When Does Preconditioning Help or Hurt Generalization?

Shun-ichi Amari[*] AMARI@BRAIN.RIKEN.JP
RIKEN CBS

Jimmy Ba JBA@CS.TORONTO.EDU
Roger Grosse RGROSSE@CS.TORONTO.EDU
University of Toronto & Vector Institute

Xuechen Li LXUECHEN@CS.TORONTO.EDU
Google Research, Brain Team

Atsushi Nitanda NITANDA@MIST.I.U-TOKYO.AC.JP
Taiji Suzuki TAIJI@MIST.I.U-TOKYO.AC.JP
University of Tokyo & RIKEN AIP

Denny Wu DENNYWU@CS.TORONTO.EDU
University of Toronto & Vector Institute

Ji Xu JIXU@CS.COLUMBIA.EDU
Columbia University

Abstract

While second order optimizers such as natural gradient descent (NGD) often speed up optimization, their effect on generalization has been called into question. This work presents a more nuanced view on how the *implicit bias* of optimizers affects the comparison of generalization properties. We provide an exact bias-variance decomposition of the generalization error of overparameterized ridgeless regression under a general class of preconditioner $\boldsymbol{P}$, and consider the inverse population Fisher information matrix as a particular example. We determine the optimal $\boldsymbol{P}$ for both the bias and variance, and find that the relative generalization performance of different optimizers depends on label noise and "shape" of the signal (true parameters): when the labels are noisy, the model is misspecified, or the signal is misaligned, NGD can achieve lower risk; conversely, GD generalizes better under clean labels, a well-specified model, or aligned signal. Based on this analysis, we discuss approaches to manage the bias-variance tradeoff, and the benefit of interpolating between first- and second-order updates. We then extend our analysis to regression in the reproducing kernel Hilbert space and demonstrate that preconditioned GD can decrease the population risk faster than GD. Lastly, we empirically compare the generalization error of first- and second-order optimizers in neural network, and observe robust trends matching our theoretical analysis.

1. Introduction

We study the generalization property of an estimator $\hat{\boldsymbol{\theta}}$ obtained by minimizing the empirical risk (or the training error) $L(f_{\boldsymbol{\theta}})$ via a preconditioned gradient update:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \boldsymbol{P}(t) \nabla_{\boldsymbol{\theta}_t} L(f_{\boldsymbol{\theta}_t}), \quad t = 0, 1, \ldots \tag{1.1}$$

Setting $\boldsymbol{P} = \boldsymbol{I}$ recovers gradient descent (GD). Choices of $\boldsymbol{P}$ which exploit second-order information include the inverse Fisher information matrix, which gives the natural gradient descent (NGD) [3]; the inverse Hessian, which leads to Newton's method; and diagonal matrices estimated from past gradients, corresponding to adaptive gradient methods [26, 44]. These preconditioners often

[*] Alphabetical ordering. Correspondence to: Denny Wu (dennywu@cs.toronto.edu).

alleviate the effect of pathological curvature and speed up *optimization*, but their *generalization* properties has been under debate. While [42, 81, 83] reported that in neural network optimization, adaptive or second-order methods generalize worse than gradient descent, other empirical studies suggested that second-order methods can achieve comparable, if not better generalization [86, 87].

The generalization property of optimizers relates to the discussion of *implicit bias* [32], i.e. preconditioning may lead to a different converged solution, as shown in Figure 1. Among many propose explanations, our starting point is the well-known observation that GD implicitly regularizes the parameter ℓ_2 norm. For instance in overparameterized least squares regression, GD and many first-order methods find the minimum ℓ_2 norm solution from zero initialization (without explicit regularization), but preconditioned updates often do not. While the minimum norm solution may generalize well in the overparameterized regime [11], it is unclear whether preconditioning leads to inferior solutions – even in the simple setting of overparameterized linear regression, *quantitative* understanding of how preconditioning affects generalization is large lacking.

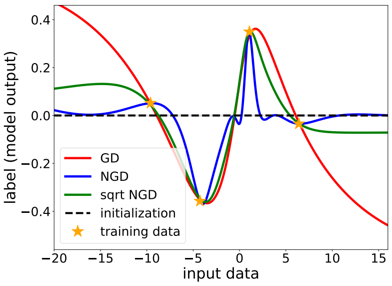


Figure 1: Function output of interpolating two-layer sigmoid network (1D) trained with preconditioned GD.

