boldsymbol{H}) + p\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{I} - \eta\boldsymbol{H}_i)\otimes(\boldsymbol{I} - \eta\boldsymbol{H}_i), \qquad (64)
\end{aligned}$$

where in the third step we add and subtract $p\boldsymbol{I} - p\eta\left(\boldsymbol{I}\otimes\boldsymbol{H} + \boldsymbol{H}\otimes\boldsymbol{I}\right)$, and in the fifth we used (7). Another useful form is the following. First, observe that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{H}_i\otimes\boldsymbol{H}_i - \boldsymbol{H}\otimes\boldsymbol{H} &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{H}_i\otimes\boldsymbol{H}_i - \boldsymbol{H}\otimes\boldsymbol{H} - \boldsymbol{H}\otimes\boldsymbol{H} + \boldsymbol{H}\otimes\boldsymbol{H}\\
&= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{H}_i\otimes\boldsymbol{H}_i - \boldsymbol{H}\otimes\boldsymbol{H} - \boldsymbol{H}\otimes\boldsymbol{H} + \boldsymbol{H}\otimes\boldsymbol{H}\\
&= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{H}_i\otimes\boldsymbol{H}_i - \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{H}_i\right)\otimes\boldsymbol{H} - \boldsymbol{H}\otimes\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{H}_i\right) + \boldsymbol{H}\otimes\boldsymbol{H}\\
&= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{H}_i\otimes\boldsymbol{H}_i - \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{H}_i\otimes\boldsymbol{H} - \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{H}\otimes\boldsymbol{H}_i + \boldsymbol{H}\otimes\boldsymbol{H}\\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{H}_i - \boldsymbol{H}\right)\otimes\left(\boldsymbol{H}_i - \boldsymbol{H}\right) \qquad (65)
\end{aligned}$$

Then, starting from the first line in (62) and using (65), we have

$$\begin{aligned}\boldsymbol{Q}(B,\eta) &= \boldsymbol{I}-\eta\left(\boldsymbol{I}\otimes\boldsymbol{H}+\boldsymbol{H}\otimes\boldsymbol{I}\right)+\eta^2\boldsymbol{H}\otimes\boldsymbol{H}+p\eta^2\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{H}_i\otimes\boldsymbol{H}_i-\boldsymbol{H}\otimes\boldsymbol{H}\right)\\ &= (\boldsymbol{I}-\eta\boldsymbol{H})\otimes(\boldsymbol{I}-\eta\boldsymbol{H})+\eta^2 p\left(\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{H}_i-\boldsymbol{H})\otimes(\boldsymbol{H}_i-\boldsymbol{H})\right)\end{aligned} \tag{66}$$

Finally, we give the $\boldsymbol{Q}$ in terms of $\boldsymbol{C}$ and $\boldsymbol{D}$. Again, we start from the first line in (62).

$$\begin{aligned}\boldsymbol{Q} &= \boldsymbol{I}-\eta\left(\boldsymbol{I}\otimes\boldsymbol{H}+\boldsymbol{H}\otimes\boldsymbol{I}\right)+\eta^2\boldsymbol{H}\otimes\boldsymbol{H}+\eta^2 p\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{H}_i\otimes\boldsymbol{H}_i-\boldsymbol{H}\otimes\boldsymbol{H}\right)\\ &= \boldsymbol{I}-\eta\left(\boldsymbol{I}\otimes\boldsymbol{H}+\boldsymbol{H}\otimes\boldsymbol{I}\right)+\eta^2\left[(1-p)\times\boldsymbol{H}\otimes\boldsymbol{H}+p\times\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{H}_i\otimes\boldsymbol{H}_i\right]\\ &= \boldsymbol{I}+2\eta\boldsymbol{C}+\eta^2\boldsymbol{D}.\end{aligned} \tag{67}$$

B.5. Covariance matrix of the gradient noise

We now develop an explicit expression for the covariance matrix of the gradient noise $\boldsymbol{v}$ of (48). We have

$$\begin{aligned}\boldsymbol{\Sigma}_{\boldsymbol{v}} = \mathbb{E}\left[\boldsymbol{v}_t\boldsymbol{v}_t^{\mathrm{T}}\right] &= \left(\frac{\eta}{B}\right)^2\mathbb{E}\left[\sum_{i,j\in\mathfrak{B}_t}\boldsymbol{g}_i\boldsymbol{g}_j^{\mathrm{T}}\right]\\ &= \left(\frac{\eta}{B}\right)^2\mathbb{E}\left[\sum_{i\neq j\in\mathfrak{B}_t}\boldsymbol{g}_i\boldsymbol{g}_j^{\mathrm{T}}+\sum_{i\in\mathfrak{B}_t}\boldsymbol{g}_i\boldsymbol{g}_i^{\mathrm{T}}\right]\\ &= \left(\frac{\eta}{B}\right)^2\left(B(B-1)\mathbb{E}\left[\boldsymbol{g}_i\boldsymbol{g}_j^{\mathrm{T}}\middle| i\neq j\in\mathfrak{B}_t\right]+\frac{B}{n}\sum_{i=1}^{n}\boldsymbol{g}_i\boldsymbol{g}_i^{\mathrm{T}}\right).\end{aligned} \tag{68}$$

Observe that

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{g}_i\boldsymbol{g}_j^{\mathrm{T}}\middle| i\neq j\in\mathfrak{B}_t\right] &= \sum_{i\neq j=1}^{n}\frac{1}{n(n-1)}\boldsymbol{g}_i\boldsymbol{g}_j^{\mathrm{T}}\\ &= \frac{1}{n(n-1)}\left(\sum_{i,j=1}^{n}\boldsymbol{g}_i\boldsymbol{g}_j^{\mathrm{T}}-\sum_{i=1}^{n}\boldsymbol{g}_i\boldsymbol{g}_i^{\mathrm{T}}\right)\\ &= \frac{1}{n(n-1)}\left(\left(\sum_{i=1}^{n}\boldsymbol{g}_i\right)\left(\sum_{i=1}^{n}\boldsymbol{g}_i\right)^{\mathrm{T}}-\sum_{i=1}^{n}\boldsymbol{g}_i\boldsymbol{g}_i^{\mathrm{T}}\right)\\ &= -\frac{1}{n(n-1)}\sum_{i=1}^{n}\boldsymbol{g}_i\boldsymbol{g}_i^{\mathrm{T}},\end{aligned} \tag{69}$$

where in the last step we used (6). Thus,

$$\begin{aligned}\boldsymbol{\Sigma}_{\boldsymbol{v}} &= \left(\frac{\eta}{B}\right)^2\left(\frac{B}{n}-\frac{B(B-1)}{n(n-1)}\right)\sum_{i=1}^{n}\boldsymbol{g}_i\boldsymbol{g}_i^{\mathrm{T}}\\ &= \left(\frac{\eta}{B}\right)^2 \times \frac{B(n-B)}{n(n-1)}\sum_{i=1}^{n}\boldsymbol{g}_i\boldsymbol{g}_i^{\mathrm{T}}\\ &= \eta^2\frac{n-B}{B(n-1)} \times \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}_i\boldsymbol{g}_i^{\mathrm{T}}\\ &= \eta^2 p\boldsymbol{\Sigma}_{\boldsymbol{g}}, \end{aligned} \tag{70}$$

where we denoted

$$\boldsymbol{\Sigma}_{\boldsymbol{g}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}_i\boldsymbol{g}_i^{\mathrm{T}}. \tag{71}$$

B.6. The Null spaces of $\boldsymbol{C}$, $\boldsymbol{D}$ and $\boldsymbol{E}$

Let us now analyze the relation between the null spaces of $\boldsymbol{C}$, $\boldsymbol{D}$ and $\boldsymbol{E}$. First, it is easy to see that under the assumption that $\boldsymbol{H}_i \in \mathcal{S}_+(\mathbb{R}^{d\times d})$ for all $i \in [n]$,

$$\mathcal{N}(\boldsymbol{H}) = \bigcap_{i=1}^{n}\mathcal{N}(\boldsymbol{H}_i), \tag{72}$$

where $\mathcal{N}(\cdot)$ denotes null space of a matrix. Here we show the following.

Lemma 15 *Assume that $\boldsymbol{H}_i \in \mathcal{S}_+(\mathbb{R}^{d\times d})$ for all $i \in [n]$ and let*

$$\begin{aligned}\boldsymbol{C} &= \frac{1}{2}\boldsymbol{H}\oplus\boldsymbol{H},\\ \boldsymbol{D} &= (1-p)\,\boldsymbol{H}\otimes\boldsymbol{H} + p\,\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{H}_i\otimes\boldsymbol{H}_i,\\ \boldsymbol{E} &= \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{H}_i-\boldsymbol{H})\otimes(\boldsymbol{H}_i-\boldsymbol{H}). \end{aligned} \tag{73}$$

Then $\mathcal{N}(\boldsymbol{C}) \subseteq \mathcal{N}(\boldsymbol{D})$ and $\mathcal{N}(\boldsymbol{C}) \subseteq \mathcal{N}(\boldsymbol{E})$.

Proof Let $\boldsymbol{u} \in \mathcal{N}(\boldsymbol{C})$ and denote $\boldsymbol{U} = \mathrm{vec}^{-1}(\boldsymbol{u})$, then

$$\boldsymbol{0} = 2\boldsymbol{C}\boldsymbol{u} = \boldsymbol{H}\oplus\boldsymbol{H}\boldsymbol{u} = (\boldsymbol{H}\otimes\boldsymbol{I}+\boldsymbol{I}\otimes\boldsymbol{H})\boldsymbol{u}. \tag{74}$$

In matrix form we get

$$\boldsymbol{U}\boldsymbol{H}+\boldsymbol{H}\boldsymbol{U} = \boldsymbol{0}. \tag{75}$$

Let us take the Frobenius norm, then

$$\|\boldsymbol{U}\boldsymbol{H}+\boldsymbol{H}\boldsymbol{U}\|_{\mathrm{F}}^2 = \|\boldsymbol{U}\boldsymbol{H}\|_{\mathrm{F}}^2 + \|\boldsymbol{H}\boldsymbol{U}\|_{\mathrm{F}}^2 + 2\mathrm{Tr}\left((\boldsymbol{U}\boldsymbol{H})^{\mathrm{T}}\boldsymbol{H}\boldsymbol{U}\right) = 0, \tag{76}$$

where

$$\mathrm{Tr}\left((\boldsymbol{U}\boldsymbol{H})^{\mathrm{T}}\boldsymbol{H}\boldsymbol{U}\right) = \mathrm{Tr}\left(\boldsymbol{H}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{U}\right) = \mathrm{Tr}\left(\boldsymbol{H}^{\frac{1}{2}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{U}\boldsymbol{H}^{\frac{1}{2}}\right) \geq 0 \tag{77}$$

because $\boldsymbol{H}^{\frac{1}{2}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{U}\boldsymbol{H}^{\frac{1}{2}} = (\boldsymbol{H}^{\frac{1}{2}}\boldsymbol{U}\boldsymbol{H}^{\frac{1}{2}})^{\mathrm{T}}(\boldsymbol{H}^{\frac{1}{2}}\boldsymbol{U}\boldsymbol{H}^{\frac{1}{2}})$ is PSD. This implies that

$$\|\boldsymbol{U}\boldsymbol{H}\|_{\mathrm{F}}^2 = \|\boldsymbol{H}\boldsymbol{U}\|_{\mathrm{F}}^2 = 0. \tag{78}$$

Thus, $\boldsymbol{u} \in \mathcal{N}(\boldsymbol{C})$ if and only if $\boldsymbol{U}\boldsymbol{H} = \boldsymbol{0}$ and $\boldsymbol{H}\boldsymbol{U} = \boldsymbol{0}$. Since the null space of $\boldsymbol{H}$ is the intersection of $\{\boldsymbol{H}_i\}$ (72) , we have that $\boldsymbol{U}$ also satisfies $\boldsymbol{H}_i\boldsymbol{U} = \boldsymbol{U}\boldsymbol{H}_i = \boldsymbol{0}$ for all $i \in [n]$. Now,

$$\boldsymbol{D}\boldsymbol{u} = (1-p)\,\boldsymbol{H}\otimes\boldsymbol{H}\boldsymbol{u} + p\,\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{H}_i\otimes\boldsymbol{H}_i\boldsymbol{u}, \tag{79}$$

and in matrix form,

$$\mathrm{vec}^{-1}(\boldsymbol{D}\boldsymbol{u}) = (1-p)\,\boldsymbol{H}\boldsymbol{U}\boldsymbol{H} + p\,\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{H}_i\boldsymbol{U}\boldsymbol{H}_i = \boldsymbol{0}. \tag{80}$$

Namely, $\boldsymbol{u} \in \mathcal{N}(\boldsymbol{D})$. Similarly,

$$\boldsymbol{E}\boldsymbol{u} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{H}_i - \boldsymbol{H})\otimes(\boldsymbol{H}_i - \boldsymbol{H})\boldsymbol{u}, \tag{81}$$

and in matrix form,

$$\mathrm{vec}^{-1}(\boldsymbol{E}\boldsymbol{u}) = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{H}_i - \boldsymbol{H})\boldsymbol{U}(\boldsymbol{H}_i - \boldsymbol{H}) = \boldsymbol{0}. \tag{82}$$

Namely, $\boldsymbol{u} \in \mathcal{N}(\boldsymbol{E})$. ■

B.7. Positivity of $\boldsymbol{C}$ and $\boldsymbol{D}$

B.7.1. POSITIVITY OF $\boldsymbol{C}$

The eigenvalues of a Kronecker *sum* are the pairwise sums of the eigenvalues of the summands (Laub, 2004, Thm. 13.16). In App. J we explicitly show this for $\boldsymbol{C}$, where we derive that the eigenvalues of $\boldsymbol{C} = \frac{1}{2}\boldsymbol{H}\oplus\boldsymbol{H}$ are $\frac{1}{2}\big(\lambda_i(\boldsymbol{H}) + \lambda_j(\boldsymbol{H})\big),\ i = 1,\ldots,d,\ j = 1,\ldots,d$. Note that $\boldsymbol{H}$ is the Hessian of the loss at a minimum, and is therefore PSD. Therefore all eigenvalues of $\boldsymbol{H}$ are nonnegative, and as a consequence, the eigenvalues of $\boldsymbol{C}$ are nonnegative, *i.e.,* $\boldsymbol{C}$ is PSD.

B.7.2. POSITIVITY OF $\boldsymbol{D}$

The eigenvalues of a Kronecker *product* are the pairwise products of the eigenvalues of the multiplicands (Laub, 2004, Thm. 13.12). This property asserts that for any PSD matrix $\boldsymbol{M}$, namely with nonnegative eigenvalues, the Kronecker product $\boldsymbol{M}\otimes\boldsymbol{M}$ is PSD. Note that $\boldsymbol{D}$ is defined as

$$\boldsymbol{D} = (1-p)\boldsymbol{H}\otimes\boldsymbol{H} + p\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{H}_i\otimes\boldsymbol{H}_i, \tag{83}$$

with $p \in [0,1]$. In our settings, *i.e.,* regular and interpolating minima, we consider Hessian matrices $\{\boldsymbol{H}_i\}$ that are PSD. By the property above, all $\{\boldsymbol{H}_i\otimes\boldsymbol{H}_i\}$ are PSD, and also $\{\boldsymbol{H}\otimes\boldsymbol{H}\}$ is PSD. Therefore, $\boldsymbol{D}$ is a convex combination of PSD matrices, which is PSD.

B.8. Proof of Theorem 5

We are now ready to prove Thm. 5.

First statement. In (38) we showed that for interpolating minima $\boldsymbol{\theta}_{t+1}^{\parallel} = \boldsymbol{\theta}_t^{\parallel}$, which completes the proof for the first statement of the theorem.

Second statement. Ma and Ying (2021) showed that the second moment $\boldsymbol{\Sigma}_t = [(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*)(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*)^{\mathrm{T}}]$ for interpolating minima evolves over time as

$$\mathrm{vec}\left(\boldsymbol{\Sigma}_{t+1}\right) = \boldsymbol{Q}\ \mathrm{vec}\left(\boldsymbol{\Sigma}_t\right), \tag{84}$$

where $\boldsymbol{Q}$ is given in (10). Since $\boldsymbol{\Sigma}_t$ is PSD by definition, we only care about the effect of $\boldsymbol{Q}$ on vectorizations of PSD matrices. Therefore, we have that $\{\boldsymbol{\Sigma}_t\}$ are bounded if and only if (see proof in (Ma and Ying, 2021))

$$\max_{\boldsymbol{\Sigma}\in\mathcal{S}_+(\mathbb{R}^{d\times d})} \frac{\|\boldsymbol{Q}(\eta,B)\ \mathrm{vec}\left(\boldsymbol{\Sigma}\right)\|}{\|\boldsymbol{\Sigma}\|_{\mathrm{F}}} \leq 1. \tag{85}$$

To obtain the stability threshold of SGD in the mean square sense we first rearrange the terms in $\boldsymbol{Q}$ as (see (64))

$$\boldsymbol{Q}(\eta,B) = (1-p)\times(\boldsymbol{I}-\eta\boldsymbol{H})\otimes(\boldsymbol{I}-\eta\boldsymbol{H}) + p\times\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{I}-\eta\boldsymbol{H}_i)\otimes(\boldsymbol{I}-\eta\boldsymbol{H}_i). \tag{86}$$

Here we explicitly see that $\boldsymbol{Q}$ can be written as a sum of Kronecker products, where each product is of a symmetric matrix with itself, as required by Thm. 14. Applying this theorem, we have that the spectral radius of $\boldsymbol{Q}$ equals its top eigenvalue, and the corresponding top eigenvector is a vectorization of a PSD matrix. Note that since $\boldsymbol{Q}$ is symmetric, its spectral radius $\rho(\boldsymbol{Q})$ is given by the *unconstrained* optimization problem

$$\rho(\boldsymbol{Q}) = \max_{\boldsymbol{\Sigma}\in\mathbb{R}^{d\times d}} \frac{\|\boldsymbol{Q}(\eta,B)\ \mathrm{vec}\left(\boldsymbol{\Sigma}\right)\|}{\|\boldsymbol{\Sigma}\|_{\mathrm{F}}}. \tag{87}$$

Theorem 14 tells us that the top eigenvector of $\boldsymbol{Q}$ maximizes this unconstrained problem, and more importantly, it always corresponds to a PSD matrix. Therefore, this top eigenvector also maximizes the objective while restricting to the subset of PSD matrices, which is given by the constraint in (85). Thus, we have that the maximizer for the constrained optimization problem in (85) is, in fact, the top eigenvalue of $\boldsymbol{Q}$. Hence, the linear system is stable if and only if $\lambda_{\max}(\boldsymbol{Q}) \leq 1$. Writing $\boldsymbol{Q}$ in terms of $\boldsymbol{C}$ and $\boldsymbol{D}$ gives (see (67))

$$\boldsymbol{Q} = \boldsymbol{I} - 2\eta\boldsymbol{C} + \eta^2\boldsymbol{D}. \tag{88}$$

Because $\boldsymbol{Q}$ is symmetric, the condition $\lambda_{\max}(\boldsymbol{Q}) \leq 1$ is equivalent to the requirement that $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{u} \leq 1$ for all $\boldsymbol{u} \in \mathbb{S}^{d^2-1}$. In App. B.6 we show that $\mathcal{N}(\boldsymbol{C}) \subseteq \mathcal{N}(\boldsymbol{D})$. Therefore, if $\boldsymbol{u} \in \mathcal{N}(\boldsymbol{C})$ then also $\boldsymbol{u} \in \mathcal{N}(\boldsymbol{D})$ and we get

$$\boldsymbol{u}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{u} = 1 - 2\eta\boldsymbol{u}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{u} + \eta^2\boldsymbol{u}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{u} = 1. \tag{89}$$

Namely, directions in the null space of $\boldsymbol{C}$ do not impose any constraint on the learning rate, and thus can be ignored. Additionally, if $\boldsymbol{u} \in \mathcal{N}(\boldsymbol{D})$ but $\boldsymbol{u} \notin \mathcal{N}(\boldsymbol{C})$, then

$$\boldsymbol{u}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{u} = 1 - 2\eta\boldsymbol{u}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{u} + \eta^2\boldsymbol{u}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{u} = 1 - 2\eta\boldsymbol{u}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{u} \leq 1, \tag{90}$$

holds for all $\eta \geq 0$, because $\boldsymbol{C}$ is PSD (see App. B.7). Now,

$$\boldsymbol{u}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{u} = 1 - 2\eta\boldsymbol{u}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{u} + \eta^2\boldsymbol{u}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{u} \leq 1 \tag{91}$$

holds for all $\boldsymbol{u} \notin \mathcal{N}(\boldsymbol{D})$ (which also results in $\boldsymbol{u} \notin \mathcal{N}(\boldsymbol{C})$), if and only if

$$\forall \boldsymbol{u} \notin \mathcal{N}(\boldsymbol{D}) \qquad \eta\boldsymbol{u}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{u} \leq 2\boldsymbol{u}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{u}. \tag{92}$$