Motivated by the observation above, in Section 2 we start with overparameterized least squares regression and analyze the stationary solution of update (1.1) under time-invariant $\boldsymbol{P}$. Extending previous analysis in the proportional limit [34], we derive the generalization error in its bias-variance decomposition. We then decide the optimal $\boldsymbol{P}$ within a general class of preconditioners for both the bias and variance, and focus on the comparison between GD ($\boldsymbol{P}=\boldsymbol{I}_d$), and NGD ($\boldsymbol{P}$ is the inverse population Fisher[1]). We find that comparison of generalization is affected by the following factors:

1. **Label Noise:** Additive noise in the labels contributes to the variance term in the risk. We show that NGD achieves the optimal variance among a general class of preconditioned updates.
2. **Model Misspecification:** Under misspecification, there does not exist $f_{\boldsymbol{\theta}}$ that perfectly learns the true function (target). We argue that this factor has similar effect as label noise.
3. **Data-Signal-Alignment:** Alignment describes how the target signal distributes among input features. We show that GD achieves lower bias for isotropic signal, whereas NGD is preferred under "misalignment" — when large directions of the features match the small signal directions.

In Section 3.1 and 3.2 we discuss how the bias-variance tradeoff can be realized by choices of $\boldsymbol{P}$ (e.g. interpolating between GD and NGD) or early stopping. In Section 3.3 we extend our analysis to regression in the RKHS and show that an update that interpolates between GD and NGD reduces the population risk faster than GD. Finally, in Section 4 we empirically show that our findings in linear model carry over to neural networks: under a student-teacher setup, we compare the generalization of GD with preconditioned updates and confirm the influence of all aforementioned factors.

2. Asymptotic Risk of Ridgeless Interpolants

We consider the following setup: given training points $\{\boldsymbol{x}_i\}_{i=1}^n$ labeled by a teacher model (target function) f^* with additive noise: $y_i = f^*(\boldsymbol{x}_i) + \varepsilon_i$, we learn a linear student $f_{\boldsymbol{\theta}}$ by minimizing $L(\boldsymbol{X}, f_{\boldsymbol{\theta}}) = \sum_{i=1}^n \big(y_i - \boldsymbol{x}_i^\top \boldsymbol{\theta}\big)^2$. We assume a random design: $\boldsymbol{x}_i = \boldsymbol{\Sigma}_{\boldsymbol{X}}^{1/2} \boldsymbol{z}_i$, where $\boldsymbol{z}_i \in \mathbb{R}^d$ is

[1] From now on we use NGD to denote the *population* Fisher-based update, and we write "sample NGD" when $\boldsymbol{P}$ is the inverse or pseudo-inverse of the sample Fisher; see Appendix B for discussion.

an i.i.d. vector with zero-mean, unit-variance, and finite 12th moment; ε is i.i.d. noise with variance σ^2. We aim to compute the population risk $R(f) = \mathbb{E}_{P_X}[(f^*(\boldsymbol{x}) - f(\boldsymbol{x}))^2]$ in the following limit:

- **(A1) Overparameterized Proportional Limit:** $n, d \to \infty$, $d/n \to \gamma \in (1, \infty)$.

(A1) entails that the number of features (or parameters) is larger than the number of samples, and there exist multiple empirical risk minimizers with potentially different generalization properties.

Denote $\boldsymbol{X} = [\boldsymbol{x}_1^\top, ..., \boldsymbol{x}_n^\top]^\top \in \mathbb{R}^{n\times d}$ the data matrix and $\boldsymbol{y} \in \mathbb{R}^n$ the corresponding label vector. We optimize the parameters $\boldsymbol{\theta}$ via a preconditioned gradient flow with preconditioner $\boldsymbol{P}(t) \in \mathbb{R}^{d\times d}$,

$$\frac{\partial \boldsymbol{\theta}(t)}{\partial t} = -\boldsymbol{P}(t)\frac{\partial L(\boldsymbol{\theta}(t))}{\partial \boldsymbol{\theta}(t)} = \frac{1}{n}\boldsymbol{P}(t)\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}(t)), \quad \boldsymbol{\theta}(0) = 0. \tag{2.1}$$

In this linear setup, many common choices of preconditioner do not change through time: under Gaussian likelihood, the sample Fisher (and also Hessian) corresponds to the sample covariance $\boldsymbol{X}^\top\boldsymbol{X}/n$ up to variance scaling, whereas the population Fisher corresponds to the population covariance $\boldsymbol{F} = \boldsymbol{\Sigma}_{\boldsymbol{X}}$. We thus limit our analysis to fixed preconditioner of the form $\boldsymbol{P}(t) =: \boldsymbol{P}$.

Write parameters at time t under update (2.1) with fixed $\boldsymbol{P}$ as $\boldsymbol{\theta}_{\boldsymbol{P}}(t)$. For positive definite $\boldsymbol{P}$, the stationary solution is given as: $\hat{\boldsymbol{\theta}}_{\boldsymbol{P}} := \lim_{t\to\infty} \boldsymbol{\theta}_{\boldsymbol{P}}(t) = \boldsymbol{P}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{P}\boldsymbol{X}^\top)^{-1}\boldsymbol{y}$. One may check that the discrete time update (with appropriate step size) and other variants that do not alter the span of gradient (e.g. stochastic gradient or momentum) converge to the same solution as well.

Remark 1 *For positive definite $\boldsymbol{P}$, $\hat{\boldsymbol{\theta}}_{\boldsymbol{P}}$ is the minimum $\|\boldsymbol{\theta}\|_{\boldsymbol{P}^{-1}}$ norm interpolant. For GD this translates to the parameter ℓ_2 norm, whereas for NGD the implicit bias is the $\|\boldsymbol{\theta}\|_{\boldsymbol{F}}$ norm. Since $\mathbb{E}_{P_X}[f(\boldsymbol{x})^2] = \|\boldsymbol{\theta}\|^2_{\boldsymbol{\Sigma}_{\boldsymbol{X}}}$, NGD finds an interpolating function with smallest norm under the data distribution. We also empirically observe this divide in neural networks (see Figure 1 and Appendix A).*

We highlight the following choices of $\boldsymbol{P}$ and the corresponding stationary solution $\hat{\boldsymbol{\theta}}_{\boldsymbol{P}}$.

- **Identity:** $\boldsymbol{P} = \boldsymbol{I}_d$ recovers GD that converges to the min ℓ_2 norm interpolant (also true for momentum GD and SGD), which we write as $\hat{\boldsymbol{\theta}}_{\boldsymbol{I}} := \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y}$ and refer to as the *GD solution*.
- **Population Fisher:** $\boldsymbol{P} = \boldsymbol{F}^{-1} = \boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}$ leads to the estimator $\hat{\boldsymbol{\theta}}_{\boldsymbol{F}^{-1}}$, which we refer to as the *NGD solution*.
- **Sample Fisher:** since the sample Fisher is rank-deficient, we may add a damping term $\boldsymbol{P} = (\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I}_d)^{-1}$ or take the pseudo-inverse $\boldsymbol{P} = (\boldsymbol{X}^\top\boldsymbol{X})^\dagger$. In both cases, the gradient is still spanned by $\boldsymbol{X}$, and thus the update finds the same min ℓ_2-norm solution $\hat{\boldsymbol{\theta}}_{\boldsymbol{I}}$, although the trajectory differs (see Figure 2).

Remark 2 *The above choices reveal a gap between sample- and population-based preconditioners: while the sample Fisher accelerates optimization [89], we demonstrate certain generalization properties only possessed by the population Fisher.*

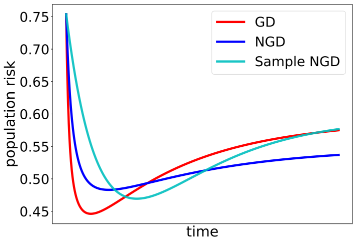
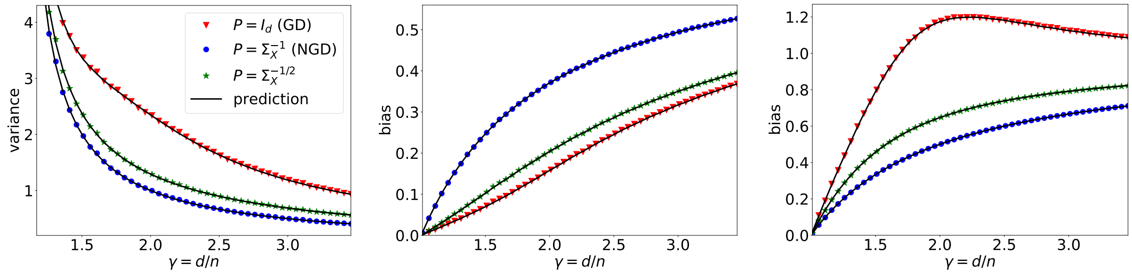


Figure 2: Population risk of preconditioned linear regression vs. time with the following $\boldsymbol{P}$: $\boldsymbol{I}$ (red), $\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}$ (blue) and $(\boldsymbol{X}^\top\boldsymbol{X})^\dagger$ (cyan). Time is rescaled differently for each curve (convergence speed is not comparable). Observe that GD and sample NGD give the same stationary risk.