Since $\boldsymbol{D}$ is PSD (see App. B.7), and we assume that $\boldsymbol{u} \notin \mathcal{N}(\boldsymbol{D})$, we can divide both sides of this equation by $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{u} > 0$ to get a condition on the learning rate as

$$0 \leq \eta \leq 2 \inf_{\boldsymbol{u}\in\mathbb{S}^{d^2-1}:\boldsymbol{u}\notin\mathcal{N}(\boldsymbol{D})} \left\{\frac{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{u}}{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{u}}\right\}. \tag{93}$$

Therefore, the stability threshold η^*_{var} is given by

$$\eta^*_{\text{var}} = 2 \inf_{\boldsymbol{u}\in\mathbb{S}^{d^2-1}:\boldsymbol{u}\notin\mathcal{N}(\boldsymbol{D})} \left\{\frac{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{u}}{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{u}}\right\} = 2\left(\sup_{\boldsymbol{u}\in\mathbb{S}^{d^2-1}:\boldsymbol{u}\notin\mathcal{N}(\boldsymbol{D})} \left\{\frac{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{u}}{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{u}}\right\}\right)^{-1}. \tag{94}$$

Note that the norm of $\boldsymbol{u}$ has no effect, and therefore we can remove the constraint $\boldsymbol{u} \in \mathbb{S}^{d^2-1}$. Additionally, we can also relax the constraint $\boldsymbol{u} \notin \mathcal{N}(\boldsymbol{D})$ to $\boldsymbol{u} \notin \mathcal{N}(\boldsymbol{C})$, because the supremum in (94) is over a nonnegative function (both $\boldsymbol{C}$ and $\boldsymbol{D}$ are PSD, see App. B.7), and will not be affected by adding to the domain points at which the function vanishes. Since $\mathcal{N}(\boldsymbol{C}) \subseteq \mathcal{N}(\boldsymbol{D})$ we have that $\mathcal{N}^{\perp}(\boldsymbol{D}) \subseteq \mathcal{N}^{\perp}(\boldsymbol{C})$ and therefore

$$\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{C})}\boldsymbol{D} = \boldsymbol{D}\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{C})} = \boldsymbol{D}, \tag{95}$$

where $\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{C})}$ is the projection matrix onto the orthogonal complement of the null space of $\boldsymbol{C}$. Additionally, $\boldsymbol{C}$ is PSD (see App. B.7), and therefore $\boldsymbol{C}^{\frac{1}{2}}$ exists and is also PSD, so that

$$\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{C})} = \left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger}\boldsymbol{C}^{\frac{1}{2}} = \boldsymbol{C}^{\frac{1}{2}}\left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger}. \tag{96}$$

Therefore,

$$\begin{aligned}
\sup_{\boldsymbol{u}\in\mathbb{S}^{d^2-1}:\boldsymbol{u}\notin\mathcal{N}(\boldsymbol{D})} \left\{\frac{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{u}}{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{u}}\right\} &= \sup_{\boldsymbol{u}\notin\mathcal{N}(\boldsymbol{C})} \left\{\frac{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{u}}{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{u}}\right\} \\
&= \sup_{\boldsymbol{u}\notin\mathcal{N}(\boldsymbol{C})} \left\{\frac{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{C})}\boldsymbol{D}\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{C})}\boldsymbol{u}}{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{u}}\right\} \\
&= \sup_{\boldsymbol{u}\notin\mathcal{N}(\boldsymbol{C})} \left\{\frac{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{C}^{\frac{1}{2}}\left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger}\boldsymbol{D}\left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger}\boldsymbol{C}^{\frac{1}{2}}\boldsymbol{u}}{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{C}^{\frac{1}{2}}\boldsymbol{C}^{\frac{1}{2}}\boldsymbol{u}}\right\} \\
&= \sup_{\boldsymbol{u}\notin\mathcal{N}(\boldsymbol{C})} \left\{\frac{\left(\boldsymbol{C}^{\frac{1}{2}}\boldsymbol{u}\right)^{\mathrm{T}}\left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger}\boldsymbol{D}\left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger}\left(\boldsymbol{C}^{\frac{1}{2}}\boldsymbol{u}\right)}{\left(\boldsymbol{C}^{\frac{1}{2}}\boldsymbol{u}\right)^{\mathrm{T}}\left(\boldsymbol{C}^{\frac{1}{2}}\boldsymbol{u}\right)}\right\},
\end{aligned} \tag{97}$$

where in the second step we used (95), and in the third step we used (96). By a simple change of variables $\boldsymbol{y} = \boldsymbol{C}^{\frac{1}{2}}\boldsymbol{u} \in \mathcal{N}^{\perp}(\boldsymbol{C})$ we get

$$\max_{\boldsymbol{y}\in\mathcal{N}^{\perp}(\boldsymbol{C})} \left\{ \frac{\boldsymbol{y}^{\mathrm{T}} \left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger} \boldsymbol{D} \left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger} \boldsymbol{y}}{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}} \right\} = \max_{\boldsymbol{y}\in\mathbb{R}^{d^2}} \left\{ \frac{\boldsymbol{y}^{\mathrm{T}} \left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger} \boldsymbol{D} \left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger} \boldsymbol{y}}{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}} \right\}$$
$$= \lambda_{\max}\left(\left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger} \boldsymbol{D} \left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger} \right), \tag{98}$$

where in the first step we used the fact that adding to $\boldsymbol{y}$ a component in $\mathcal{N}(\boldsymbol{C})$ will increase the denominator by $\|\boldsymbol{P}_{\mathcal{N}(\boldsymbol{C})}\boldsymbol{y}\|^2$ but will not affect the numerator. Namely, the optimum cannot be attained by $\boldsymbol{y} \notin \mathcal{N}^{\perp}(\boldsymbol{C})$. Now, let $(\lambda_i,\ \boldsymbol{y}_i)$ be an eigenpair of $(\boldsymbol{C}^{\frac{1}{2}})^{\dagger}\boldsymbol{D}(\boldsymbol{C}^{\frac{1}{2}})^{\dagger}$, then we have

$$\lambda_i \boldsymbol{y}_i = \left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger} \boldsymbol{D} \left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger} \boldsymbol{y}_i. \tag{99}$$

Since we only care about nonzero eigenvalues, we can assume that $\lambda_i \neq 0$, and therefore $\boldsymbol{y}_i \notin \mathcal{N}(\boldsymbol{C})$. Multiplying by $(\boldsymbol{C}^{\frac{1}{2}})^{\dagger}$ from the left we get

$$\lambda_i \left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger} \boldsymbol{y}_i = \left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger} \left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger} \boldsymbol{D} \left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger} \boldsymbol{y}_i = \boldsymbol{C}^{\dagger}\boldsymbol{D} \left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger} \boldsymbol{y}_i. \tag{100}$$

Namely, $(\boldsymbol{C}^{\frac{1}{2}})^{\dagger}\boldsymbol{y}_i \neq \boldsymbol{0}$ is an eigenvector of $\boldsymbol{C}^{\dagger}\boldsymbol{D}$ with eigenvalue λ_i. Thus we have that if $\lambda_i \neq 0$ is an eigenvalue of $(\boldsymbol{C}^{\frac{1}{2}})^{\dagger}\boldsymbol{D}(\boldsymbol{C}^{\frac{1}{2}})^{\dagger}$, then it is also an eigenvalue of $\boldsymbol{C}^{\dagger}\boldsymbol{D}$. Similarly, we can prove vice versa, *i.e.,* that if $\lambda_i \neq 0$ is an eigenvalue of $\boldsymbol{C}^{\dagger}\boldsymbol{D}$, then it is also an eigenvalue of $(\boldsymbol{C}^{\frac{1}{2}})^{\dagger}\boldsymbol{D}(\boldsymbol{C}^{\frac{1}{2}})^{\dagger}$. This means that $(\boldsymbol{C}^{\frac{1}{2}})^{\dagger}\boldsymbol{D}(\boldsymbol{C}^{\frac{1}{2}})^{\dagger}$ and $\boldsymbol{C}^{\dagger}\boldsymbol{D}$ have the same eigenvalues. Therefore,

$$\lambda_{\max}\left(\left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger} \boldsymbol{D} \left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger} \right) = \lambda_{\max}\left(\boldsymbol{C}^{\dagger}\boldsymbol{D}\right). \tag{101}$$

Overall, we showed that the condition in (11) is equivalent to

$$\eta \leq \frac{2}{\lambda_{\max}\left(\boldsymbol{C}^{\dagger}\boldsymbol{D}\right)}. \tag{102}$$

This completes the proof for the second statement of the theorem.

Third statement. For the third statement of the theorem, note that from (55) we have that $\forall t \geq 0$

$$\mathrm{vec}\left(\boldsymbol{\Sigma}_{t+1}^{\perp}\right) = \left(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})} \otimes \boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\right) \boldsymbol{Q}\, \mathrm{vec}\left(\boldsymbol{\Sigma}_{t}^{\perp}\right). \tag{103}$$

Namely, $\mathrm{vec}\left(\boldsymbol{\Sigma}_{t}^{\perp}\right) = \left[(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})} \otimes \boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})})\boldsymbol{Q}\right]^{t}\mathrm{vec}\left(\boldsymbol{\Sigma}_{0}^{\perp}\right)$, and thus

$$\left\|\mathrm{vec}\left(\boldsymbol{\Sigma}_{t}^{\perp}\right)\right\| = \left\|\left[(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})} \otimes \boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})})\boldsymbol{Q}\right]^{t}\mathrm{vec}\left(\boldsymbol{\Sigma}_{0}^{\perp}\right)\right\| \leq \left\|(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})} \otimes \boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})})\boldsymbol{Q}\right\|^{t} \left\|\mathrm{vec}\left(\boldsymbol{\Sigma}_{0}^{\perp}\right)\right\|. \tag{104}$$

Here

$$\begin{aligned}\big(\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})}\big)\boldsymbol{Q} &= \big(\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})}\big)\\ &\qquad\left[(1-p)(\boldsymbol{I}-\eta\boldsymbol{H})\otimes(\boldsymbol{I}-\eta\boldsymbol{H})+p\times\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{I}-\eta\boldsymbol{H}_i)\otimes(\boldsymbol{I}-\eta\boldsymbol{H}_i)\right]\\ &= (1-p)\big(\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})}-\eta\boldsymbol{H}\big)\otimes\big(\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})}-\eta\boldsymbol{H}\big)\\ &\quad+p\times\frac{1}{n}\sum_{i=1}^{n}\big(\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})}-\eta\boldsymbol{H}_i\big)\otimes\big(\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})}-\eta\boldsymbol{H}_i\big),\end{aligned} \tag{105}$$

where we used (64) for the value of $\boldsymbol{Q}$. We see that $\big(\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})}\big)\boldsymbol{Q}$ is a sum of Kronecker products, where each product is a symmetric matrix multiplied by itself. This means that Thm. 14 applies to $\big(\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})}\big)\boldsymbol{Q}$, and thus we have that $\big\|\big(\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})}\big)\boldsymbol{Q}\big\| = \lambda_{\max}\big((\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})})\boldsymbol{Q}\big)$. Moreover, it is easy to show that $\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})} = \boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}$, and $\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}\boldsymbol{C} = \boldsymbol{C}\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}$. Combining this with (67), we have

$$(\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})})\boldsymbol{Q} = \boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})} - 2\eta\boldsymbol{C}\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})} + \eta^2\boldsymbol{D}. \tag{106}$$

Thus, for all $\boldsymbol{u}\in\mathcal{N}(\boldsymbol{D})$ we have

$$(\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})})\boldsymbol{Q}\boldsymbol{u} = \boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}\boldsymbol{u} - 2\eta\boldsymbol{C}\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}\boldsymbol{u} + \eta^2\boldsymbol{D}\boldsymbol{u} = \boldsymbol{0}. \tag{107}$$

Since the eigenvectors of symmetric matrices are orthogonal, and $\mathcal{N}(\boldsymbol{D})$ is an eigenspace, we get that the top eigenvector of $(\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})})\boldsymbol{Q}$ should be in $\mathcal{N}^\perp(\boldsymbol{D})$. Now, for $\boldsymbol{u}\in\mathcal{N}^\perp(\boldsymbol{D})\cap\mathbb{S}^{d^2-1}$

$$\begin{aligned}\boldsymbol{u}^{\mathrm{T}}(\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})})\boldsymbol{Q}\boldsymbol{u} &= \boldsymbol{u}^{\mathrm{T}}\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}\boldsymbol{u} - 2\eta\boldsymbol{u}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}\boldsymbol{u} + \eta^2\boldsymbol{u}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{u}\\ &= 1 - 2\eta\boldsymbol{u}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{u} + \eta^2\boldsymbol{u}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{u},\end{aligned} \tag{108}$$

where in the second step we used the fact that $\boldsymbol{u}\in\mathcal{N}^\perp(\boldsymbol{D})$, and therefore $\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}\boldsymbol{u} = \boldsymbol{u}$. Additionally, note that

$$\inf_{\boldsymbol{u}\in\mathbb{S}^{d^2-1}:\boldsymbol{u}\notin\mathcal{N}(\boldsymbol{D})}\left\{\frac{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{u}}{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{u}}\right\} = \inf_{\boldsymbol{u}\in\mathbb{S}^{d^2-1}\cap\mathcal{N}^\perp(\boldsymbol{D})}\left\{\frac{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{u}}{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{u}}\right\}. \tag{109}$$

Namely, having a component of $\boldsymbol{u}$ in $\mathcal{N}(\boldsymbol{D})$ can only be non-optimal, since the denominator is invariant to vectors in $\mathcal{N}(\boldsymbol{D})$, while the numerator can only increase ($\boldsymbol{C}$ is PSD, see App. B.7). Now, assuming $\eta > 0$ we have from the derivation of η^*_{var} in the second statement (see (94))

$$\begin{aligned}
& \eta < \eta^*_{\text{var}} \\
\Leftrightarrow \quad & \eta < 2 \inf_{\boldsymbol{u}\in\mathbb{S}^{d^2-1}:\boldsymbol{u}\notin\mathcal{N}(\boldsymbol{D})} \left\{\frac{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{u}}{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{u}}\right\} \\
\Leftrightarrow \quad & \eta < 2 \inf_{\boldsymbol{u}\in\mathbb{S}^{d^2-1}\cap\mathcal{N}^{\perp}(\boldsymbol{D})} \left\{\frac{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{u}}{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{u}}\right\} \\
\Leftrightarrow \quad & \eta < 2\frac{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{u}}{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{u}} \qquad \forall \boldsymbol{u}\in\mathbb{S}^{d^2-1}\cap\mathcal{N}^{\perp}(\boldsymbol{D}) \\
\Leftrightarrow \quad & \eta\boldsymbol{u}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{u} < 2\boldsymbol{u}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{u} \qquad \forall \boldsymbol{u}\in\mathbb{S}^{d^2-1}\cap\mathcal{N}^{\perp}(\boldsymbol{D}) \qquad (\boldsymbol{D}\text{ is PSD}) \\
\Leftrightarrow \quad & \eta^2\boldsymbol{u}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{u} < 2\eta\boldsymbol{u}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{u} \qquad \forall \boldsymbol{u}\in\mathbb{S}^{d^2-1}\cap\mathcal{N}^{\perp}(\boldsymbol{D}) \qquad (\eta > 0) \\
\Leftrightarrow \quad & -2\eta\boldsymbol{u}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{u} + \eta^2\boldsymbol{u}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{u} < 0 \qquad \forall \boldsymbol{u}\in\mathbb{S}^{d^2-1}\cap\mathcal{N}^{\perp}(\boldsymbol{D}) \\
\Leftrightarrow \quad & 1 - 2\eta\boldsymbol{u}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{u} + \eta^2\boldsymbol{u}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{u} < 1 \qquad \forall \boldsymbol{u}\in\mathbb{S}^{d^2-1}\cap\mathcal{N}^{\perp}(\boldsymbol{D}) \\
\Leftrightarrow \quad & \boldsymbol{u}^{\mathrm{T}}\left(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\right)\boldsymbol{Q}\boldsymbol{u} < 1 \qquad \forall \boldsymbol{u}\in\mathbb{S}^{d^2-1}\cap\mathcal{N}^{\perp}(\boldsymbol{D}) \\
\Leftrightarrow \quad & \lambda_{\max}\left(\left(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\right)\boldsymbol{Q}\right) < 1
\end{aligned} \tag{110}$$

where in the fourth step we used the fact that $\boldsymbol{D}$ is PSD (see App. B.7), and in the penultimate step we used (108). Overall we have that $0 < \eta < \eta^*_{\text{var}}$ if and only if $\lambda_{\max}\big((\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})})\boldsymbol{Q}\big) < 1$ (we will use this fact in later sections). Therefore, when $\eta < \eta^*_{\text{var}}$ then

$$\left\|\left(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\right)\boldsymbol{Q}\right\| = \lambda_{\max}\big((\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})})\boldsymbol{Q}\big) < 1. \tag{111}$$

Hence, from (104) we get

$$\lim_{t\to\infty}\|\text{vec}\,(\boldsymbol{\Sigma}_t^{\perp})\| \leq \left\|\left(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\right)\boldsymbol{Q}\right\|^t \|\text{vec}\,(\boldsymbol{\Sigma}_0^{\perp})\| = 0, \tag{112}$$

which proves the statement.

B.9. Proof of Theorem 11

First statement. Let us start by proving the first statement. In (37) we showed that if the minimum is regular then

$$\boldsymbol{\theta}_{t+1}^{\parallel} - \boldsymbol{\theta}^{*\parallel} = \boldsymbol{\theta}_t^{\parallel} - \boldsymbol{\theta}^{*\parallel} - \frac{\eta}{B}\sum_{i\in\mathfrak{B}_t}\boldsymbol{g}_i^{\parallel}. \tag{113}$$

Let us compute the expected squared norm. We have

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{\theta}^{\parallel}_{t+1}-\boldsymbol{\theta}^{*\parallel}\right\|^2\right] &= \mathbb{E}\left[\left\|\boldsymbol{\theta}^{\parallel}_{t}-\boldsymbol{\theta}^{*\parallel}-\frac{\eta}{B}\sum_{i\in\mathfrak{B}_t}\boldsymbol{g}_i^{\parallel}\right\|^2\right]\\
&= \mathbb{E}\left[\left\|\boldsymbol{\theta}^{\parallel}_{t}-\boldsymbol{\theta}^{*\parallel}\right\|^2\right]+\mathbb{E}\left[\left\|\frac{\eta}{B}\sum_{i\in\mathfrak{B}_t}\boldsymbol{g}_i^{\parallel}\right\|^2\right]-2\mathbb{E}\left[\left(\boldsymbol{\theta}^{\parallel}_{t}-\boldsymbol{\theta}^{*\parallel}\right)^{\mathrm{T}}\left(\frac{\eta}{B}\sum_{i\in\mathfrak{B}_t}\boldsymbol{g}_i^{\parallel}\right)\right]\\
&= \mathbb{E}\left[\left\|\boldsymbol{\theta}^{\parallel}_{t}-\boldsymbol{\theta}^{*\parallel}\right\|^2\right]+\mathbb{E}\left[\left\|\frac{\eta}{B}\sum_{i\in\mathfrak{B}_t}\boldsymbol{g}_i^{\parallel}\right\|^2\right]-2\mathbb{E}\left[\boldsymbol{\theta}^{\parallel}_{t}-\boldsymbol{\theta}^{*\parallel}\right]^{\mathrm{T}}\mathbb{E}\left[\frac{\eta}{B}\sum_{i\in\mathfrak{B}_t}\boldsymbol{g}_i^{\parallel}\right]\\
&= \mathbb{E}\left[\left\|\boldsymbol{\theta}^{\parallel}_{t}-\boldsymbol{\theta}^{*\parallel}\right\|^2\right]+\mathbb{E}\left[\left\|\frac{\eta}{B}\sum_{i\in\mathfrak{B}_t}\boldsymbol{g}_i^{\parallel}\right\|^2\right], \qquad (114)\end{aligned}$$

where in the third step we used the fact that $\boldsymbol{\theta}^{\parallel}_{t}$ is independent of $\mathfrak{B}_t$ and in the last we used the fact that

$$\mathbb{E}\left[\frac{\eta}{B}\sum_{i\in\mathfrak{B}_t}\boldsymbol{g}_i^{\parallel}\right]=\frac{\eta}{n}\sum_{i=1}^{n}\boldsymbol{g}_i^{\parallel}=\boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})}\frac{\eta}{n}\sum_{i=1}^{n}\boldsymbol{g}_i=\boldsymbol{0}. \qquad (115)$$

Calculating the right term in the last line of (114) using the definition of $\boldsymbol{v}_t$ (see (48)) gives

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{\eta}{B}\sum_{i\in\mathfrak{B}_t}\boldsymbol{g}_i^{\parallel}\right\|^2\right] &= \mathbb{E}\left[\left\|\boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})}\boldsymbol{v}_t\right\|^2\right]\\
&= \mathrm{Tr}\left(\boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})}\mathbb{E}\left[\boldsymbol{v}_t\boldsymbol{v}_t^{\mathrm{T}}\right]\boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})}\right)\\
&= \eta^2\frac{n-B}{B(n-1)}\frac{1}{n}\sum_{i=1}^{n}\mathrm{Tr}\left(\boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})}\boldsymbol{g}_i\boldsymbol{g}_i^{\mathrm{T}}\boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})}\right)\\
&= \eta^2 p\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})}\boldsymbol{g}_i\right\|^2\\
&= \eta^2 p\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{g}_i^{\parallel}\right\|^2, \qquad (116)\end{aligned}$$

where in the third step we used (70). Unrolling (114) we have that

$$\mathbb{E}\left[\left\|\boldsymbol{\theta}^{\parallel}_{t}-\boldsymbol{\theta}^{*\parallel}\right\|^2\right]=\mathbb{E}\left[\left\|\boldsymbol{\theta}^{\parallel}_{0}-\boldsymbol{\theta}^{*\parallel}\right\|^2\right]+t\times\eta^2 p\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{g}_i^{\parallel}\right\|^2. \qquad (117)$$

Thus, $\lim_{t\to\infty}\mathbb{E}[\|\boldsymbol{\theta}^{\parallel}_{t}-\boldsymbol{\theta}^{*\parallel}\|^2]=\infty$ if and only if $\sum_{i=1}^{n}\left\|\boldsymbol{g}_i^{\parallel}\right\|^2>0$.

Second and third statements. Next, we turn to prove the second and third statements of the theorem. In App. B.10 we show the following.

Lemma 16 *Assume that $\boldsymbol{\theta}^*$ is a twice differentiable regular minimum. Consider the linear dynamics of $\{\boldsymbol{\theta}_t\}$ from Def. 1.*

1. *If* $\lambda_{\max}\big((\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})} \otimes \boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})})\boldsymbol{Q}\big) < 1$ *then* $\limsup_{t\to\infty} \mathbb{E}[\|\boldsymbol{\theta}_t^\perp - \boldsymbol{\theta}^{*\perp}\|^2]$ *is finite.*

2. *If* $\limsup_{t\to\infty} \mathbb{E}[\|\boldsymbol{\theta}_t^\perp - \boldsymbol{\theta}^{*\perp}\|^2]$ *is finite then* $\lambda_{\max}\big((\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})} \otimes \boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})})\boldsymbol{Q}\big) \leq 1$.

3. *Let* $\boldsymbol{z}_{\max}$ *denote the top eigenvector of* $(\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})} \otimes \boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})})\boldsymbol{Q}$*, and assume that* $\boldsymbol{z}_{\max}^{\mathrm{T}}\mathrm{vec}(\boldsymbol{\Sigma}_{\boldsymbol{g}}^\perp) \neq 0$*. If* $\limsup_{t\to\infty} \mathbb{E}[\|\boldsymbol{\theta}_t^\perp - \boldsymbol{\theta}^{*\perp}\|^2]$ *is finite then* $\lambda_{\max}\big((\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})} \otimes \boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})})\boldsymbol{Q}\big) < 1$.

In (110) we showed that $\lambda_{\max}\big((\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})} \otimes \boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})})\boldsymbol{Q}\big) < 1$ if and only if $0 < \eta < \eta^*_{\mathrm{var}}$, which proves the second and third statements. Note that under the mild assumption that $\boldsymbol{z}_{\max}^{\mathrm{T}}\mathrm{vec}(\boldsymbol{\Sigma}_{\boldsymbol{g}}^\perp) \neq 0$ we get that $\limsup_{t\to\infty} \mathbb{E}[\|\boldsymbol{\theta}_t^\perp - \boldsymbol{\theta}^{*\perp}\|^2]$ is finite if and only if $0 \leq \eta < \eta^*_{\mathrm{var}}$.

B.10. Proof of Lemma 16

First statement. Here we assume $\lambda_{\max}\big((\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})} \otimes \boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})})\boldsymbol{Q}\big) < 1$, and show this implies that $\limsup_{t\to\infty} \mathbb{E}[\|\boldsymbol{\theta}_t^\perp - \boldsymbol{\theta}^{*\perp}\|^2]$ is finite. The (projected) transition matrix that governs the dynamics of $\boldsymbol{\Sigma}_t^\perp$ and $\boldsymbol{\mu}_t^\perp$ in (56) is given by

$$\boldsymbol{\Xi} = \begin{pmatrix} \boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})} - \eta\boldsymbol{H} & \boldsymbol{0} \\ -\left(\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})} \otimes \boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})}\right)\left(\mathbb{E}\left[\boldsymbol{v}_t^\perp \otimes \boldsymbol{A}_t\right] + \mathbb{E}\left[\boldsymbol{A}_t \otimes \boldsymbol{v}_t^\perp\right]\right) & \left(\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})} \otimes \boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})}\right)\boldsymbol{Q} \end{pmatrix}. \tag{118}$$

Since this matrix is a block lower triangular matrix, its eigenvalues are

$$\{\lambda_j(\boldsymbol{\Xi})\} = \Big\{\lambda_j\big(\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})} - \eta\boldsymbol{H}\big)\Big\} \bigcup \Big\{\lambda_j\left(\big(\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})} \otimes \boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})}\big)\boldsymbol{Q}\right)\Big\}. \tag{119}$$

In Lemma 17 we show that if $\rho\big((\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})} \otimes \boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})})\boldsymbol{Q}\big) < 1$ then $\rho(\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})} - \eta\boldsymbol{H}) < 1$ (see proof in App. B.11). Therefore, all the eigenvalues of $\boldsymbol{\Xi}$ are less than 1 in absolute value. Therefore, $\|\mathrm{vec}(\boldsymbol{\Sigma}_t^\perp)\|_2 = \|\boldsymbol{\Sigma}_t^\perp\|_{\mathrm{F}}$ is bounded. Since $\boldsymbol{\Sigma}_t^\perp$ is PSD we have

$$\|\boldsymbol{\Sigma}_t^\perp\|_{\mathrm{F}} = \sqrt{\sum_{j=1}^d \lambda_j^2(\boldsymbol{\Sigma}_t^\perp)} \geq \frac{1}{\sqrt{d}}\sum_{j=1}^d \lambda_j(\boldsymbol{\Sigma}_t^\perp) = \frac{1}{\sqrt{d}}\mathrm{Tr}(\boldsymbol{\Sigma}_t^\perp) = \frac{1}{\sqrt{d}}\mathbb{E}\left[\|\boldsymbol{\theta}_t^\perp - \boldsymbol{\theta}^{*\perp}\|^2\right]. \tag{120}$$

Therefore, $\mathbb{E}[\|\boldsymbol{\theta}_t^\perp - \boldsymbol{\theta}^{*\perp}\|^2]$ is bounded.

Second statement. Here we assume that $\limsup_{t\to\infty} \mathbb{E}[\|\boldsymbol{\theta}_t^\perp - \boldsymbol{\theta}^{*\perp}\|^2]$ is finite, then we show $\lambda_{\max}\big((\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})} \otimes \boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})})\,\boldsymbol{Q}\big) \leq 1$. The matrix $\big(\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})} \otimes \boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})}\big)\boldsymbol{Q}$ can be written as

$$\begin{aligned}\left(\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})} \otimes \boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})}\right)\boldsymbol{Q} = &\left(\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})} - \eta\boldsymbol{H}\right) \otimes \left(\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})} - \eta\boldsymbol{H}\right) \\ &+ \eta^2 p\left(\frac{1}{n}\sum_{i=1}^n (\boldsymbol{H}_i - \boldsymbol{H}) \otimes (\boldsymbol{H}_i - \boldsymbol{H})\right),\end{aligned} \tag{121}$$

where we used (66) for the value of $\boldsymbol{Q}$. This expression is a sum of Kronecker products, where each product is a symmetric matrix with itself. Therefore, according to Thm. 14, we get

$$\lambda_{\max}\big((\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})})\boldsymbol{Q}\big)=\rho\big((\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})})\boldsymbol{Q}\big) \tag{122}$$

and $\boldsymbol{Z}_{\max}=\mathrm{vec}^{-1}(\boldsymbol{z}_{\max})$ is a PSD matrix, where $\boldsymbol{z}_{\max}$ is a normalized top eigenvector of $\big(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\big)\boldsymbol{Q}$. Now, set[3] $\boldsymbol{\Sigma}_0^{\perp}=\boldsymbol{Z}_{\max}$ and $\boldsymbol{\mu}_0=\boldsymbol{0}$, then in this case $\boldsymbol{\mu}_t^{\perp}=(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}-\eta\boldsymbol{H})^t\boldsymbol{\mu}_0^{\perp}=\boldsymbol{0}$ for all $t>0$. Therefore, from (56)

$$\mathrm{vec}\left(\boldsymbol{\Sigma}_{t+1}\right)=(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})})\boldsymbol{Q}\mathrm{vec}\left(\boldsymbol{\Sigma}_t^{\perp}\right)+\mathrm{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{v}}^{\perp}\right). \tag{123}$$

Thus, taking an inner product w.r.t. $\boldsymbol{z}_{\max}$ on both sides we get

$$\begin{aligned}\boldsymbol{z}_{\max}^{\mathrm{T}}\mathrm{vec}\left(\boldsymbol{\Sigma}_{t+1}^{\perp}\right)&=\boldsymbol{z}_{\max}^{\mathrm{T}}(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})})\boldsymbol{Q}\mathrm{vec}\left(\boldsymbol{\Sigma}_t^{\perp}\right)+\boldsymbol{z}_{\max}^{\mathrm{T}}\mathrm{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{v}}^{\perp}\right)\\&=\lambda_{\max}\big((\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})})\boldsymbol{Q}\big)\boldsymbol{z}_{\max}^{\mathrm{T}}\mathrm{vec}\left(\boldsymbol{\Sigma}_t^{\perp}\right)+\mathrm{Tr}(\boldsymbol{Z}_{\max}\boldsymbol{\Sigma}_{\boldsymbol{v}}^{\perp})\\&\geq\lambda_{\max}\big((\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})})\boldsymbol{Q}\big)\boldsymbol{z}_{\max}^{\mathrm{T}}\mathrm{vec}\left(\boldsymbol{\Sigma}_t^{\perp}\right),\end{aligned} \tag{124}$$

where in the last step we used the fact that $\boldsymbol{Z}_{\max}^{\frac{1}{2}}\boldsymbol{\Sigma}_{\boldsymbol{v}}^{\perp}\boldsymbol{Z}_{\max}^{\frac{1}{2}}$ is a PSD matrix, and thus

$$\mathrm{Tr}\left(\boldsymbol{Z}_{\max}\boldsymbol{\Sigma}_{\boldsymbol{v}}^{\perp}\right)=\mathrm{Tr}\left(\boldsymbol{Z}_{\max}^{\frac{1}{2}}\boldsymbol{\Sigma}_{\boldsymbol{v}}^{\perp}\boldsymbol{Z}_{\max}^{\frac{1}{2}}\right)\geq 0. \tag{125}$$

Therefore, using (124) t times we have

$$\begin{aligned}\boldsymbol{z}_{\max}^{\mathrm{T}}\mathrm{vec}\left(\boldsymbol{\Sigma}_t^{\perp}\right)&\geq\lambda_{\max}^t\big((\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})})\boldsymbol{Q}\big)\boldsymbol{z}_{\max}^{\mathrm{T}}\mathrm{vec}\left(\boldsymbol{\Sigma}_0\right)\\&=\lambda_{\max}^t\big((\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})})\boldsymbol{Q}\big),\end{aligned} \tag{126}$$

where in the last step we used $\boldsymbol{\Sigma}_0=\boldsymbol{Z}_{\max}$ and $\|\boldsymbol{Z}_{\max}\|_{\mathrm{F}}=1$. Additionally, for all $t>0$

$$\begin{aligned}\boldsymbol{z}_{\max}^{\mathrm{T}}\mathrm{vec}\left(\boldsymbol{\Sigma}_t^{\perp}\right)&\leq\|\boldsymbol{z}_{\max}\|_2\left\|\mathrm{vec}\left(\boldsymbol{\Sigma}_t^{\perp}\right)\right\|_2\\&=\|\boldsymbol{\Sigma}_t^{\perp}\|_{\mathrm{F}}\\&=\sqrt{\sum_{j=1}^{d}\lambda_j^2\big(\boldsymbol{\Sigma}_t^{\perp}\big)}\\&\leq\sum_{j=1}^{d}\lambda_j\big(\boldsymbol{\Sigma}_t^{\perp}\big)\\&=\mathrm{Tr}\big(\boldsymbol{\Sigma}_t^{\perp}\big)\\&=\mathbb{E}\left[\|\boldsymbol{\theta}_t^{\perp}-\boldsymbol{\theta}^{*\perp}\|^2\right].\end{aligned} \tag{127}$$

Overall, combining (126) and (127) results with

$$\lambda_{\max}^t\big((\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})})\boldsymbol{Q}\big)\leq\mathbb{E}\left[\|\boldsymbol{\theta}_t^{\perp}-\boldsymbol{\theta}^{*\perp}\|^2\right]. \tag{128}$$

Since $\mathbb{E}[\|\boldsymbol{\theta}_t^{\perp}-\boldsymbol{\theta}^{*\perp}\|^2]$ is bounded then $\lambda_{\max}\big((\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})})\boldsymbol{Q}\big)\leq 1$.

[3] Since the eigenvectors of symmetric matrices are orthogonal, and $\mathcal{N}(\boldsymbol{D})$ is an eigenspace, we get that the top eigenvector of $(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})})\boldsymbol{Q}$ should be in $\mathcal{N}^{\perp}(\boldsymbol{D})$, *i.e.*, $\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\boldsymbol{Z}_{\max}\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}=\boldsymbol{Z}_{\max}$. Therefore this initialization is possible.

Third statement. Furthermore, if $\boldsymbol{z}_{\max}^{\mathrm{T}}\mathrm{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{v}}^{\perp}\right) \neq 0$ we get from (125)

$$\boldsymbol{z}_{\max}^{\mathrm{T}}\mathrm{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{v}}^{\perp}\right) > 0. \tag{129}$$

Assume by contradiction that $\lambda_{\max}\big((\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})} \otimes \boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})})\boldsymbol{Q}\big) = 1$, then (124) gives

$$\boldsymbol{z}_{\max}^{\mathrm{T}}\mathrm{vec}\left(\boldsymbol{\Sigma}_{t+1}^{\perp}\right) = \boldsymbol{z}_{\max}^{\mathrm{T}}\mathrm{vec}\left(\boldsymbol{\Sigma}_{t}^{\perp}\right) + \boldsymbol{z}_{\max}^{\mathrm{T}}\mathrm{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{v}}^{\perp}\right). \tag{130}$$

Unrolling this equation gives $\boldsymbol{z}_{\max}^{\mathrm{T}}\mathrm{vec}\left(\boldsymbol{\Sigma}_{t}^{\perp}\right) = t\boldsymbol{z}_{\max}^{\mathrm{T}}\mathrm{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{v}}^{\perp}\right)$. Then, by (127) we get

$$\mathbb{E}\left[\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|^2\right] \geq \boldsymbol{z}_{\max}^{\mathrm{T}}\mathrm{vec}\left(\boldsymbol{\Sigma}_{t}^{\perp}\right) = t\boldsymbol{z}_{\max}^{\mathrm{T}}\mathrm{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{v}}^{\perp}\right). \tag{131}$$

Since $\boldsymbol{z}_{\max}^{\mathrm{T}}\mathrm{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{v}}^{\perp}\right) > 0$, then $\mathbb{E}[\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|^2] \to \infty$ and we have a contradiction. Therefore $\lambda_{\max}\big((\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})} \otimes \boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})})\boldsymbol{Q}\big) < 1$.

B.11. Proof of Lemma 17

Lemma 17 *Let $\boldsymbol{H} \in \mathcal{S}_+(\mathbb{R}^{d\times d})$ and $\boldsymbol{Q}$ defined as in (10). If $\rho((\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})} \otimes \boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})})\boldsymbol{Q}) < 1$ then $\rho(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})} - \eta\boldsymbol{H}) < 1$.*

The matrix $(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})} \otimes \boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})})\boldsymbol{Q}$ is

$$\begin{aligned}\left(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})} \otimes \boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\right)\boldsymbol{Q} = \left(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})} - \eta\boldsymbol{H}\right) \otimes \left(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})} - \eta\boldsymbol{H}\right)&\\ + \eta^2 p\left(\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{H}_i - \boldsymbol{H}) \otimes (\boldsymbol{H}_i - \boldsymbol{H})\right)&,\end{aligned} \tag{132}$$

where we used (66) for the value of $\boldsymbol{Q}$. Note that $\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})} - \eta\boldsymbol{H}$ is a symmetric matrix, and thus each of its dominant eigenvectors $\tilde{\boldsymbol{v}} \in \mathbb{S}^{d-1}$ satisfies $\rho(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})} - \eta\boldsymbol{H}) = |\tilde{\boldsymbol{v}}^{\mathrm{T}}(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})} - \eta\boldsymbol{H})\tilde{\boldsymbol{v}}|$. Additionally, $\|\tilde{\boldsymbol{v}} \otimes \tilde{\boldsymbol{v}}\| = \left\|\tilde{\boldsymbol{v}}\tilde{\boldsymbol{v}}^{\mathrm{T}}\right\|_{\mathrm{F}} = 1$, *i.e.*, $\tilde{\boldsymbol{v}} \otimes \tilde{\boldsymbol{v}} \in \mathbb{S}^{d^2-1}$. Now, let the spectral radius $\rho\big((\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})} \otimes \boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})})\boldsymbol{Q}\big) < 1$ then

$$\begin{aligned}1 &> \rho\big((\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})} \otimes \boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})})\boldsymbol{Q}\big)\\ &\geq \left|[\tilde{\boldsymbol{v}} \otimes \tilde{\boldsymbol{v}}]^{\mathrm{T}}\big(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})} \otimes \boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\big)\boldsymbol{Q}[\tilde{\boldsymbol{v}} \otimes \tilde{\boldsymbol{v}}]\right|\\ &= \left(\tilde{\boldsymbol{v}}^{\mathrm{T}}\left(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})} - \eta\boldsymbol{H}\right)\tilde{\boldsymbol{v}}\right)^2 + \eta^2 p\frac{1}{n}\sum_{i=1}^{n}\left(\tilde{\boldsymbol{v}}^{\mathrm{T}}\left(\boldsymbol{H}_i - \boldsymbol{H}\right)\tilde{\boldsymbol{v}}\right)^2\\ &= \rho^2(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})} - \eta\boldsymbol{H}) + \eta^2 p\frac{1}{n}\sum_{i=1}^{n}\left(\tilde{\boldsymbol{v}}^{\mathrm{T}}\left(\boldsymbol{H}_i - \boldsymbol{H}\right)\tilde{\boldsymbol{v}}\right)^2\\ &\geq \rho^2(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})} - \eta\boldsymbol{H}),\end{aligned} \tag{133}$$

where in the third step we used (132).