We compare the population risk of the GD solution $\hat{\boldsymbol{\theta}}_{\boldsymbol{I}}$ and the NGD solution $\hat{\boldsymbol{\theta}}_{\boldsymbol{F}^{-1}}$ in its bias-variance decomposition w.r.t. the label noise, and discuss the two components separately:

$$R(\boldsymbol{\theta}) = \underbrace{\mathbb{E}_{P_X}[(f^*(\boldsymbol{x}) - \boldsymbol{x}^\top\mathbb{E}_{P_\varepsilon}[\boldsymbol{\theta}])^2]}_{B(\boldsymbol{\theta}),\ \text{bias}} + \underbrace{\mathrm{tr}(\mathrm{Cov}(\boldsymbol{\theta})\boldsymbol{\Sigma}_{\boldsymbol{X}})}_{V(\boldsymbol{\theta}),\ \text{variance}}.$$



(a) variance. (b) well-specified bias (isotropic). (c) well-specified bias (misaligned).

Figure 3: We set eigenvalues of $\mathbf{\Sigma}_{\boldsymbol{X}}$ as two point masses with $\kappa_X = 20$ and $\|\mathbf{\Sigma}_{\boldsymbol{X}}\|_F^2 = d$; empirical values (dots) are computed with $n = 300$. (a) NGD (blue) achieves minimum variance. (b) GD (red) achieves lower bias under isotropic signal: $\mathbf{\Sigma}_{\boldsymbol{\theta}} = \boldsymbol{I}_d$. (c) NGD achieves lower bias under "misalignment": $\mathbf{\Sigma}_{\boldsymbol{X}} = \mathbf{\Sigma}_{\boldsymbol{\theta}}^{-1}$.

2.1. The Variance Term: NGD is Optimal

We first characterize the stationary variance which only depends on the label noise but not the teacher model f^*. We restrict ourselves to preconditioners satisfying the following condition:

- **(A2) Converging Eigenvalues:** $\boldsymbol{P}$ is positive definite and as $n, d \to \infty$, the spectral distribution of $\mathbf{\Sigma}_{\boldsymbol{XP}} := \boldsymbol{P}^{1/2}\mathbf{\Sigma}_{\boldsymbol{X}}\boldsymbol{P}^{1/2}$ converges weakly to $\boldsymbol{H}_{\boldsymbol{XP}}$ supported on $[c, C]$ for $c, C > 0$.

The following theorem characterizes the asymptotic variance and the corresponding optimal $\boldsymbol{P}$.

Theorem 1 *Given (A1-2),* $V(\hat{\boldsymbol{\theta}}_{\boldsymbol{P}}) \to \sigma^2\left(\lim_{\lambda\to 0_+} m'(-\lambda)m^{-2}(-\lambda) - 1\right)$*, where* $m(z) > 0$ *is the Stieltjes transform of the limiting distribution of eigenvalues of* $\frac{1}{n}\boldsymbol{XPX}^\top$ *(for* z *beyond its support) defined as the solution to* $m^{-1}(z) = -z + \gamma\int \tau(1 + \tau m(z))^{-1}\mathrm{d}\boldsymbol{H}_{\boldsymbol{XP}}(\tau)$.

Furthermore, under (A1-2), $V(\hat{\boldsymbol{\theta}}_{\boldsymbol{P}}) \geq \sigma^2(\gamma - 1)^{-1}$*, and the equality is achieved by* $\boldsymbol{P} = \boldsymbol{F}^{-1}$.