Appendix C. Proof of Theorem 14

First statement. Let $\{\boldsymbol{Y}_i\}$ be symmetric matrices in $\mathbb{R}^{d\times d}$, and let

$$\boldsymbol{Q} = \sum_{i=1}^{M} \boldsymbol{Y}_i \otimes \boldsymbol{Y}_i. \tag{134}$$

First, note that $\boldsymbol{Q} \in \mathbb{R}^{d^2\times d^2}$ is symmetric.

$$\boldsymbol{Q}^{\mathrm{T}} = \left(\sum_{i=1}^{M} \boldsymbol{Y}_i \otimes \boldsymbol{Y}_i\right)^{\mathrm{T}} = \sum_{i=1}^{M} (\boldsymbol{Y}_i \otimes \boldsymbol{Y}_i)^{\mathrm{T}} = \sum_{i=1}^{M} \boldsymbol{Y}_i^{\mathrm{T}} \otimes \boldsymbol{Y}_i^{\mathrm{T}} = \sum_{i=1}^{M} \boldsymbol{Y}_i \otimes \boldsymbol{Y}_i = \boldsymbol{Q}, \tag{135}$$

where in the third step we used (P2) property of the Kronecker product, and in the fourth we used the fact that $\{\boldsymbol{Y}_i\}$ are symmetric. Then, by the spectral theorem, we have that all its eigenvectors $\{\boldsymbol{z}_j\}$ and eigenvalues $\{\lambda_j\}$ are real. Given an eigenvector $\boldsymbol{z} \in \mathbb{R}^{d^2}$ of $\boldsymbol{Q}$, we can examine its matrix form $\boldsymbol{Z} = \mathrm{vec}^{-1}(\boldsymbol{z})$, where $\boldsymbol{Z} \in \mathbb{R}^{d\times d}$. Here we show that $\boldsymbol{Q}$ always has a set of d^2 eigenvectors that correspond only to either symmetric or skew-symmetric matrices $\{\boldsymbol{Z}_j\}$. Let $(\lambda, \boldsymbol{z})$ be an eigenpair of $\boldsymbol{Q}$, *i.e.,* $\lambda\boldsymbol{z} = \boldsymbol{Q}\boldsymbol{z}$, and set $\boldsymbol{Z} = \mathrm{vec}^{-1}(\boldsymbol{z})$. Then,

$$\begin{aligned}\lambda\boldsymbol{Z} = \mathrm{vec}^{-1}(\lambda\boldsymbol{z}) = \mathrm{vec}^{-1}(\boldsymbol{Q}\boldsymbol{z}) = \mathrm{vec}^{-1}\left(\sum_{i=1}^{M} \boldsymbol{Y}_i \otimes \boldsymbol{Y}_i\boldsymbol{z}\right) &= \sum_{i=1}^{M} \mathrm{vec}^{-1}\left(\boldsymbol{Y}_i \otimes \boldsymbol{Y}_i\boldsymbol{z}\right) \\ &= \sum_{i=1}^{M} \boldsymbol{Y}_i\boldsymbol{Z}\boldsymbol{Y}_i^{\mathrm{T}},\end{aligned} \tag{136}$$

where in the penultimate step we used (P1) property of the Kronecker product. By taking a transpose on both ends of this equation we have

$$\lambda\boldsymbol{Z}^{\mathrm{T}} = \left(\sum_{i=1}^{M} \boldsymbol{Y}_i\boldsymbol{Z}\boldsymbol{Y}_i^{\mathrm{T}}\right)^{\mathrm{T}} = \sum_{i=1}^{M} \left(\boldsymbol{Y}_i\boldsymbol{Z}\boldsymbol{Y}_i^{\mathrm{T}}\right)^{\mathrm{T}} = \sum_{i=1}^{M} \boldsymbol{Y}_i\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Y}_i^{\mathrm{T}}. \tag{137}$$

Thus, we have that $\mathrm{vec}(\boldsymbol{Z}^{\mathrm{T}})$ is also an eigenvector of $\boldsymbol{Q}$. If λ has multiplicity one, then it must be that $\boldsymbol{Z}^{\mathrm{T}} = \pm\boldsymbol{Z}$, *i.e.,* symmetric or skew-symmetric matrix. If the multiplicity is greater than one and $\boldsymbol{Z}^{\mathrm{T}} \neq \pm\boldsymbol{Z}$, then any linear combination of $\boldsymbol{Z}$ and $\boldsymbol{Z}^{\mathrm{T}}$ is also an eigenvector corresponding to λ. In particular,

$$\hat{\boldsymbol{Z}}_1 = \frac{1}{2}\left(\boldsymbol{Z} + \boldsymbol{Z}^{\mathrm{T}}\right) \qquad \text{and} \qquad \hat{\boldsymbol{Z}}_2 = \frac{1}{2}\left(\boldsymbol{Z} - \boldsymbol{Z}^{\mathrm{T}}\right). \tag{138}$$

By construction, $\hat{\boldsymbol{Z}}_1$ and $\hat{\boldsymbol{Z}}_2$ are symmetric and skew-symmetric eigenvectors, corresponding to λ. This procedure can be repeated while projecting the next eigenvectors of λ onto the orthogonal complement of the already found vectors, until we find all eigenvectors of λ. In this way, we can find a set of eigenvectors comprised solely of vectors that correspond to symmetric or skew-symmetric.

Second and third statement. Using the first statement, we can consider a complete set of eigenvectors for $\boldsymbol{Q}$ that is comprised solely of vectors that correspond to either symmetric or skew-symmetric matrices. Our next step is to show that there is at least one dominant eigenvector of $\boldsymbol{Q}$ that corresponds to a symmetric matrix. Our final step will be to show that among the dominant eigenvectors that correspond to symmetric matrices, at least one corresponds to a PDS matrix.

To this end, we first bring the eigenvalues of $\boldsymbol{Q}$, denoted by $\{\boldsymbol{Z}_j\}$, to a normal (canonical) form. Here we assume without loss of generality that the eigenvectors are normalized, that is $\|\boldsymbol{Z}_j\|_{\mathrm{F}} = 1$ for all $j \in [d^2]$. For symmetric matrix $\boldsymbol{Z}$, we have the spectral decomposition theorem, and thus $\boldsymbol{Z} = \boldsymbol{V}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}$, where $\boldsymbol{V}$ is an orthogonal matrix and $\boldsymbol{S}$ is diagonal. We can also bring a skew-symmetric matrix to a similar form of $\boldsymbol{Z} = \boldsymbol{V}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}$ with orthogonal $\boldsymbol{V}$, where $\boldsymbol{S}$ is a block diagonal matrix, with $\lfloor d/2 \rfloor$ blocks of size 2×2. Specifically, these blocks are in the form of (Zumino, 1962)

$$\begin{bmatrix} 0 & s_\ell \\ -s_\ell & 0 \end{bmatrix}. \tag{139}$$

If the dimension d is odd, then the last row and column of $\boldsymbol{S}$ are the zero vectors. For numerical purposes, this normal (canonical) form can be computed using the real Schur decomposition.

For symmetric matrices, we define the vector $\boldsymbol{s}_{\text{sym}} \in \mathbb{R}^d$ to be the diagonal of $\boldsymbol{S}$, and for skew-symmetric matrices we define $\boldsymbol{s}_{\text{skew}} \in \mathbb{R}^{\lfloor d/2 \rfloor}$ to be $[s_1, s_2, \cdots, s_{\lfloor d/2 \rfloor}]^{\mathrm{T}}$. In App. C.1 we show that for a symmetric matrix $\boldsymbol{Z}$, its corresponding vector form $\boldsymbol{z}$ satisfies

$$\boldsymbol{z}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{z} = \boldsymbol{s}_{\text{sym}}^{\mathrm{T}} \sum_{i=1}^{M} \boldsymbol{M}_i^{\odot 2} \boldsymbol{s}_{\text{sym}}, \tag{140}$$

where $\boldsymbol{M}_i = \boldsymbol{V}^{\mathrm{T}}\boldsymbol{Y}_i\boldsymbol{V}$, the superscript ${}^{\odot k}$ denotes the Hadamard power and $\|\boldsymbol{s}_{\text{sym}}\| = 1$. For skew-symmetric matrices, we define a set of matrices $\{\boldsymbol{T}_i\}$ in $\mathbb{R}^{\lfloor d/2 \rfloor \times \lfloor d/2 \rfloor}$, where

$$\boldsymbol{T}_{i\,[\ell,p]} = \boldsymbol{M}_{i\,[2\ell-1,2p-1]}\boldsymbol{M}_{i\,[2\ell,2p]} - \boldsymbol{M}_{i\,[2\ell-1,2p]}\boldsymbol{M}_{i\,[2\ell,2p-1]}, \tag{141}$$

for all $1 \leq \ell, p \leq \lfloor d/2 \rfloor$. Namely, $\boldsymbol{T}_i$ is the determinant of each 2×2 block of $\boldsymbol{M}_i$ without overlap. We show in App. C.1 that for a skew-symmetric matrix $\boldsymbol{Z}$, its corresponding vector form $\boldsymbol{z}$ satisfies

$$\boldsymbol{z}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{z} = 2\boldsymbol{s}_{\text{skew}}^{\mathrm{T}} \sum_{i=1}^{M} \boldsymbol{T}_i \boldsymbol{s}_{\text{skew}}, \tag{142}$$

where $\|\boldsymbol{s}_{\text{skew}}\| = 1/\sqrt{2}$. Let us define the projection matrix $\boldsymbol{P} \in \mathbb{R}^{\lfloor d/2 \rfloor \times d}$ as

$$\boldsymbol{P} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & & & \ddots & & & & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 1 & 1 \end{bmatrix}. \tag{143}$$

If d is odd, then the last column of $\boldsymbol{P}$ is the zero vector. This matrix is semi-orthogonal, *i.e.,* it satisfies $\boldsymbol{P}\boldsymbol{P}^{\mathrm{T}} = \boldsymbol{I}$. Note that

$$\left[\boldsymbol{P}\boldsymbol{M}_i^{\odot 2}\boldsymbol{P}^{\mathrm{T}}\right]_{[\ell,p]} = \frac{1}{2}\left(\boldsymbol{M}_{i\,[2\ell-1,2p-1]}^2 + \boldsymbol{M}_{i\,[2\ell-1,2p]}^2 + \boldsymbol{M}_{i\,[2\ell,2p-1]}^2 + \boldsymbol{M}_{i\,[2\ell,2p]}^2\right). \tag{144}$$

Therefore, for all $1 \leq \ell, p \leq \lfloor d/2 \rfloor$ and $i \in [M]$ we have

$$\begin{aligned} \left|\boldsymbol{T}_{i\,[\ell,p]}\right| &= \left|\boldsymbol{M}_{i\,[2\ell-1,2p-1]}\boldsymbol{M}_{i\,[2\ell,2p]} - \boldsymbol{M}_{i\,[2\ell-1,2p]}\boldsymbol{M}_{i\,[2\ell,2p-1]}\right| \\ &\leq \left|\boldsymbol{M}_{i\,[2\ell-1,2p-1]}\boldsymbol{M}_{i\,[2\ell,2p]}\right| + \left|\boldsymbol{M}_{i\,[2\ell-1,2p]}\boldsymbol{M}_{i\,[2\ell,2p-1]}\right| \\ &\leq \frac{1}{2}\left(\boldsymbol{M}_{i\,[2\ell-1,2p-1]}^2 + \boldsymbol{M}_{i\,[2\ell,2p]}^2\right) + \frac{1}{2}\left(\boldsymbol{M}_{i\,[2\ell-1,2p]}^2 + \boldsymbol{M}_{i\,[2\ell,2p-1]}^2\right) \\ &= \left[\boldsymbol{P}\boldsymbol{M}_i^{\odot 2}\boldsymbol{P}^{\mathrm{T}}\right]_{[\ell,p]}, \end{aligned} \tag{145}$$

where in the first step we used (141), in the second we used the triangle inequality, in the third we used $|ab| \leq \frac{1}{2}(a^2+b^2)$ twice, and in the last step we used (144).

Note that any pair of orthogonal matrix $\boldsymbol{V}$ and a vector $\boldsymbol{s}_{\text{skew}} \in \mathbb{R}^{\lfloor d/2 \rfloor}$, such that $\|\boldsymbol{s}_{\text{skew}}\| = 1/\sqrt{2}$, define a skew-symmetric matrix $\boldsymbol{Z}_{\text{skew}}$ with a vectorization $\boldsymbol{z}_{\text{skew}}$. Similarly, any pair of orthogonal matrix $\boldsymbol{V}$ and a vector $\boldsymbol{s}_{\text{sym}} \in \mathbb{R}^d$, such that $\|\boldsymbol{s}_{\text{sym}}\| = 1$, correspond to a symmetric matrix $\boldsymbol{Z}_{\text{sym}}$ with a vectorization $\boldsymbol{z}_{\text{sym}}$. Here we will show that given $\boldsymbol{V}$, there exists $\boldsymbol{s}_{\text{sym}} \in \mathbb{S}^{d-1}$ s.t.

$$\left|\boldsymbol{z}_{\text{skew}}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{z}_{\text{skew}}\right| \leq \boldsymbol{z}_{\text{sym}}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{z}_{\text{sym}} \tag{146}$$

for any $\boldsymbol{s}_{\text{skew}} \in \mathbb{R}^{\lfloor d/2 \rfloor}$ for which $\|\boldsymbol{s}_{\text{skew}}\| = 1/\sqrt{2}$. To this end, we set $\boldsymbol{r} = \sqrt{2}\boldsymbol{s}_{\text{skew}} \in \mathbb{S}^{\lfloor d/2 \rfloor}$, then

$$\begin{aligned}
\left|2\boldsymbol{s}_{\text{skew}}^{\mathrm{T}}\sum_{i=1}^{M}\boldsymbol{T}_i\boldsymbol{s}_{\text{skew}}\right| &= \left|\boldsymbol{r}^{\mathrm{T}}\sum_{i=1}^{M}\boldsymbol{T}_i\boldsymbol{r}\right| \\
&\leq \sum_{\ell=1}^{d}\sum_{p=1}^{d}\sum_{i=1}^{M}\left|\boldsymbol{r}_{[\ell]}\right|\left|\boldsymbol{T}_{i\,[\ell,p]}\right|\left|\boldsymbol{r}_{[p]}\right| \\
&\leq \sum_{\ell=1}^{d}\sum_{p=1}^{d}\sum_{i=1}^{M}\left|\boldsymbol{r}_{[\ell]}\right|\left[\boldsymbol{P}\boldsymbol{M}_i^{\odot 2}\boldsymbol{P}^{\mathrm{T}}\right]_{[\ell,p]}\left|\boldsymbol{r}_{[p]}\right| \\
&= \sum_{\ell=1}^{d}\sum_{p=1}^{d}\left|\boldsymbol{r}_{[\ell]}\right|\left[\sum_{i=1}^{M}\boldsymbol{P}\boldsymbol{M}_i^{\odot 2}\boldsymbol{P}^{\mathrm{T}}\right]_{[\ell,p]}\left|\boldsymbol{r}_{[p]}\right| \\
&\leq \lambda_{\max}\left(\sum_{i=1}^{M}\boldsymbol{P}\boldsymbol{M}_i^{\odot 2}\boldsymbol{P}^{\mathrm{T}}\right) \\
&= \lambda_{\max}\left(\boldsymbol{P}\left(\sum_{i=1}^{M}\boldsymbol{M}_i^{\odot 2}\right)\boldsymbol{P}^{\mathrm{T}}\right) \\
&\leq \lambda_{\max}\left(\sum_{i=1}^{M}\boldsymbol{M}_i^{\odot 2}\right),
\end{aligned} \tag{147}$$

where in the second step we used the triangle inequality, in the third step we used (145), in the fifth we used the fact that $\|\boldsymbol{r}\| = 1$ and bound the quadratic form with the top eigenvalue (note that $\{\boldsymbol{M}_i\}$ are symmetric and thus the top eigenvalue is real), and in the last step we used the Cauchy interlacing theorem (a.k.a. Poincaré separation theorem). Now, take $\boldsymbol{s}_{\text{sym}} \in \mathbb{S}^{d-1}$ to be the top eigenvector (normalized) of $\sum_{i=1}^{M}\boldsymbol{M}_i^{\odot 2}$ (note that $\{\boldsymbol{M}_i\}$ are symmetric and thus this top eigenvector is real), and pair it with the same basis $\boldsymbol{V}$ of $\boldsymbol{Z}_{\text{skew}}$ to get a symmetric matrix $\boldsymbol{Z}_{\text{sym}}$ that satisfies

$$\left|\boldsymbol{z}_{\text{skew}}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{z}_{\text{skew}}\right| = \left|2\boldsymbol{s}_{\text{skew}}^{\mathrm{T}}\sum_{i=1}^{M}\boldsymbol{T}_i\boldsymbol{s}_{\text{skew}}\right| \leq \lambda_{\max}\left(\sum_{i=1}^{M}\boldsymbol{M}_i^{\odot 2}\right) = \boldsymbol{s}_{\text{sym}}^{\mathrm{T}}\sum_{i=1}^{M}\boldsymbol{M}_i^{\odot 2}\boldsymbol{s}_{\text{sym}} = \boldsymbol{z}_{\text{sym}}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{z}_{\text{sym}}, \tag{148}$$

where in the first step we used (142), in the second we used (147), in the third we used the fact that $\boldsymbol{s}_{\text{sym}} \in \mathbb{S}^{d-1}$ is the top eigenvector of $\sum_{i=1}^{M}\boldsymbol{M}_i^{\odot 2}$, and in the last step we used (140). Since this is true for any orthogonal $\boldsymbol{V}$, we get that at least one dominant eigenvector of $\boldsymbol{Q}$ corresponds to a symmetric matrix rather than a skew-symmetric one. Hence, from here onwards we can use the fact that there exists a dominant eigenvector of $\boldsymbol{Q}$ that corresponds to a symmetric matrix.

Let $\tilde{\boldsymbol{z}}$ be a dominant eigenvector of $\boldsymbol{Q}$ which correspond to a symmetric matrix, and let $\tilde{\boldsymbol{Z}} = \tilde{\boldsymbol{V}}\tilde{\boldsymbol{S}}\tilde{\boldsymbol{V}}^{\mathrm{T}}$ be its spectral decomposition. Set

$$\boldsymbol{\Psi} = \sum_{i=1}^{M} \tilde{\boldsymbol{M}}_i^{\odot 2}, \qquad \text{s.t.} \qquad \tilde{\boldsymbol{M}}_i = \tilde{\boldsymbol{V}}^{\mathrm{T}}\boldsymbol{Y}_i\tilde{\boldsymbol{V}}. \tag{149}$$

Since $\boldsymbol{Q}$ is symmetric, then by the spectral theorem we have that all its eigenvectors and eigenvalues are real, and they are given by the quadratic form using the corresponding eigenvectors. Thus,

$$\begin{aligned}
\rho(\boldsymbol{Q}) &= \left|\tilde{\boldsymbol{z}}^{\mathrm{T}}\boldsymbol{Q}\tilde{\boldsymbol{z}}\right| \\
&= \left|\tilde{\boldsymbol{s}}_{\text{sym}}^{\mathrm{T}} \sum_{i=1}^{M} \tilde{\boldsymbol{M}}_i^{\odot 2} \tilde{\boldsymbol{s}}_{\text{sym}}\right| \\
&= \left|\tilde{\boldsymbol{s}}_{\text{sym}}^{\mathrm{T}}\boldsymbol{\Psi}\tilde{\boldsymbol{s}}_{\text{sym}}\right| \\
&= \left|\sum_{\ell=1}^{d}\sum_{p=1}^{d} \tilde{\boldsymbol{s}}_{\text{sym}\,[\ell]}\boldsymbol{\Psi}_{[\ell,p]}\tilde{\boldsymbol{s}}_{\text{sym}\,[p]}\right| \\
&\leq \sum_{\ell=1}^{d}\sum_{p=1}^{d} \left|\tilde{\boldsymbol{s}}_{\text{sym}\,[\ell]}\right| \boldsymbol{\Psi}_{[\ell,p]} \left|\tilde{\boldsymbol{s}}_{\text{sym}\,[p]}\right| \\
&= \left[\tilde{\boldsymbol{s}}_{\text{sym}}^{\text{abs}}\right]^{\mathrm{T}} \boldsymbol{\Psi} \left[\tilde{\boldsymbol{s}}_{\text{sym}}^{\text{abs}}\right] \\
&= \left[\tilde{\boldsymbol{z}}^{\text{abs}}\right]^{\mathrm{T}} \boldsymbol{Q} \left[\tilde{\boldsymbol{z}}^{\text{abs}}\right],
\end{aligned} \tag{150}$$

where $\tilde{\boldsymbol{s}}_{\text{sym}}^{\text{abs}}$ is the element-wise absolute value of $\tilde{\boldsymbol{s}}_{\text{sym}}$, and $\tilde{\boldsymbol{z}}^{\text{abs}} = \text{vec}(\tilde{\boldsymbol{V}}\tilde{\boldsymbol{S}}^{\text{abs}}\tilde{\boldsymbol{V}}^{\mathrm{T}})$. Namely, the vector $\tilde{\boldsymbol{z}}^{\text{abs}}$ that corresponds to the matrix built from the element-wise absolute value of the spectrum of $\tilde{\boldsymbol{Z}}$ yields a greater or equal result than the spectral radius of $\boldsymbol{Q}$, while still having a unit Euclidean norm. Thus, either $\tilde{\boldsymbol{s}}_{\text{sym}}^{\text{abs}} = \tilde{\boldsymbol{s}}_{\text{sym}}$, in which case $\tilde{\boldsymbol{Z}}$ is PSD, or $\tilde{\boldsymbol{s}}_{\text{sym}}^{\text{abs}} \neq \tilde{\boldsymbol{s}}_{\text{sym}}$, and then both $\tilde{\boldsymbol{z}}$ and $\tilde{\boldsymbol{z}}^{\text{abs}}$ are dominant eigenvectors (or else we get a contradiction). Note that $\text{vec}^{-1}(\tilde{\boldsymbol{z}}^{\text{abs}})$ is in fact a PSD matrix. Therefore, there is always a dominant eigenvector for $\boldsymbol{Q}$ which corresponds to a PSD matrix. Additionally, since $\max_j |\lambda_j(\boldsymbol{Q})| = \rho(\boldsymbol{Q}) = [\tilde{\boldsymbol{z}}^{\text{abs}}]^{\mathrm{T}}\boldsymbol{Q}[\tilde{\boldsymbol{z}}^{\text{abs}}]$, then $[\tilde{\boldsymbol{z}}^{\text{abs}}]$ is also a top eigenvector which corresponds to $\lambda_{\max}(\boldsymbol{Q})$, *i.e.,* $\rho(\boldsymbol{Q}) = \lambda_{\max}(\boldsymbol{Q})$ ($\lambda_{\max}(\boldsymbol{Q})$ is a dominant eigenvalue).

C.1. Quadratic form calculation for symmetric and skew-symmetric matrices

Let $\boldsymbol{z} = \text{vec}(\boldsymbol{Z})$, and assume that $\boldsymbol{Z} = \boldsymbol{V}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}$ where $\boldsymbol{V}$ is orthogonal matrix. Then

$$\begin{aligned}
\boldsymbol{z}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{z} &= \left[\text{vec}\left(\boldsymbol{V}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}\right)\right]^{\mathrm{T}} \boldsymbol{Q}\, \text{vec}\left(\boldsymbol{V}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}\right) \\
&= \left[(\boldsymbol{V}\otimes\boldsymbol{V})\,\text{vec}\,(\boldsymbol{S})\right]^{\mathrm{T}} \boldsymbol{Q}\,(\boldsymbol{V}\otimes\boldsymbol{V})\,\text{vec}\,(\boldsymbol{S}) \\
&= \left[\text{vec}\,(\boldsymbol{S})\right]^{\mathrm{T}} \left(\boldsymbol{V}^{\mathrm{T}}\otimes\boldsymbol{V}^{\mathrm{T}}\right) \sum_{i=1}^{M} \boldsymbol{Y}_i\otimes\boldsymbol{Y}_i \Big(\boldsymbol{V}\otimes\boldsymbol{V}\Big)\text{vec}\,(\boldsymbol{S}) \\
&= \left[\text{vec}\,(\boldsymbol{S})\right]^{\mathrm{T}} \sum_{i=1}^{M} \left(\boldsymbol{V}^{\mathrm{T}}\boldsymbol{Y}_i\boldsymbol{V}\right)\otimes\left(\boldsymbol{V}^{\mathrm{T}}\boldsymbol{Y}_i\boldsymbol{V}\right)\text{vec}\,(\boldsymbol{S}) \\
&= \sum_{i=1}^{M} \left[\text{vec}\,(\boldsymbol{S})\right]^{\mathrm{T}} \boldsymbol{M}_i\otimes\boldsymbol{M}_i\ \text{vec}\,(\boldsymbol{S})\,.
\end{aligned} \tag{151}$$

Writing the quadratic form explicitly for each $i \in [M]$ we have

$$[\mathrm{vec}\,(\boldsymbol{S})]^{\mathrm{T}}\, \boldsymbol{M}_i \otimes \boldsymbol{M}_i\ \mathrm{vec}\,(\boldsymbol{S}) = \sum_{m=1}^{d^2} \sum_{k=1}^{d^2} [\boldsymbol{M}_i \otimes \boldsymbol{M}_i]_{[m,k]}\ \mathrm{vec}\,(\boldsymbol{S})_{[m]}\ \mathrm{vec}\,(\boldsymbol{S})_{[k]}\,. \tag{152}$$

Set $m = d(m_2 - 1) + m_1$ and $k = d(k_2 - 1) + k_1$ where $m_1, m_2, k_1, k_2 \in [d]$, then

$$[\boldsymbol{M}_i \otimes \boldsymbol{M}_i]_{[m,k]} = [\boldsymbol{M}_i \otimes \boldsymbol{M}_i]_{[d(m_2-1)+m_1, d(k_2-1)+k_1]} = \boldsymbol{M}_{i\,[m_1,k_1]} \boldsymbol{M}_{i\,[m_2,k_2]}. \tag{153}$$

Moreover,

$$\begin{aligned} [\mathrm{vec}\,(\boldsymbol{S})]_{[m]} &= [\mathrm{vec}\,(\boldsymbol{S})]_{[d(m_2-1)+m_1]} = \boldsymbol{S}_{[m_1,m_2]}, \\ [\mathrm{vec}\,(\boldsymbol{S})]_{[k]} &= [\mathrm{vec}\,(\boldsymbol{S})]_{[d(k_2-1)+k_1]} = \boldsymbol{S}_{[k_1,k_2]}. \end{aligned} \tag{154}$$

Therefore,

$$\begin{aligned} &[\mathrm{vec}\,(\boldsymbol{S})]^{\mathrm{T}}\, \boldsymbol{M}_i \otimes \boldsymbol{M}_i\ \mathrm{vec}\,(\boldsymbol{S}) \\ &= \sum_{m=1}^{d^2} \sum_{k=1}^{d^2} [\boldsymbol{M}_i \otimes \boldsymbol{M}_i]_{[m,k]}\ [\mathrm{vec}\,(\boldsymbol{S})]_{[m]}\ [\mathrm{vec}\,(\boldsymbol{S})]_{[k]} \\ &= \sum_{m_2=1}^{d} \sum_{m_1=1}^{d} \sum_{k_2=1}^{d} \sum_{k_1=1}^{d} [\boldsymbol{M}_i \otimes \boldsymbol{M}_i]_{[d(m_2-1)+m_1, d(k_2-1)+k_1]}\ [\mathrm{vec}\,(\boldsymbol{S})]_{[d(m_2-1)+m_1]}\ [\mathrm{vec}\,(\boldsymbol{S})]_{[d(k_2-1)+k_1]} \\ &= \sum_{m_2=1}^{d} \sum_{m_1=1}^{d} \sum_{k_2=1}^{d} \sum_{k_1=1}^{d} \boldsymbol{M}_{i\,[m_1,k_1]} \boldsymbol{M}_{i\,[m_2,k_2]} \boldsymbol{S}_{[m_1,m_2]} \boldsymbol{S}_{[k_1,k_2]}, \end{aligned} \tag{155}$$

where in the last step we used (153) and (154).

C.1.1. SYMMETRIC EIGENVECTORS

Assume that $\boldsymbol{Z}$ is symmetric, then $\boldsymbol{S}$ is a diagonal matrix. Therefore, we only need to consider the terms in the series above for which $m_1 = m_2 = \ell$ and $k_1 = k_2 = p$.

$$\sum_{\ell=1}^{d} \sum_{p=1}^{d} \boldsymbol{M}_{i\,[\ell,p]} \boldsymbol{M}_{i\,[\ell,p]} \boldsymbol{S}_{[\ell,\ell]} \boldsymbol{S}_{[p,p]} = \sum_{\ell=1}^{d} \sum_{p=1}^{d} \boldsymbol{M}^2_{i\,[\ell,p]} \boldsymbol{s}_{\mathrm{sym}\,[\ell]} \boldsymbol{s}_{\mathrm{sym}\,[p]} = \boldsymbol{s}_{\mathrm{sym}}^{\mathrm{T}} \boldsymbol{M}_i^{\odot 2} \boldsymbol{s}_{\mathrm{sym}}. \tag{156}$$

Overall,

$$\boldsymbol{z}^{\mathrm{T}} \boldsymbol{Q} \boldsymbol{z} = \boldsymbol{s}_{\mathrm{sym}}^{\mathrm{T}} \sum_{i=1}^{M} \boldsymbol{M}_i^{\odot 2} \boldsymbol{s}_{\mathrm{sym}}. \tag{157}$$

C.1.2. SKEW-SYMMETRIC EIGENVECTORS

Assume that $\boldsymbol{Z}$ is skew-symmetric, then $\boldsymbol{S}$ is a block diagonal matrix, where each block is 2×2 in the form of

$$\begin{bmatrix} 0 & s_\ell \\ -s_\ell & 0 \end{bmatrix}. \tag{158}$$

If the dimension d is odd, then $\boldsymbol{S}$ has a row and column at the end filled with zeros. Here, the nonzero elements are located above and below the main diagonal of $\boldsymbol{S}$ and they come in pairs. Specifically, the following relations hold.

$$\boldsymbol{S}_{[2\ell-1,2\ell]} = s_\ell = \boldsymbol{s}_{\text{skew}\,[\ell]}, \qquad \boldsymbol{S}_{[2\ell,2\ell-1]} = -s_\ell = -\boldsymbol{s}_{\text{skew}\,[\ell]}. \tag{159}$$

Therefore, in the skew-symmetric scenario, we have four different cases to consider in the last line of (155). The four cases are as follows.

Case I: $m_1 = 2\ell-1,\ m_2 = 2\ell,\ k_1 = 2p-1,\ k_2 = 2p$**.**

$$\sum_{\ell=1}^{\lfloor d/2\rfloor}\sum_{p=1}^{\lfloor d/2\rfloor} \boldsymbol{M}_{i\,[2\ell-1,2p-1]}\boldsymbol{M}_{i\,[2\ell,2p]}\boldsymbol{S}_{[2\ell-1,2\ell]}\boldsymbol{S}_{[2p-1,2p]} = \sum_{\ell=1}^{\lfloor d/2\rfloor}\sum_{p=1}^{\lfloor d/2\rfloor} \boldsymbol{M}_{i\,[2\ell-1,2p-1]}\boldsymbol{M}_{i\,[2\ell,2p]}\boldsymbol{s}_{\text{skew}\,[\ell]}\boldsymbol{s}_{\text{skew}\,[p]}. \tag{160}$$

Case II: $m_1 = 2\ell,\ m_2 = 2\ell-1,\ k_1 = 2p-1,\ k_2 = 2p$**.**

$$\sum_{\ell=1}^{\lfloor d/2\rfloor}\sum_{p=1}^{\lfloor d/2\rfloor} \boldsymbol{M}_{i\,[2\ell,2p-1]}\boldsymbol{M}_{i\,[2\ell-1,2p]}\boldsymbol{S}_{[2\ell,2\ell-1]}\boldsymbol{S}_{[2p-1,2p]} = -\sum_{\ell=1}^{\lfloor d/2\rfloor}\sum_{p=1}^{\lfloor d/2\rfloor} \boldsymbol{M}_{i\,[2\ell,2p-1]}\boldsymbol{M}_{i\,[2\ell-1,2p]}\boldsymbol{s}_{\text{skew}\,[\ell]}\boldsymbol{s}_{\text{skew}\,[p]}. \tag{161}$$

Case III: $m_1 = 2\ell-1,\ m_2 = 2\ell,\ k_1 = 2p,\ k_2 = 2p-1$**.**

$$\sum_{\ell=1}^{\lfloor d/2\rfloor}\sum_{p=1}^{\lfloor d/2\rfloor} \boldsymbol{M}_{i\,[2\ell-1,2p]}\boldsymbol{M}_{i\,[2\ell,2p-1]}\boldsymbol{S}_{[2\ell-1,2\ell]}\boldsymbol{S}_{[2p,2p-1]} = -\sum_{\ell=1}^{\lfloor d/2\rfloor}\sum_{p=1}^{\lfloor d/2\rfloor} \boldsymbol{M}_{i\,[2\ell-1,2p]}\boldsymbol{M}_{i\,[2\ell,2p-1]}\boldsymbol{s}_{\text{skew}\,[\ell]}\boldsymbol{s}_{\text{skew}\,[p]}. \tag{162}$$

Case IV: $m_1 = 2\ell,\ m_2 = 2\ell-1,\ k_1 = 2p,\ k_2 = 2p-1$**.**

$$\sum_{\ell=1}^{\lfloor d/2\rfloor}\sum_{p=1}^{\lfloor d/2\rfloor} \boldsymbol{M}_{i\,[2\ell,2p]}\boldsymbol{M}_{i\,[2\ell-1,2p-1]}\boldsymbol{S}_{[2\ell,2\ell-1]}\boldsymbol{S}_{[2p,2p-1]} = \sum_{\ell=1}^{\lfloor d/2\rfloor}\sum_{p=1}^{\lfloor d/2\rfloor} \boldsymbol{M}_{i\,[2\ell,2p]}\boldsymbol{M}_{i\,[2\ell-1,2p-1]}\boldsymbol{s}_{\text{skew}\,[\ell]}\boldsymbol{s}_{\text{skew}\,[p]}. \tag{163}$$

Summing over all these cases we get

$$\begin{aligned}
\boldsymbol{z}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{z} &= \sum_{i=1}^{M}\sum_{\ell=1}^{\lfloor d/2\rfloor}\sum_{p=1}^{\lfloor d/2\rfloor} 2\left(\boldsymbol{M}_{i\,[2\ell-1,2p-1]}\boldsymbol{M}_{i\,[2\ell,2p]} - \boldsymbol{M}_{i\,[2\ell-1,2p]}\boldsymbol{M}_{i\,[2\ell,2p-1]}\right)\boldsymbol{s}_{\text{skew}\,[\ell]}\boldsymbol{s}_{\text{skew}\,[p]} \\
&= 2\sum_{i=1}^{M}\sum_{\ell=1}^{\lfloor d/2\rfloor}\sum_{p=1}^{\lfloor d/2\rfloor} \boldsymbol{T}_{i\,[\ell,p]}\boldsymbol{s}_{\text{skew}\,[\ell]}\boldsymbol{s}_{\text{skew}\,[p]} \\
&= 2\boldsymbol{s}_{\text{skew}}^{\mathrm{T}}\sum_{i=1}^{M}\boldsymbol{T}_i\boldsymbol{s}_{\text{skew}}.
\end{aligned} \tag{164}$$

Appendix D. Proof of Proposition 6

Here we focus on interpolating minima for simplicity. A similar proof can be derived for regular minima. To begin with, note that (see (101))

$$\lambda_{\max}\left(\boldsymbol{C}^{\dagger}\boldsymbol{D}\right)=\lambda_{\max}\left(\left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger}\boldsymbol{D}\left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger}\right). \tag{165}$$

Additionally,

$$\begin{aligned}\boldsymbol{D}&=(1-p)\,\boldsymbol{H}\otimes\boldsymbol{H}+p\,\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{H}_i\otimes\boldsymbol{H}_i\\&=\boldsymbol{H}\otimes\boldsymbol{H}+p\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{H}_i\otimes\boldsymbol{H}_i-\boldsymbol{H}\otimes\boldsymbol{H}\right)\\&=\boldsymbol{H}\otimes\boldsymbol{H}+p\left(\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{H}_i-\boldsymbol{H})\otimes(\boldsymbol{H}_i-\boldsymbol{H})\right)\\&=\boldsymbol{H}\otimes\boldsymbol{H}+p\boldsymbol{E},\end{aligned} \tag{166}$$

where in the third step we used (65), and at the last step $\boldsymbol{E}\triangleq\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{H}_i-\boldsymbol{H})\otimes(\boldsymbol{H}_i-\boldsymbol{H})$. Let $\boldsymbol{y}\in\mathbb{S}^{d^2-1}$ be the top eigenvector of $(\boldsymbol{C}^{\frac{1}{2}})^{\dagger}\boldsymbol{D}(\boldsymbol{C}^{\frac{1}{2}})^{\dagger}$, then since $(\boldsymbol{C}^{\frac{1}{2}})^{\dagger}\boldsymbol{D}(\boldsymbol{C}^{\frac{1}{2}})^{\dagger}$ is symmetric we have

$$\begin{aligned}\frac{\partial}{\partial p}\lambda_{\max}\left(\left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger}\boldsymbol{D}\left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger}\right)&=\boldsymbol{y}^{\mathrm{T}}\left[\left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger}\left(\frac{\partial}{\partial p}\boldsymbol{D}\right)\left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger}\right]\boldsymbol{y}\\&=\boldsymbol{y}^{\mathrm{T}}\left[\left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger}\boldsymbol{E}\left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger}\right]\boldsymbol{y}.\end{aligned} \tag{167}$$

In App. D.1 we show that $\boldsymbol{y}$ has the form of $\boldsymbol{y}=\boldsymbol{C}^{\frac{1}{2}}\boldsymbol{u}$ such that $\mathrm{vec}^{-1}(\boldsymbol{u})\in\mathcal{S}_{+}(\mathbb{R}^{d\times d})$. Plugging this into the equation above we get

$$\begin{aligned}\boldsymbol{y}^{\mathrm{T}}\left[\left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger}\boldsymbol{E}\left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger}\right]\boldsymbol{y}&=\boldsymbol{u}^{\mathrm{T}}\boldsymbol{C}^{\frac{1}{2}}\left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger}\boldsymbol{E}\left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger}\boldsymbol{C}^{\frac{1}{2}}\boldsymbol{u}\\&=\boldsymbol{u}^{\mathrm{T}}\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{C})}\boldsymbol{E}\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{C})}\boldsymbol{u}\\&=\boldsymbol{u}^{\mathrm{T}}\boldsymbol{E}\boldsymbol{u},\end{aligned} \tag{168}$$

where in the first step we used the fact that $\boldsymbol{C}$ is symmetric. Additionally, in the second step, we used the fact that $\boldsymbol{C}^{\frac{1}{2}}(\boldsymbol{C}^{\frac{1}{2}})^{\dagger}$ and $(\boldsymbol{C}^{\frac{1}{2}})^{\dagger}\boldsymbol{C}^{\frac{1}{2}}$ are projection matrices onto the column space of $\boldsymbol{C}$. Since the null space of $\boldsymbol{E}$ contains the null space of $\boldsymbol{C}$, we have that these projections can be removed (see App. B.6). Note that $\mathrm{vec}^{-1}(\boldsymbol{u})$ is PSD, and let $\boldsymbol{V}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}$ be its spectral decomposition, then in App. C.1 we show that in this case

$$\boldsymbol{u}^{\mathrm{T}}\boldsymbol{E}\boldsymbol{u}=\boldsymbol{s}^{\mathrm{T}}\sum_{i=1}^{n}\boldsymbol{M}_i^{\odot 2}\boldsymbol{s}, \tag{169}$$

where $\boldsymbol{M}_i = \boldsymbol{V}^{\mathrm{T}}(\boldsymbol{H}_i - \boldsymbol{H})\boldsymbol{V}$ and $\boldsymbol{s}$ is a vector containing the eigenvalues of $\mathrm{vec}^{-1}(\boldsymbol{u})$. Since $\mathrm{vec}^{-1}(\boldsymbol{u})$ is PSD, we have the right-hand side of (169) is a sum over nonnegative terms. Namely,

$$\frac{\partial}{\partial p}\lambda_{\max}\left(\left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger}\boldsymbol{D}\left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger}\right) = \boldsymbol{u}^{\mathrm{T}}\boldsymbol{E}\boldsymbol{u} = \boldsymbol{s}^{\mathrm{T}}\sum_{i=1}^{n}\boldsymbol{M}_i^{\odot 2}\boldsymbol{s} \geq 0. \tag{170}$$

Therefore, $\lambda_{\max}\left(\boldsymbol{C}^{\dagger}\boldsymbol{D}\right)$ is monotonically non-decreasing in p, which means that $\eta^{*}_{\mathrm{var}} = 2/\lambda_{\max}\left(\boldsymbol{C}^{\dagger}\boldsymbol{D}\right)$ is monotonically non-decreasing with B.