The above risk formula is a direct extension of Hastie et al. [34, Thorem 4] and can also be obtained from earlier random matrix theoretical results [24, 46]. Formula (**??**) is a direct extension of Hastie et al. [34, Thorem 4], which can be obtained from Dobriban et al. [24, Thorem 2.1] or Ledoit and Péché [46, Thorem 1.2]. We note that the eigenvalue condition in (A2) may also be relaxed as in [85]. Theorem 1 implies that preconditioning with the inverse population Fisher $\boldsymbol{F}$ results in the optimal stationary variance, which is supported by Figure 3(a). In other words, when the labels are noisy so that the risk is dominated by the variance term, we expect NGD to generalize better upon convergence. We emphasize that this advantage is only present when the population Fisher is used, but not its sample-based counterpart (which converges to $\hat{\boldsymbol{\theta}}_{\boldsymbol{I}}$ as commented above).

2.2. The Bias Term: Well-specified Case

We now analyze the bias term under linear teacher: $f^*(\boldsymbol{x}) = \boldsymbol{x}^\top\boldsymbol{\theta}^*$. Extending the setting in Dobriban et al. [24], we assume a general prior $\mathbb{E}[\boldsymbol{\theta}^*\boldsymbol{\theta}^{*\top}] = d^{-1}\mathbf{\Sigma}_{\boldsymbol{\theta}}$ and the following joint relations:

- **(A3) Joint Convergence:** $\mathbf{\Sigma}_{\boldsymbol{X}}$ and $\boldsymbol{P}$ share the same eigenvector matrix $\boldsymbol{U}$. The empirical distributions of elements of $(\boldsymbol{e}_x, \boldsymbol{e}_\theta, \boldsymbol{e}_{xp})$ jointly converge to random variables $(\upsilon_x, \upsilon_\theta, \upsilon_{xp})$ supported on $[c', C']$ for $c', C' > 0$, where $\boldsymbol{e}_x$, $\boldsymbol{e}_{xp}$ are eigenvalues of $\mathbf{\Sigma}_{\boldsymbol{X}}$, $\mathbf{\Sigma}_{\boldsymbol{XP}}$, and $\boldsymbol{e}_\theta = \mathrm{diag}\left(\boldsymbol{U}^\top\mathbf{\Sigma}_\theta\boldsymbol{U}\right)$.

When $\boldsymbol{P} = \boldsymbol{I}_d$, Hastie et al. [34], Xu and Hsu [85] considered the special case of isotropic prior $\mathbf{\Sigma}_{\boldsymbol{\theta}} = \boldsymbol{I}_d$. We remark that our generalized prior $\mathbf{\Sigma}_{\boldsymbol{\theta}}$ gives rise to interesting phenomena that are not captured by simplified settings, such as non-monotonic bias and variance (see Figure 12), and epoch-wise double descent (see Figure 10). Under the general setup, we have the following result:

Theorem 2 *Under (A1)(A3), the expected bias $B(\hat{\boldsymbol{\theta}}_{\boldsymbol{P}}) := \mathbb{E}_{\boldsymbol{\theta}^*}[B(\hat{\boldsymbol{\theta}}_{\boldsymbol{P}})]$ is given as*

$$B(\hat{\boldsymbol{\theta}}_{\boldsymbol{P}}) \to \lim_{\lambda\to 0_+} m'(-\lambda)m^{-2}(-\lambda)\mathbb{E}\left[\upsilon_x\upsilon_\theta(1+\upsilon_{xp}m(-\lambda))^{-2}\right],$$

where expectation is taken over υ and $m(z)$ is the Stieltjes transform defined in Theorem 1.

Furthermore, among all $\boldsymbol{P}$ satisfying (A3), the optimal bias is achieved by $\boldsymbol{P} = \boldsymbol{U}\,\mathrm{diag}\,(\boldsymbol{e}_\theta)\boldsymbol{U}^\top$.

Note that the optimal $\boldsymbol{P}$ depends on orientation of the teacher $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$, which is usually not known. This result can thus be interpreted as a *no-free-lunch* characterization in choosing an optimal preconditioner: when parameters of the teacher model have roughly equal magnitude (isotropic), GD achieves lower bias (see Figure 3(b) where $\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \boldsymbol{I}_d$). On the other hand, when $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ is "misaligned" with $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$, i.e. when the most varying directions of the input features contain little information about $\boldsymbol{\theta}^*$ (see Figure 3(c) where $\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}$), NGD results in lower bias. In the analogy of source condition, the extent of "misalignment" translates to the hardness of learning; the above discussion thus suggests that GD is beneficial in "easy" tasks, whereas NGD is preferable under "difficult" teacher.

2.3. Misspecification ≈ Label Noise

Under misspecification, there does not exist a linear predictor achieving zero bias. In this case, we may decompose the teacher into a linear component and its residual: $f^*(\boldsymbol{x}) = \boldsymbol{x}^\top\boldsymbol{\theta}^* + f_c^*(\boldsymbol{x})$.