D.1. Top eigenvector of $(\boldsymbol{C}^{\frac{1}{2}})^{\dagger}\boldsymbol{D}(\boldsymbol{C}^{\frac{1}{2}})^{\dagger}$

Using the stability condition of Ma and Ying (2021), we have that $\{\mathbb{E}[\|\boldsymbol{\theta}_t - \boldsymbol{\theta}\|^2]\}$ is bounded if and only if (see proof in (Ma and Ying, 2021))

$$\max_{\boldsymbol{\Sigma}\in\mathcal{S}_+(\mathbb{R}^{d\times d})}\frac{\|\boldsymbol{Q}(\eta,B)\ \mathrm{vec}\left(\boldsymbol{\Sigma}\right)\|}{\|\boldsymbol{\Sigma}\|_{\mathrm{F}}} \leq 1. \tag{171}$$

Let us repeat the same steps from the proof of Thm. 5 in App. B.8 but *without* relaxing the constraint of PSD matrices. Specifically, repeating the steps in equations in (91)-(94) without invoking Thm. 14 gives us that

$$\boldsymbol{u}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{u} = 1 - 2\eta\boldsymbol{u}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{u} + \eta^2\boldsymbol{u}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{u} \leq 1 \tag{172}$$

holds for any $\boldsymbol{u}\in\mathbb{S}^{d^2-1}$ such that $\mathrm{vec}^{-1}(\boldsymbol{u})\in\mathcal{S}_+(\mathbb{R}^{d\times d})$ and $\boldsymbol{u}\notin\mathcal{N}(\boldsymbol{D})$, if and only if

$$\eta \leq \frac{2}{\lambda^*} = \eta^*_{\mathrm{var}}, \tag{173}$$

where

$$\lambda^* = \sup_{\boldsymbol{u}\in\mathbb{S}^{d^2-1}}\left\{\frac{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{u}}{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{u}}\right\} \quad \text{s.t.} \quad \mathrm{vec}^{-1}(\boldsymbol{u})\in\mathcal{S}_+(\mathbb{R}^{d\times d}) \text{ and } \boldsymbol{u}\notin\mathcal{N}(\boldsymbol{D}). \tag{174}$$

(Note that the case $\boldsymbol{u}\in\mathcal{N}(\boldsymbol{D})$ do not contribute any conditions on the learning rate, and therefore can be ignored - see App. B.8). Using change of variables (see (97) and (98)) results with

$$\lambda^* = \max_{\boldsymbol{y}\in\mathbb{S}^{d^2-1}}\boldsymbol{y}^{\mathrm{T}}\left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger}\boldsymbol{D}\left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger}\boldsymbol{y} \quad \text{s.t.} \quad \boldsymbol{y} = (\boldsymbol{C}^{\frac{1}{2}}\boldsymbol{u}) \text{ and } \mathrm{vec}^{-1}(\boldsymbol{u})\in\mathcal{S}_+(\mathbb{R}^{d\times d}). \tag{175}$$

Since the alternative form of η^*_{var} in (173) has to be equal to the definition in (13) (or else we will get a contradiction), we get

$$\lambda_{\max}\left(\left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger}\boldsymbol{D}\left(\boldsymbol{C}^{\frac{1}{2}}\right)^{\dagger}\right) = \lambda^*. \tag{176}$$

Namely, the top eigenvector $\boldsymbol{y}$ of $(\boldsymbol{C}^{\frac{1}{2}})^{\dagger}\boldsymbol{D}(\boldsymbol{C}^{\frac{1}{2}})^{\dagger}$ has the form of $\boldsymbol{y} = \boldsymbol{C}^{\frac{1}{2}}\boldsymbol{u}$ such that $\mathrm{vec}^{-1}(\boldsymbol{u})\in\mathcal{S}_+(\mathbb{R}^{d\times d})$.

Appendix E. Proof of Proposition 7

Here we focus on interpolating minima for simplicity. A similar proof can be derived for regular minima. Let $\{\beta_t\}$ and $\{\kappa_t\}$ be i.i.d. random variables such that $\beta_t \sim \text{Bernoulli}(p)$ and $\kappa_t \sim \mathcal{U}(\{1, ..., n\})$, then

$$\mathfrak{B}_t = \begin{cases} \kappa_t & \text{if } \beta_t = 1, \\ \{1, ..., n\} & \text{otherwise.} \end{cases} \tag{177}$$

Let us consider the following stochastic loss function

$$\hat{\mathcal{L}}_t(\boldsymbol{\theta}) = \frac{1}{|\mathfrak{B}_t|} \sum_{i \in \mathfrak{B}_t} \ell_i(\boldsymbol{\theta}), \tag{178}$$

where $|\mathfrak{B}_t|$ denotes the size of $\mathfrak{B}_t$ (either 1 or n), and define the following notation

$$\mathcal{A}_t = \boldsymbol{I} - \frac{\eta}{|\mathfrak{B}_t|} \sum_{i \in \mathfrak{B}_t} \boldsymbol{H}_i. \tag{179}$$

First, for interpolating minima we have

$$\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^* = \left(\boldsymbol{I} - \frac{\eta}{|\mathfrak{B}_t|} \sum_{i \in \mathfrak{B}_t} \boldsymbol{H}_i \right) (\boldsymbol{\theta}_t - \boldsymbol{\theta}^*) = \mathcal{A}_t (\boldsymbol{\theta}_t - \boldsymbol{\theta}^*). \tag{180}$$

Thus,

$$\begin{aligned} \boldsymbol{\Sigma}_{t+1} &= \mathbb{E}\left[(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^*) (\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}^*)^{\mathrm{T}} \right] \\ &= \mathbb{E}\left[\mathcal{A}_t (\boldsymbol{\theta}_t - \boldsymbol{\theta}^*) (\boldsymbol{\theta}_t - \boldsymbol{\theta}^*)^{\mathrm{T}} \mathcal{A}_t \right] \\ &= \mathbb{E}\left[\mathcal{A}_t \mathbb{E}\left[(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*) (\boldsymbol{\theta}_t - \boldsymbol{\theta}^*)^{\mathrm{T}} \middle| \mathfrak{B}_t \right] \mathcal{A}_t \right] \\ &= \mathbb{E}\left[\mathcal{A}_t \mathbb{E}\left[(\boldsymbol{\theta}_t - \boldsymbol{\theta}^*) (\boldsymbol{\theta}_t - \boldsymbol{\theta}^*)^{\mathrm{T}} \right] \mathcal{A}_t \right] \\ &= \mathbb{E}\left[\mathcal{A}_t \boldsymbol{\Sigma}_t \mathcal{A}_t \right], \end{aligned} \tag{181}$$

where in the second step we used (180), in the third we used the law of total expectation, and in the fourth we used the fact that $\boldsymbol{\theta}_t$ is statistically independent of $\mathfrak{B}_t$. Using vectorization we get

$$\text{vec}\left(\boldsymbol{\Sigma}_{t+1} \right) = \mathbb{E}\left[\mathcal{A}_t \otimes \mathcal{A}_t \right] \text{vec}\left(\boldsymbol{\Sigma}_t \right). \tag{182}$$

In (64) we show that for any given (fixed) batch size,

$$\boldsymbol{Q}(B, \eta) = \left(1 - \frac{n - B}{B(n-1)} \right) (\boldsymbol{I} - \eta \boldsymbol{H}) \otimes (\boldsymbol{I} - \eta \boldsymbol{H}) + \frac{n - B}{B(n-1)} \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{I} - \eta \boldsymbol{H}_i) \otimes (\boldsymbol{I} - \eta \boldsymbol{H}_i). \tag{183}$$

Specifically,

$$\boldsymbol{Q}(\eta, B = 1) = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{I} - \eta \boldsymbol{H}_i) \otimes (\boldsymbol{I} - \eta \boldsymbol{H}_i), \qquad \boldsymbol{Q}(\eta, B = n) = (\boldsymbol{I} - \eta \boldsymbol{H}) \otimes (\boldsymbol{I} - \eta \boldsymbol{H}). \tag{184}$$

Using this result, let us compute the term in (182).

$$\begin{aligned}\mathbb{E}\left[\mathcal{A}_t \otimes \mathcal{A}_t\right] &= \mathbb{P}\left(\beta_t = 0\right)\mathbb{E}\left[\mathcal{A}_t \otimes \mathcal{A}_t | \beta_t = 0\right] + \mathbb{P}\left(\beta_t = 1\right)\mathbb{E}\left[\mathcal{A}_t \otimes \mathcal{A}_t | \beta_t = 1\right] \\ &= (1-p)\boldsymbol{Q}(\eta, B = n) + p\boldsymbol{Q}(\eta, B = 1) \\ &= (1-p) \times (\boldsymbol{I} - \eta\boldsymbol{H}) \otimes (\boldsymbol{I} - \eta\boldsymbol{H}) + p \times \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{I} - \eta\boldsymbol{H}_i) \otimes (\boldsymbol{I} - \eta\boldsymbol{H}_i). \qquad (185)\end{aligned}$$

This is the same matrix that we had for mini-batch SGD with batch size B such that $p = \frac{n-B}{B(n-1)}$ (see (183)). Namely, the covariance matrix of the parameters for the mixed process evolves in the same way as mini-batch SGD, with a corresponding batch size. Therefore, the stability threshold of both algorithms is the same.

Appendix F. Proof of Proposition 8

First, since η^*_{var} is monotonically non-decreasing with B (Thm. 6), and for $B = n$ we have $\eta^*_{\text{var}} = \eta^*_{\text{mean}}$ (App. J), we get that $\eta^*_{\text{var}} \leq \eta^*_{\text{mean}}$ for all values of B. Now, set $\varepsilon \in (0, 1)$, then $(1-\varepsilon)\eta^*_{\text{mean}} \leq \eta^*_{\text{var}}$ holds whenever

$$\begin{aligned}(1-\varepsilon)\frac{2}{\lambda_{\max}(\boldsymbol{H})} &\leq \frac{2}{\lambda_{\max}(\boldsymbol{C}^\dagger\boldsymbol{D})} \\ \Leftrightarrow \quad (1-\varepsilon)\lambda_{\max}(\boldsymbol{C}^\dagger\boldsymbol{D}) &\leq \lambda_{\max}(\boldsymbol{H}). \qquad (186)\end{aligned}$$

Note that $\boldsymbol{D} = \boldsymbol{H} \otimes \boldsymbol{H} + p\boldsymbol{E}$ (see (166)), then

$$\begin{aligned}\lambda_{\max}\left(\boldsymbol{C}^\dagger\boldsymbol{D}\right) &= \lambda_{\max}\left(\boldsymbol{C}^\dagger\boldsymbol{H} \otimes \boldsymbol{H} + p\boldsymbol{C}^\dagger\boldsymbol{E}\right) \\ &\leq \lambda_{\max}\left(\boldsymbol{C}^\dagger\boldsymbol{H} \otimes \boldsymbol{H}\right) + p\lambda_{\max}\left(\boldsymbol{C}^\dagger\boldsymbol{E}\right) \\ &= \lambda_{\max}(\boldsymbol{H}) + p\lambda_{\max}\left(\boldsymbol{C}^\dagger\boldsymbol{E}\right), \qquad (187)\end{aligned}$$

where we used $\lambda_{\max}\left(\boldsymbol{C}^\dagger\boldsymbol{H} \otimes \boldsymbol{H}\right) = \lambda_{\max}(\boldsymbol{H})$ (see App. J). Using the fact that $\lambda_{\max}\left(\boldsymbol{C}^\dagger\boldsymbol{E}\right)$ is nonnegative (see App. F.1) and that

$$p = \frac{1}{B}\frac{n-B}{n-1} \leq \frac{1}{B}, \qquad (188)$$

we can further bound (187) from above by

$$\begin{aligned}\lambda_{\max}\left(\boldsymbol{C}^\dagger\boldsymbol{D}\right) &\leq \lambda_{\max}(\boldsymbol{H}) + p\lambda_{\max}\left(\boldsymbol{C}^\dagger\boldsymbol{E}\right) \\ &\leq \lambda_{\max}(\boldsymbol{H}) + \frac{1}{B}\lambda_{\max}\left(\boldsymbol{C}^\dagger\boldsymbol{E}\right). \qquad (189)\end{aligned}$$

Therefore, if

$$(1-\varepsilon)\left(\lambda_{\max}(\boldsymbol{H}) + \frac{1}{B}\lambda_{\max}\left(\boldsymbol{C}^\dagger\boldsymbol{E}\right)\right) \leq \lambda_{\max}(\boldsymbol{H}) \qquad (190)$$

then (186) holds. It is easy to show that (190) is equivalent to

$$B \geq \frac{1-\varepsilon}{\varepsilon}\frac{\lambda_{\max}(\boldsymbol{C}^\dagger\boldsymbol{E})}{\lambda_{\max}(\boldsymbol{H})}. \qquad (191)$$

Overall, if the batch size satisfies this inequality then $(1-\varepsilon)\eta^*_{\text{mean}} \leq \eta^*_{\text{var}} \leq \eta^*_{\text{mean}}$.

F.1. Proof that $\lambda_{\max}\left(\boldsymbol{C}^{\dagger}\boldsymbol{E}\right)$ is nonnegative

Since $\eta^*_{\text{var}} \leq \eta^*_{\text{mean}}$ for all values of B (see the beginning of this section), then $\lambda_{\max}(\boldsymbol{H}) \leq \lambda_{\max}(\boldsymbol{C}^{\dagger}\boldsymbol{D})$. Therefore,

$$0 \leq \lambda_{\max}(\boldsymbol{C}^{\dagger}\boldsymbol{D}) - \lambda_{\max}(\boldsymbol{H}) \leq p\lambda_{\max}\left(\boldsymbol{C}^{\dagger}\boldsymbol{E}\right), \tag{192}$$

where in the last step we used both ends of (187).

Appendix G. Proof of Proposition 9

The stability threshold given by Thm. 5 and Thm. 11 is

$$\eta^*_{\text{var}} = \frac{2}{\lambda_{\max}\left(\boldsymbol{C}^{\dagger}\boldsymbol{D}\right)} \tag{193}$$

where

$$\boldsymbol{C} = \frac{1}{2}\boldsymbol{H} \oplus \boldsymbol{H}, \qquad \boldsymbol{D} = (1-p)\,\boldsymbol{H} \otimes \boldsymbol{H} + p\,\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{H}_i \otimes \boldsymbol{H}_i. \tag{194}$$

This threshold corresponds to a necessary and sufficient condition for stability. Here we derive simplified necessary conditions for stability. In App. B.8 we show that (see (94))

$$\frac{2}{\lambda_{\max}\left(\boldsymbol{C}^{\dagger}\boldsymbol{D}\right)} = 2 \inf_{\boldsymbol{u}\in\mathbb{S}^{d^2-1}:\boldsymbol{u}\notin\mathcal{N}(\boldsymbol{D})} \left\{\frac{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{u}}{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{u}}\right\}. \tag{195}$$

We shall upper bound the stability threshold by considering non-optimal yet interesting vectors $\boldsymbol{u}$. Specifically, in the following we look at $\boldsymbol{u} = \boldsymbol{v}_{\max} \otimes \boldsymbol{v}_{\max}$, where $\boldsymbol{v}_{\max}$ is the top eigenvector of $\boldsymbol{H}$, and $\boldsymbol{u} = \text{vec}(\boldsymbol{I})$ to obtain the results of Proposition 9.

G.1. Setting $\boldsymbol{u} = \boldsymbol{v}_{\max} \otimes \boldsymbol{v}_{\max}$

Let $\boldsymbol{u} = \boldsymbol{v} \otimes \boldsymbol{v} \notin \mathcal{N}(\boldsymbol{D})$ where $\|\boldsymbol{v}\| = 1$, then

$$\begin{aligned}
\boldsymbol{u}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{u} &= \frac{1}{2}\boldsymbol{u}^{\mathrm{T}}\left(\boldsymbol{H}\otimes\boldsymbol{I} + \boldsymbol{I}\otimes\boldsymbol{H}\right)\boldsymbol{u} \\
&= \frac{1}{2}\left[\left(\boldsymbol{v}^{\mathrm{T}}\otimes\boldsymbol{v}^{\mathrm{T}}\right)\left(\boldsymbol{H}\otimes\boldsymbol{I}\right)\left(\boldsymbol{v}\otimes\boldsymbol{v}\right) + \left(\boldsymbol{v}^{\mathrm{T}}\otimes\boldsymbol{v}^{\mathrm{T}}\right)\left(\boldsymbol{I}\otimes\boldsymbol{H}\right)\left(\boldsymbol{v}\otimes\boldsymbol{v}\right)\right] \\
&= \frac{1}{2}\left[\left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{v}\right)\otimes\left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}\right) + \left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}\right)\otimes\left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{v}\right)\right] \\
&= \frac{1}{2}\left[\left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{v}\right)\otimes 1 + 1\otimes\left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{v}\right)\right] \\
&= \boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{v}.
\end{aligned} \tag{196}$$

Similarly,

$$\begin{aligned}
\boldsymbol{u}^{\mathrm{T}}\left(\boldsymbol{H}\otimes\boldsymbol{H}\right)\boldsymbol{u} &= \left(\boldsymbol{v}^{\mathrm{T}}\otimes\boldsymbol{v}^{\mathrm{T}}\right)\left(\boldsymbol{H}\otimes\boldsymbol{H}\right)\left(\boldsymbol{v}\otimes\boldsymbol{v}\right) \\
&= \left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{v}\right)\otimes\left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{v}\right) \\
&= \left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{v}\right)\otimes\left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{v}\right) \\
&= \left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{v}\right)^2.
\end{aligned} \tag{197}$$

And again

$$\boldsymbol{u}^{\mathrm{T}}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{H}_i\otimes\boldsymbol{H}_i\right)\boldsymbol{u} = \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{v}^{\mathrm{T}}\otimes\boldsymbol{v}^{\mathrm{T}}\right)\left(\boldsymbol{H}_i\otimes\boldsymbol{H}_i\right)\left(\boldsymbol{v}\otimes\boldsymbol{v}\right)$$
$$= \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}_i\boldsymbol{v}\right)\otimes\left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}_i\boldsymbol{v}\right)$$
$$= \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}_i\boldsymbol{v}\right)^2. \tag{198}$$

Thus,

$$\boldsymbol{u}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{u} = (1-p)\boldsymbol{u}^{\mathrm{T}}\left(\boldsymbol{H}\otimes\boldsymbol{H}\right)\boldsymbol{u} + p\boldsymbol{u}^{\mathrm{T}}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{H}_i\otimes\boldsymbol{H}_i\right)\boldsymbol{u}$$
$$= (1-p)\left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{v}\right)^2 + p\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}_i\boldsymbol{v}\right)^2$$
$$= \left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{v}\right)^2 + p\left[\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}_i\boldsymbol{v}\right)^2 - \left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{v}\right)^2\right]$$
$$= \left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{v}\right)^2 + p\left[\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}_i\boldsymbol{v}\right)^2 - 2\left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{v}\right)\left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{v}\right) + \left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{v}\right)^2\right]$$
$$= \left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{v}\right)^2 + p\frac{1}{n}\sum_{i=1}^{n}\left[\left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}_i\boldsymbol{v}\right)^2 - 2\left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}_i\boldsymbol{v}\right)\left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{v}\right) + \left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{v}\right)^2\right]$$
$$= \left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{v}\right)^2 + p\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}_i\boldsymbol{v} - \left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{v}\right)\right)^2. \tag{199}$$

Therefore, for general $\boldsymbol{u}=\boldsymbol{v}\otimes\boldsymbol{v}$ we get

$$\eta^*_{\text{var}} \le 2\frac{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{u}}{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{u}} = \frac{2\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{v}}{(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{v})^2 + \frac{p}{n}\sum_{i=1}^{n}(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}_i\boldsymbol{v} - \boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{v})^2}. \tag{200}$$

Specifically, for $\boldsymbol{u}=\boldsymbol{v}_{\max}\otimes\boldsymbol{v}_{\max}$ we get

$$\eta^*_{\text{var}} \le \frac{2\lambda_{\max}(\boldsymbol{H})}{\lambda^2_{\max}(\boldsymbol{H}) + p\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{v}^{\mathrm{T}}_{\max}\boldsymbol{H}_i\boldsymbol{v}_{\max} - \lambda_{\max}(\boldsymbol{H})\right)^2}. \tag{201}$$

Finally, from (200) we get the following result which we used in Sec. 4.

$$\lambda_{\max}\left(\boldsymbol{C}^{\dagger}\boldsymbol{D}\right) = \frac{2}{\eta^*_{\text{var}}} \ge \boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{v} + p\frac{\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}_i\boldsymbol{v} - \boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{v})^2}{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{v}}. \tag{202}$$

Since this inequality holds for every $\boldsymbol{v}\notin\mathcal{N}(\boldsymbol{H})$, we can take the maximum to obtain (27).

G.2. Setting $\boldsymbol{u} = \text{vec}\,(\boldsymbol{I})$

Let $\boldsymbol{u} = \text{vec}(\boldsymbol{I}) \notin \mathcal{N}(\boldsymbol{D})$, then

$$\begin{aligned}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{u} &= \frac{1}{2}\boldsymbol{u}^{\mathrm{T}}\left(\boldsymbol{H}\otimes\boldsymbol{I}+\boldsymbol{I}\otimes\boldsymbol{H}\right)\boldsymbol{u}\\ &= \frac{1}{2}\left([\text{vec}\,(\boldsymbol{I})]^{\mathrm{T}}\left(\boldsymbol{H}\otimes\boldsymbol{I}\right)\text{vec}\,(\boldsymbol{I})+[\text{vec}\,(\boldsymbol{I})]^{\mathrm{T}}\left(\boldsymbol{I}\otimes\boldsymbol{H}\right)\text{vec}\,(\boldsymbol{I})\right)\\ &= \frac{1}{2}\left(\text{Tr}\big(\boldsymbol{H}^{\mathrm{T}}\big)+\text{Tr}\big(\boldsymbol{H}\big)\right)\\ &= \text{Tr}(\boldsymbol{H}),\end{aligned} \tag{203}$$

where in the third step we used (P4). Moreover, using (P4) we have

$$\boldsymbol{u}^{\mathrm{T}}\left(\boldsymbol{H}\otimes\boldsymbol{H}\right)\boldsymbol{u} = [\text{vec}\,(\boldsymbol{I})]^{\mathrm{T}}\left(\boldsymbol{H}\otimes\boldsymbol{H}\right)\text{vec}\,(\boldsymbol{I}) = \text{Tr}(\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}) = \|\boldsymbol{H}\|_{\mathrm{F}}^{2}\,. \tag{204}$$

Similarly,

$$\boldsymbol{u}^{\mathrm{T}}\left(\boldsymbol{H}_i\otimes\boldsymbol{H}_i\right)\boldsymbol{u} = [\text{vec}\,(\boldsymbol{I})]^{\mathrm{T}}\left(\boldsymbol{H}_i\otimes\boldsymbol{H}_i\right)\text{vec}\,(\boldsymbol{I}) = \text{Tr}(\boldsymbol{H}_i\boldsymbol{H}_i^{\mathrm{T}}) = \|\boldsymbol{H}_i\|_{\mathrm{F}}^{2}\,. \tag{205}$$

Then,

$$\begin{aligned}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{u} &= (1-p)\boldsymbol{u}^{\mathrm{T}}\left(\boldsymbol{H}\otimes\boldsymbol{H}\right)\boldsymbol{u}+p\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{H}_i\otimes\boldsymbol{H}_i\boldsymbol{u}\\ &= (1-p)\left\|\boldsymbol{H}\right\|_{\mathrm{F}}^{2}+p\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{H}_i\right\|_{\mathrm{F}}^{2}.\end{aligned} \tag{206}$$

Therefore,

$$\eta_{\text{var}}^{*} \le 2\frac{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{u}}{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{u}} = \frac{2\text{Tr}(\boldsymbol{H})}{(1-p)\left\|\boldsymbol{H}\right\|_{\mathrm{F}}^{2}+p\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{H}_i\right\|_{\mathrm{F}}^{2}}. \tag{207}$$

Appendix H. Proof of Theorem 12

In this section, we use the following result on the Moore–Penrose inverse of a sum of two matrices.

Theorem 18 (Fill and Fishkind (2000), Thm. 3) *Let* $\boldsymbol{X},\boldsymbol{Y}\in\mathbb{R}^{p\times p}$ *with* $\text{rank}(\boldsymbol{X}+\boldsymbol{Y}) = \text{rank}(\boldsymbol{X})+\text{rank}(\boldsymbol{Y})$. *Then*

$$(\boldsymbol{X}+\boldsymbol{Y})^{\dagger} = (\boldsymbol{I}-\boldsymbol{L})\boldsymbol{X}^{\dagger}(\boldsymbol{I}-\boldsymbol{O})+\boldsymbol{L}\boldsymbol{Y}^{\dagger}\boldsymbol{O}, \tag{208}$$

where

$$\boldsymbol{L} = \left(\boldsymbol{P}_{\mathcal{R}(\boldsymbol{Y}^{\mathrm{T}})}\boldsymbol{P}_{\mathcal{R}^{\perp}(\boldsymbol{X}^{\mathrm{T}})}\right)^{\dagger} \qquad \textit{and} \qquad \boldsymbol{O} = \left(\boldsymbol{P}_{\mathcal{R}^{\perp}(\boldsymbol{X})}\boldsymbol{P}_{\mathcal{R}(\boldsymbol{Y})}\right)^{\dagger}. \tag{209}$$

Moreover, we use the following relations.

$$\begin{aligned}&\mathcal{R}(\boldsymbol{D}) = \mathcal{R}(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}),\\ &\mathcal{R}(\boldsymbol{C}) = \mathcal{R}(\boldsymbol{I}-\boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})}),\\ &\mathcal{R}(\boldsymbol{P}_{\mathcal{N}(\boldsymbol{D})}\boldsymbol{C}) = \mathcal{R}(\boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}+\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})}).\end{aligned} \tag{210}$$

The dynamics of $\boldsymbol{\mu}_t^\perp$ and $\boldsymbol{\Sigma}_t^\perp$ are given by (see (56))

$$\begin{pmatrix} \boldsymbol{\mu}_{t+1}^\perp \\ \operatorname{vec}\left(\boldsymbol{\Sigma}_{t+1}^\perp\right) \end{pmatrix} = \boldsymbol{\Xi} \begin{pmatrix} \boldsymbol{\mu}_{t}^\perp \\ \operatorname{vec}\left(\boldsymbol{\Sigma}_{t}^\perp\right) \end{pmatrix} + \begin{pmatrix} \mathbf{0} \\ \operatorname{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{v}}^\perp\right) \end{pmatrix}. \tag{211}$$

where

$$\begin{aligned} \boldsymbol{\Xi} &= \begin{pmatrix} \boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})} - \eta \boldsymbol{H} & \mathbf{0} \\ -\left(\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})} \otimes \boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})}\right)\left(\mathbb{E}\left[\boldsymbol{v}_t^\perp \otimes \boldsymbol{A}_t\right] + \mathbb{E}\left[\boldsymbol{A}_t \otimes \boldsymbol{v}_t^\perp\right]\right) & \left(\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})} \otimes \boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})}\right)\boldsymbol{Q} \end{pmatrix} \\ &\triangleq \begin{pmatrix} \boldsymbol{\Xi}_{1,1} & \boldsymbol{\Xi}_{1,2} \\ \boldsymbol{\Xi}_{2,1} & \boldsymbol{\Xi}_{2,2} \end{pmatrix}. \end{aligned} \tag{212}$$

In App. B.9 we show that if $0 < \eta < \eta_{\text{var}}^*$ then the spectral radius of $\boldsymbol{\Xi}$ is less then one. Therefore, the dynamical system is stable, and the asymptotic values of $\boldsymbol{\mu}_t^\perp$ and $\boldsymbol{\Sigma}_t^\perp$ as $t \to \infty$ are given by

$$\lim_{t\to\infty} \begin{pmatrix} \boldsymbol{\mu}_{t}^\perp \\ \operatorname{vec}\left(\boldsymbol{\Sigma}_{t}^\perp\right) \end{pmatrix} = \left(\boldsymbol{I} - \boldsymbol{\Xi}\right)^{-1} \begin{pmatrix} \mathbf{0} \\ \operatorname{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{v}}^\perp\right) \end{pmatrix}. \tag{213}$$

Using the inversion formula for block matrix and the fact that $\boldsymbol{\Xi}_{1,2} = \mathbf{0}$ we have that

$$\begin{aligned} \left(\boldsymbol{I} - \boldsymbol{\Xi}\right)^{-1} &= \begin{pmatrix} \boldsymbol{I} - \boldsymbol{\Xi}_{1,1} & -\boldsymbol{\Xi}_{1,2} \\ -\boldsymbol{\Xi}_{2,1} & \boldsymbol{I} - \boldsymbol{\Xi}_{2,2} \end{pmatrix}^{-1} \\ &= \begin{pmatrix} \left(\boldsymbol{I} - \boldsymbol{\Xi}_{1,1} - \boldsymbol{\Xi}_{1,2}\left(\boldsymbol{I} - \boldsymbol{\Xi}_{2,2}\right)^{-1}\boldsymbol{\Xi}_{2,1}\right)^{-1} & \mathbf{0} \\ \mathbf{0} & \left(\boldsymbol{I} - \boldsymbol{\Xi}_{2,2} - \boldsymbol{\Xi}_{2,1}\left(\boldsymbol{I} - \boldsymbol{\Xi}_{1,1}\right)^{-1}\boldsymbol{\Xi}_{1,2}\right)^{-1} \end{pmatrix} \\ &\quad \begin{pmatrix} \boldsymbol{I} & \boldsymbol{\Xi}_{1,2}\left(\boldsymbol{I} - \boldsymbol{\Xi}_{2,2}\right)^{-1} \\ \boldsymbol{\Xi}_{2,1}\left(\boldsymbol{I} - \boldsymbol{\Xi}_{1,1}\right)^{-1} & \boldsymbol{I} \end{pmatrix} \\ &= \begin{pmatrix} \left(\boldsymbol{I} - \boldsymbol{\Xi}_{1,1}\right)^{-1} & \mathbf{0} \\ \mathbf{0} & \left(\boldsymbol{I} - \boldsymbol{\Xi}_{2,2}\right)^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{I} & \mathbf{0} \\ \boldsymbol{\Xi}_{2,1}\left(\boldsymbol{I} - \boldsymbol{\Xi}_{1,1}\right)^{-1} & \boldsymbol{I} \end{pmatrix}. \end{aligned} \tag{214}$$

Therefore,

$$\begin{aligned} \lim_{t\to\infty} \begin{pmatrix} \boldsymbol{\mu}_{t}^\perp \\ \operatorname{vec}\left(\boldsymbol{\Sigma}_{t}^\perp\right) \end{pmatrix} &= \begin{pmatrix} \left(\boldsymbol{I} - \boldsymbol{\Xi}_{1,1}\right)^{-1} & \mathbf{0} \\ \mathbf{0} & \left(\boldsymbol{I} - \boldsymbol{\Xi}_{2,2}\right)^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{I} & \mathbf{0} \\ \boldsymbol{\Xi}_{2,1}\left(\boldsymbol{I} - \boldsymbol{\Xi}_{1,1}\right)^{-1} & \boldsymbol{I} \end{pmatrix} \begin{pmatrix} \mathbf{0} \\ \operatorname{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{v}}^\perp\right) \end{pmatrix} \\ &= \begin{pmatrix} \left(\boldsymbol{I} - \boldsymbol{\Xi}_{1,1}\right)^{-1} & \mathbf{0} \\ \mathbf{0} & \left(\boldsymbol{I} - \boldsymbol{\Xi}_{2,2}\right)^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{0} \\ \operatorname{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{v}}^\perp\right) \end{pmatrix}. \end{aligned} \tag{215}$$

Namely,

$$\lim_{t\to\infty} \boldsymbol{\mu}_t^\perp = \mathbf{0} \qquad \text{and} \qquad \lim_{t\to\infty} \operatorname{vec}\left(\boldsymbol{\Sigma}_t^\perp\right) = \left(\boldsymbol{I} - \boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}\boldsymbol{Q}\right)^{-1} \operatorname{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{v}}^\perp\right). \tag{216}$$

Now,

$$\begin{aligned} \left(\boldsymbol{I} - \boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}\boldsymbol{Q}\right)^{-1} &= \left(\boldsymbol{I} - \boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}\boldsymbol{Q}\right)^{\dagger} \\ &= \left(\boldsymbol{P}_{\mathcal{N}(\boldsymbol{D})} + \boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})} - \boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}\boldsymbol{Q}\right)^{\dagger} \\ &= \left(\boldsymbol{P}_{\mathcal{N}(\boldsymbol{D})} + \boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}\left(\boldsymbol{I} - \boldsymbol{Q}\right)\right)^{\dagger}. \end{aligned} \tag{217}$$

Let us apply Thm. 18 on $(\boldsymbol{P}_{\boldsymbol{D}}+\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}(\boldsymbol{I}-\boldsymbol{Q}))^\dagger$. Here, $\boldsymbol{X}_1=\boldsymbol{P}_{\mathcal{N}(\boldsymbol{D})}$ and $\boldsymbol{Y}_1=\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}(\boldsymbol{I}-\boldsymbol{Q})$. Note that $\mathcal{R}(\boldsymbol{X}_1)=\mathcal{R}^\perp(\boldsymbol{D})$ and $\mathcal{R}(\boldsymbol{Y}_1)=\mathcal{R}(\boldsymbol{D})$ and therefore $\operatorname{rank}(\boldsymbol{X}_1+\boldsymbol{Y}_1)=\operatorname{rank}(\boldsymbol{X}_1)+\operatorname{rank}(\boldsymbol{Y}_1)$. Additionally,

$$\boldsymbol{P}_{\mathcal{R}(\boldsymbol{Y}_1^{\mathrm{T}})}=\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})},\quad \boldsymbol{P}_{\mathcal{R}^\perp(\boldsymbol{X}_1^{\mathrm{T}})}=\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})},\quad \boldsymbol{P}_{\mathcal{R}^\perp(\boldsymbol{X}_1)}=\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})},\quad \boldsymbol{P}_{\mathcal{R}(\boldsymbol{Y}_1)}=\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}. \tag{218}$$

Hence,

$$\begin{aligned}\boldsymbol{L}_1&=\left(\boldsymbol{P}_{\mathcal{R}(\boldsymbol{Y}_1^{\mathrm{T}})}\boldsymbol{P}_{\mathcal{R}^\perp(\boldsymbol{X}_1^{\mathrm{T}})}\right)^\dagger=\left(\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}\right)^\dagger=\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})},\\ \boldsymbol{O}_1&=\left(\boldsymbol{P}_{\mathcal{R}^\perp(\boldsymbol{X}_1)}\boldsymbol{P}_{\mathcal{R}(\boldsymbol{Y}_1)}\right)^\dagger=\left(\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}\right)^\dagger=\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}.\end{aligned} \tag{219}$$

Therefore,

$$\begin{aligned}\left(\boldsymbol{I}-\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}\boldsymbol{Q}\right)^{-1}&=\left(\boldsymbol{P}_{\mathcal{N}(\boldsymbol{D})}+\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}\left(\boldsymbol{I}-\boldsymbol{Q}\right)\right)^\dagger\\ &=(\boldsymbol{X}_1+\boldsymbol{Y}_1)^\dagger\\ &=(\boldsymbol{I}-\boldsymbol{L}_1)\boldsymbol{X}_1^\dagger(\boldsymbol{I}-\boldsymbol{O}_1)+\boldsymbol{L}_1\boldsymbol{Y}_1^\dagger\boldsymbol{O}_1\\ &=(\boldsymbol{I}-\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})})(\boldsymbol{P}_{\mathcal{N}(\boldsymbol{D})})^\dagger(\boldsymbol{I}-\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})})\\ &\qquad+\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}\left(\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}(\boldsymbol{I}-\boldsymbol{Q})\right)^\dagger\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}\\ &=(\boldsymbol{I}-\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})})\boldsymbol{P}_{\mathcal{N}(\boldsymbol{D})}(\boldsymbol{I}-\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})})\\ &\qquad+\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}\left(\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}(\boldsymbol{I}-\boldsymbol{Q})\right)^\dagger\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}\\ &=\boldsymbol{P}_{\mathcal{N}(\boldsymbol{D})}+\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}\left(\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}(\boldsymbol{I}-\boldsymbol{Q})\right)^\dagger\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})},\end{aligned} \tag{220}$$

where in the third step we used Thm. 18. Thus we get the following intermediate result

$$\begin{aligned}\lim_{t\to\infty}\operatorname{vec}\left(\boldsymbol{\Sigma}_t^\perp\right)&=\left(\boldsymbol{I}-\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}\boldsymbol{Q}\right)^{-1}\operatorname{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{v}}^\perp\right)\\ &=\left(\boldsymbol{P}_{\mathcal{N}(\boldsymbol{D})}+\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}\left(\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}(\boldsymbol{I}-\boldsymbol{Q})\right)^\dagger\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}\right)\operatorname{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{v}}^\perp\right)\\ &=\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}\left(\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}(\boldsymbol{I}-\boldsymbol{Q})\right)^\dagger\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}\operatorname{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{v}}^\perp\right),\end{aligned} \tag{221}$$

where in the final step we used $\boldsymbol{P}_{\mathcal{N}(\boldsymbol{D})}\operatorname{vec}(\boldsymbol{\Sigma}_{\boldsymbol{v}}^\perp)=\boldsymbol{0}$. Now, note that $\mathcal{R}(\boldsymbol{P}_{\mathcal{N}(\boldsymbol{D})}\boldsymbol{C})=\mathcal{R}(\boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})}+\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})})$, whereas $\operatorname{vec}(\boldsymbol{\Sigma}_{\boldsymbol{v}}^\perp)\in\mathcal{R}(\boldsymbol{D})=\mathcal{R}(\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})})$ and therefore $(\boldsymbol{P}_{\mathcal{N}(\boldsymbol{D})}\boldsymbol{C})^\dagger\operatorname{vec}(\boldsymbol{\Sigma}_{\boldsymbol{v}}^\perp)=\boldsymbol{0}$. Hence,

$$\begin{aligned}\lim_{t\to\infty}\operatorname{vec}\left(\boldsymbol{\Sigma}_t^\perp\right)&=\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}\left(\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}(\boldsymbol{I}-\boldsymbol{Q})\right)^\dagger\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}\operatorname{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{v}}^\perp\right)\\ &=\left(\left(2\eta\boldsymbol{P}_{\mathcal{N}(\boldsymbol{D})}\boldsymbol{C}\right)^\dagger+\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}\left(\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}(\boldsymbol{I}-\boldsymbol{Q})\right)^\dagger\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{D})}\right)\operatorname{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{v}}^\perp\right).\end{aligned} \tag{222}$$

Let us apply again Thm. 18 but in the other direction. This time, $\boldsymbol{X}_2 = 2\eta \boldsymbol{P}_{\mathcal{N}(\boldsymbol{D})}\boldsymbol{C}$ and $\boldsymbol{Y}_2 = \boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{D})}(\boldsymbol{I}-\boldsymbol{Q})$. Note that $\mathcal{R}(\boldsymbol{X}_2) = \mathcal{R}(\boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}+\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})})$ and $\mathcal{R}(\boldsymbol{Y}_2) = \mathcal{R}(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})})$ and therefore $\operatorname{rank}(\boldsymbol{X}_2+\boldsymbol{Y}_2) = \operatorname{rank}(\boldsymbol{X}_2)+\operatorname{rank}(\boldsymbol{Y}_2)$. Additionally,

$$\begin{aligned}
\boldsymbol{P}_{\mathcal{R}(\boldsymbol{Y}_2^{\mathrm{T}})} &= \boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})},\\
\boldsymbol{P}_{\mathcal{R}(\boldsymbol{Y}_2)} &= \boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})},\\
\boldsymbol{P}_{\mathcal{R}^{\perp}(\boldsymbol{X}_2^{\mathrm{T}})} &= \boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})}+\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})},\\
\boldsymbol{P}_{\mathcal{R}^{\perp}(\boldsymbol{X}_2)} &= \boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})}+\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}.
\end{aligned} \tag{223}$$

Hence,

$$\begin{aligned}
\boldsymbol{L}_2 &= \left(\boldsymbol{P}_{\mathcal{R}(\boldsymbol{Y}_2^{\mathrm{T}})}\boldsymbol{P}_{\mathcal{R}^{\perp}(\boldsymbol{X}_2^{\mathrm{T}})}\right)^{\dagger}\\
&= \left(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\left(\boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})}+\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\right)\right)^{\dagger}\\
&= \boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\\
&= \boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{D})},\\
\boldsymbol{O}_2 &= \left(\boldsymbol{P}_{\mathcal{R}^{\perp}(\boldsymbol{X}_2)}\boldsymbol{P}_{\mathcal{R}(\boldsymbol{Y}_2)}\right)^{\dagger}\\
&= \left(\left(\boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})}+\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\right)\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\right)^{\dagger}\\
&= \boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})} = \boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{D})}.
\end{aligned} \tag{224}$$

Moreover, since $\mathcal{R}(\boldsymbol{X}_2) = \mathcal{R}(\boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}+\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})})$ and $\mathcal{R}^{\perp}(\boldsymbol{Y}_2) = \mathcal{R}(\boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}+\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}(\boldsymbol{H})}+\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\otimes\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})})$ we have that $\mathcal{R}(\boldsymbol{X}_2)\subseteq\mathcal{R}^{\perp}(\boldsymbol{Y}_2) = \mathcal{N}(\boldsymbol{D})$. Therefore,

$$(\boldsymbol{I}-\boldsymbol{L}_2)\boldsymbol{X}_2^{\dagger}(\boldsymbol{I}-\boldsymbol{O}_2) = \left(\boldsymbol{I}-\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{D})}\right)\boldsymbol{X}_2^{\dagger}\left(\boldsymbol{I}-\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{D})}\right) = \boldsymbol{P}_{\mathcal{N}(\boldsymbol{D})}\boldsymbol{X}_2^{\dagger}\boldsymbol{P}_{\mathcal{N}(\boldsymbol{D})} = \boldsymbol{X}_2^{\dagger}. \tag{225}$$

Therefore, applying Thm. 18 we get

$$\begin{aligned}
(2\eta\boldsymbol{P}_{\mathcal{N}(\boldsymbol{D})}\boldsymbol{C})^{\dagger} + \boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{D})}\left(\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{D})}(\boldsymbol{I}-\boldsymbol{Q})\right)^{\dagger}\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{D})}&\\
&= \boldsymbol{X}_2^{\dagger}+\boldsymbol{L}_2\boldsymbol{Y}_2^{\dagger}\boldsymbol{O}_2\\
&= (\boldsymbol{I}-\boldsymbol{L}_2)\boldsymbol{X}_2^{\dagger}(\boldsymbol{I}-\boldsymbol{O}_2)+\boldsymbol{L}_2\boldsymbol{Y}_2^{\dagger}\boldsymbol{O}_2\\
&= (\boldsymbol{X}_2+\boldsymbol{Y}_2)^{\dagger}\\
&= \left(2\eta\boldsymbol{P}_{\mathcal{N}(\boldsymbol{D})}\boldsymbol{C}+\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{D})}(\boldsymbol{I}-\boldsymbol{Q})\right)^{\dagger}\\
&= \left(2\eta\boldsymbol{P}_{\mathcal{N}(\boldsymbol{D})}\boldsymbol{C}+2\eta\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{D})}\boldsymbol{C}-\eta^2\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{D})}\boldsymbol{D}\right)^{\dagger}\\
&= \left(2\eta\boldsymbol{C}-\eta^2\boldsymbol{D}\right)^{\dagger}.
\end{aligned} \tag{226}$$

where in the second step we used (225), and in the third step we used Thm. 18. Overall, together with (70) we get

$$\begin{aligned}\lim_{t\to\infty} \operatorname{vec}\left(\boldsymbol{\Sigma}_t^{\perp}\right) &= \left(2\eta\boldsymbol{C}-\eta^2\boldsymbol{D}\right)^{\dagger}\operatorname{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{v}}^{\perp}\right)\\ &= \left(\frac{1}{\eta}\left(2\boldsymbol{C}-\eta\boldsymbol{D}\right)^{\dagger}\right)\left(\eta^2 p\operatorname{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{g}}^{\perp}\right)\right)\\ &= \eta p\left(2\boldsymbol{C}-\eta\boldsymbol{D}\right)^{\dagger}\operatorname{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{g}}^{\perp}\right). \end{aligned} \tag{227}$$

Appendix I. Proof of Corollary 13

From Thm. 12 we have that if $0<\eta<\eta_{\text{var}}^*$ then

$$\lim_{t\to\infty} \operatorname{vec}\left(\boldsymbol{\Sigma}_t^{\perp}\right) = \eta p\left(2\boldsymbol{C}-\eta\boldsymbol{D}\right)^{\dagger}\operatorname{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{g}}^{\perp}\right), \tag{228}$$

Using this result, we prove Corollary 13.

First statement. If $0<\eta<\eta_{\text{var}}^*$ then by Prop. 12

$$\begin{aligned}\lim_{t\to\infty}\mathbb{E}\left[\left\|\boldsymbol{\theta}_t^{\perp}-\boldsymbol{\theta}^{*\perp}\right\|^2\right] &= \left(\operatorname{vec}\left(\boldsymbol{I}\right)\right)^{\mathrm{T}}\lim_{t\to\infty}\operatorname{vec}\left(\boldsymbol{\Sigma}_t^{\perp}\right)\\ &= \left(\operatorname{vec}\left(\boldsymbol{I}\right)\right)^{\mathrm{T}}\left(\eta p\left(2\boldsymbol{C}-\eta\boldsymbol{D}\right)^{\dagger}\operatorname{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{g}}^{\perp}\right)\right)\\ &= \eta p(\operatorname{vec}\left(\boldsymbol{I}\right))^{\mathrm{T}}\left(2\boldsymbol{C}-\eta\boldsymbol{D}\right)^{\dagger}\operatorname{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{g}}^{\perp}\right). \end{aligned} \tag{229}$$

Second statement. Similarly, let us compute the limit of the expected value of the loss function to obtain point 2.

$$\begin{aligned}\lim_{t\to\infty}\mathbb{E}\left[\tilde{\mathcal{L}}(\boldsymbol{\theta}_t)\right]-\mathcal{L}(\boldsymbol{\theta}^*) &= \frac{1}{2}\lim_{t\to\infty}\mathbb{E}\left[(\boldsymbol{\theta}_t-\boldsymbol{\theta}^*)^{\mathrm{T}}\boldsymbol{H}(\boldsymbol{\theta}_t-\boldsymbol{\theta}^*)\right]\\ &= \frac{1}{2}\lim_{t\to\infty}\mathbb{E}\left[(\boldsymbol{\theta}_t-\boldsymbol{\theta}^*)^{\mathrm{T}}\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}\boldsymbol{H}\boldsymbol{P}_{\mathcal{N}^{\perp}(\boldsymbol{H})}(\boldsymbol{\theta}_t-\boldsymbol{\theta}^*)\right]\\ &= \frac{1}{2}\lim_{t\to\infty}\mathbb{E}\left[(\boldsymbol{\theta}_t^{\perp}-\boldsymbol{\theta}^{*\perp})^{\mathrm{T}}\boldsymbol{H}(\boldsymbol{\theta}_t^{\perp}-\boldsymbol{\theta}^{*\perp})\right]\\ &= \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{H}\lim_{t\to\infty}\mathbb{E}\left[(\boldsymbol{\theta}_t^{\perp}-\boldsymbol{\theta}^{*\perp})(\boldsymbol{\theta}_t^{\perp}-\boldsymbol{\theta}^{*\perp})^{\mathrm{T}}\right]\right)\\ &= \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{H}\lim_{t\to\infty}\boldsymbol{\Sigma}_t^{\perp}\right)\\ &= \frac{1}{2}\left(\operatorname{vec}\left(\boldsymbol{H}\right)\right)^{\mathrm{T}}\lim_{t\to\infty}\operatorname{vec}\left(\boldsymbol{\Sigma}_t^{\perp}\right)\\ &= \frac{1}{2}(\operatorname{vec}\left(\boldsymbol{H}\right))^{\mathrm{T}}\left(\eta p\left(2\boldsymbol{C}-\eta\boldsymbol{D}\right)^{\dagger}\operatorname{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{g}}^{\perp}\right)\right)\\ &= \frac{1}{2}\eta p(\operatorname{vec}\left(\boldsymbol{H}\right))^{\mathrm{T}}\left(2\boldsymbol{C}-\eta\boldsymbol{D}\right)^{\dagger}\operatorname{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{g}}^{\perp}\right). \end{aligned} \tag{230}$$

Third statement. Finally, we prove point 3. The gradient of the second-order Taylor expansion of the loss is given by

$$\nabla\tilde{\mathcal{L}}(\boldsymbol{\theta}) = \boldsymbol{H}\left(\boldsymbol{\theta}-\boldsymbol{\theta}^*\right). \tag{231}$$

Therefore

$$\begin{aligned}
\lim_{t\to\infty} \mathbb{E}\left[\left\|\nabla\tilde{\mathcal{L}}(\boldsymbol{\theta}_t)\right\|^2\right] &= \lim_{t\to\infty} \mathbb{E}\left[(\boldsymbol{\theta}_t-\boldsymbol{\theta}^*)^{\mathrm{T}}\boldsymbol{H}^2(\boldsymbol{\theta}_t-\boldsymbol{\theta}^*)\right] \\
&= \lim_{t\to\infty} \mathbb{E}\left[(\boldsymbol{\theta}_t-\boldsymbol{\theta}^*)^{\mathrm{T}}\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})}\boldsymbol{H}^2\boldsymbol{P}_{\mathcal{N}^\perp(\boldsymbol{H})}(\boldsymbol{\theta}_t-\boldsymbol{\theta}^*)\right] \\
&= \lim_{t\to\infty} \mathbb{E}\left[(\boldsymbol{\theta}_t^\perp-\boldsymbol{\theta}^{*\perp})^{\mathrm{T}}\boldsymbol{H}^2(\boldsymbol{\theta}_t^\perp-\boldsymbol{\theta}^{*\perp})\right] \\
&= \mathrm{Tr}\left(\boldsymbol{H}^2 \lim_{t\to\infty} \mathbb{E}\left[(\boldsymbol{\theta}_t^\perp-\boldsymbol{\theta}^{*\perp})(\boldsymbol{\theta}_t^\perp-\boldsymbol{\theta}^{*\perp})^{\mathrm{T}}\right]\right) \\
&= \mathrm{Tr}\left(\boldsymbol{H}^2 \lim_{t\to\infty} \boldsymbol{\Sigma}_t^\perp\right) \\
&= \left(\mathrm{vec}\left(\boldsymbol{H}^2\right)\right)^{\mathrm{T}} \lim_{t\to\infty} \mathrm{vec}\left(\boldsymbol{\Sigma}_t^\perp\right) \\
&= \left(\mathrm{vec}\left(\boldsymbol{H}^2\right)\right)^{\mathrm{T}} \left(\eta p\left(2\boldsymbol{C}-\eta\boldsymbol{D}\right)^\dagger \mathrm{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{g}}^\perp\right)\right) \\
&= \eta p\left(\mathrm{vec}\left(\boldsymbol{H}^2\right)\right)^{\mathrm{T}}\left(2\boldsymbol{C}-\eta\boldsymbol{D}\right)^\dagger \mathrm{vec}\left(\boldsymbol{\Sigma}_{\boldsymbol{g}}^\perp\right). \qquad (232)
\end{aligned}$$

Appendix J. Recovering GD's stability condition

In this section, we show how our stability condition for SGD reduces to GD's when $B = n$. In this case $p = 0$ and thus

$$\boldsymbol{C} = \frac{1}{2}\boldsymbol{H}\oplus\boldsymbol{H}, \qquad \boldsymbol{D} = \boldsymbol{H}\otimes\boldsymbol{H}. \qquad (233)$$

Let $\boldsymbol{H} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}}$ be the eigenvalue decomposition of $\boldsymbol{H}$, where $\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}} = \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \boldsymbol{I}$, then

$$\begin{aligned}
\boldsymbol{C} &= \frac{1}{2}\boldsymbol{H}\oplus\boldsymbol{H} \\
&= \frac{1}{2}\left(\boldsymbol{H}\otimes\boldsymbol{I}+\boldsymbol{H}\otimes\boldsymbol{I}\right) \\
&= \frac{1}{2}\left(\left(\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}}\right)\otimes\left(\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}\right)+\left(\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}\right)\otimes\left(\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}}\right)\right) \\
&= \frac{1}{2}\left(\left(\boldsymbol{V}\otimes\boldsymbol{V}\right)\left(\boldsymbol{\Lambda}\otimes\boldsymbol{I}\right)\left(\boldsymbol{V}^{\mathrm{T}}\otimes\boldsymbol{V}^{\mathrm{T}}\right)+\left(\boldsymbol{V}\otimes\boldsymbol{V}\right)\left(\boldsymbol{I}\otimes\boldsymbol{\Lambda}\right)\left(\boldsymbol{V}^{\mathrm{T}}\otimes\boldsymbol{V}^{\mathrm{T}}\right)\right) \\
&= \left(\boldsymbol{V}\otimes\boldsymbol{V}\right)\left(\frac{1}{2}\boldsymbol{\Lambda}\otimes\boldsymbol{I}+\frac{1}{2}\boldsymbol{I}\otimes\boldsymbol{\Lambda}\right)\left(\boldsymbol{V}\otimes\boldsymbol{V}\right)^{\mathrm{T}}. \qquad (234)
\end{aligned}$$

Note that

$$\left(\boldsymbol{V}\otimes\boldsymbol{V}\right)^{\mathrm{T}}\left(\boldsymbol{V}\otimes\boldsymbol{V}\right) = \left(\boldsymbol{V}^{\mathrm{T}}\otimes\boldsymbol{V}^{\mathrm{T}}\right)\left(\boldsymbol{V}\otimes\boldsymbol{V}\right) = \left(\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\right)\otimes\left(\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\right) = \boldsymbol{I}\otimes\boldsymbol{I} = \boldsymbol{I}, \qquad (235)$$

i.e., $(\boldsymbol{V}\otimes\boldsymbol{V})$ is an orthogonal matrix. Since $\frac{1}{2}(\boldsymbol{\Lambda}\otimes\boldsymbol{I}+\boldsymbol{I}\otimes\boldsymbol{\Lambda})$ is diagonal, then the last result in (234) is an eigenvalue decomposition of $\boldsymbol{C}$. Similarly,

$$\begin{aligned}
\boldsymbol{D} &= \boldsymbol{H}\otimes\boldsymbol{H} \\
&= \left(\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}}\right)\otimes\left(\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}}\right) \\
&= \left(\boldsymbol{V}\otimes\boldsymbol{V}\right)\left(\boldsymbol{\Lambda}\otimes\boldsymbol{\Lambda}\right)\left(\boldsymbol{V}^{\mathrm{T}}\otimes\boldsymbol{V}^{\mathrm{T}}\right) \\
&= \left(\boldsymbol{V}\otimes\boldsymbol{V}\right)\left(\boldsymbol{\Lambda}\otimes\boldsymbol{\Lambda}\right)\left(\boldsymbol{V}\otimes\boldsymbol{V}\right)^{\mathrm{T}}, \qquad (236)
\end{aligned}$$

where the last result here is the eigenvalue decomposition of $\boldsymbol{D}$. We have that $\boldsymbol{C}$ and $\boldsymbol{D}$ have the same set of eigenvectors, given by $\boldsymbol{V} \otimes \boldsymbol{V}$. This means that we can look only at the eigenvalues. Thus, set $\lambda_\ell = \boldsymbol{\Lambda}_{[\ell,\ell]} = \lambda_\ell(\boldsymbol{H})$, and define the Moore–Penrose inverse for scalars

$$\forall x \in \mathbb{R} \qquad [x]^\dagger \triangleq \begin{cases} \frac{1}{x}, & x \neq 0, \\ 0, & x = 0. \end{cases} \tag{237}$$

Then

$$\lambda_{\max}\left(\boldsymbol{C}^\dagger \boldsymbol{D}\right) = \max_{\ell,p\in[d]} \left\{ \lambda_\ell \lambda_p \left[\frac{1}{2}(\lambda_\ell + \lambda_p)\right]^\dagger \right\}. \tag{238}$$

Note that the objective vanishes whenever $\lambda_\ell = 0$ or $\lambda_p = 0$. Restricting to only positive eigenvalues gives

$$\lambda_{\max}\left(\boldsymbol{C}^\dagger \boldsymbol{D}\right) = \max_{\lambda_\ell,\lambda_p>0} \left\{ \frac{\lambda_\ell \lambda_p}{\frac{1}{2}(\lambda_\ell + \lambda_p)} \right\}. \tag{239}$$

Additionally $\sqrt{\lambda_\ell \lambda_p} \leq \frac{1}{2}(\lambda_\ell + \lambda_p)$ holds for all $\lambda_\ell, \lambda_p > 0$, therefore

$$\begin{aligned} \frac{\lambda_\ell \lambda_p}{\frac{1}{2}(\lambda_\ell + \lambda_p)} &= \sqrt{\lambda_\ell \lambda_p} \frac{\sqrt{\lambda_\ell \lambda_p}}{\frac{1}{2}(\lambda_\ell + \lambda_p)} \\ &\leq \sqrt{\lambda_\ell \lambda_p} \\ &\leq \lambda_{\max}. \end{aligned} \tag{240}$$

Yet for $\lambda_\ell = \lambda_p = \lambda_{\max}$ we have that

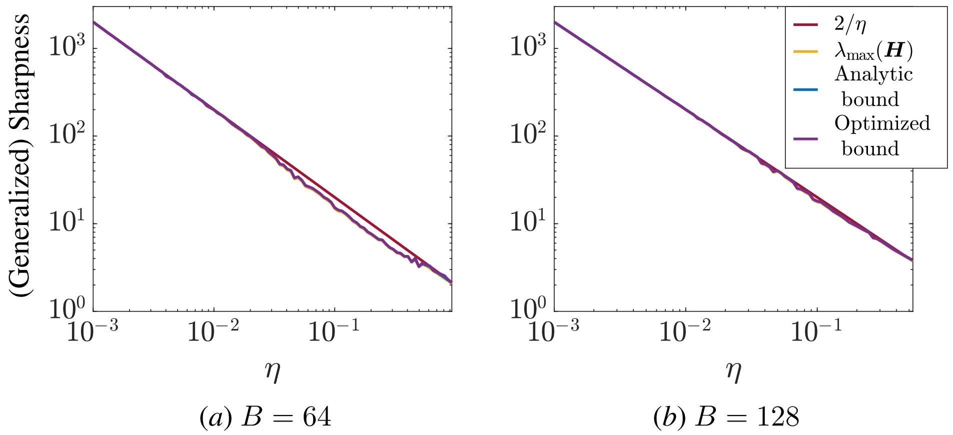
$$\frac{\lambda_\ell \lambda_p}{\frac{1}{2}(\lambda_\ell + \lambda_p)} = \lambda_{\max}. \tag{241}$$

Hence we have

$$\lambda_{\max}\left(\boldsymbol{C}^\dagger \boldsymbol{D}\right) = \max_{\lambda_\ell,\lambda_p>0} \left\{ \frac{\lambda_\ell \lambda_p}{\frac{1}{2}(\lambda_\ell + \lambda_p)} \right\} = \lambda_{\max}(\boldsymbol{H}). \tag{242}$$

Appendix K. Additional experimental results and detail

In this section, we complete the technical detail of the experiment shown in Sec. 4. For the experiment, we used a single-hidden layer ReLU network with fully connected layers (with bias vectors). The number of neurons is 1024, and the total number of parameters is $807,940$. We used four classes from MNIST, 256 samples from each class, with a total of 1024 samples. To get large initialization, we used standard torch initialization and multiplied the initial weights by a factor of 15. The maximal number of epochs was set to 4×10^4. If SGD did not converge within this number of epochs, then we removed this run from the plots.



(a) $B = 64$

(b) $B = 128$

Figure 3: **Sharpness vs. learning rate.** Additional results for the experiment in Sec. 4. These two figures complete the results of Fig. 2. Here we see that SGD with big batch sizes behaves like GD.