For simplicity, we first consider f_c^* to be a linear function on unobserved features (similar to [34]): $y_i = \boldsymbol{x}_i^\top\boldsymbol{\theta}^* + \boldsymbol{x}_{c,i}^\top\boldsymbol{\theta}^c + \varepsilon_i$, where $\boldsymbol{x}_{c,i} \in \mathbb{R}^{d_c}$ is independent to $\boldsymbol{x}_i$ with covariance $\boldsymbol{\Sigma}_{\boldsymbol{X}}^c$, and $\mathbb{E}[\boldsymbol{\theta}^c\boldsymbol{\theta}^{c\top}] = d_c^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^c$. In this setting, the misspecification bias is analogous to the variance:

Proposition 3 *For the above unobserved features model, given (A1-3), the bias can be written as $B(\hat{\boldsymbol{\theta}}) = B_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{P}}) + B_c(\hat{\boldsymbol{\theta}}_{\boldsymbol{P}})$, where $B_{\boldsymbol{\theta}}$ is the well-specified bias in Thm. 2, and $B_c = d_c^{-1}\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{X}}^c\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^c)(V(\hat{\boldsymbol{\theta}}_{\boldsymbol{P}})+1)$, where $V(\hat{\boldsymbol{\theta}}_{\boldsymbol{P}})$ is the variance in Thm. 1.*

Misspecification can thus be interpreted as additional label noise (also noted in Hastie et al. [34, Thorem 4]), for which NGD is beneficial due to Theorem 1. While Proposition 3 only describes one example of misspecified model, we expect such finding to hold under broader settings. In particular, Mei and Montanari [60, Remark 5] suggests that for many f_c^*, the misspecified bias is the same as variance due to label noise. This result is only rigorously shown for isotropic data, but we empirically verify similar phenomenon under general covariances in Figure 4, in which f_c^* is a quadratic function: $f_c^*(\boldsymbol{x}) = \alpha(\boldsymbol{x}^\top\boldsymbol{x} - \mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{X}}))$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \boldsymbol{I}_d$, where α controls the extent of nonlinearity. Observe that NGD achieves lower bias as we further misspecify the model.

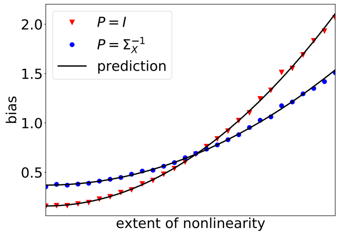


Figure 4: Misspecified bias.

3. Bias-variance Tradeoff

Our characterization of the stationary risk suggests that preconditioners that achieve the optimal bias and variance are in general different. This section discusses how the bias-variance tradeoff can be realized by interpolating between preconditioners or by early stopping. In addition, we analyze regression in the RKHS and show that by balancing the bias and variance, a preconditioned update that interpolates between GD and NGD also decreases the population risk faster than GD.

3.1. Interpolating between Preconditioners

Intuitively, given $\boldsymbol{P}_1$ that minimizes the bias and $\boldsymbol{P}_2$ that minimizes the variance, we may expect a preconditioner that interpolates between $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ to balance the bias and variance and thus generalize better. We confirm this intuition in a setup of general $\boldsymbol{\Sigma_X}$ and isotropic $\boldsymbol{\Sigma_\theta}$, for which GD achieves optimal stationary bias and NGD achieves optimal variance.

Proposition 4 (Informal) *Let $\boldsymbol{\Sigma_X} \neq \boldsymbol{I}_d$ and $\boldsymbol{\Sigma_\theta} = \boldsymbol{I}_d$. Consider the following preconditioners: (i) $\boldsymbol{P}_\alpha = \alpha\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1} + (1-\alpha)\boldsymbol{I}_d$, (ii) $\boldsymbol{P}_\alpha = (\alpha\boldsymbol{\Sigma_X} + (1-\alpha)\boldsymbol{I}_d)^{-1}$, (iii) $\boldsymbol{P}_\alpha = \boldsymbol{\Sigma}_{\boldsymbol{X}}^{-\alpha}$. The stationary variance monotonically decreases with $\alpha \in [0,1]$ for all three choices. For (i), the bias monotonically increases with $\alpha \in [0,1]$, and for (ii)(iii), the monotonicity holds for α in certain range.*

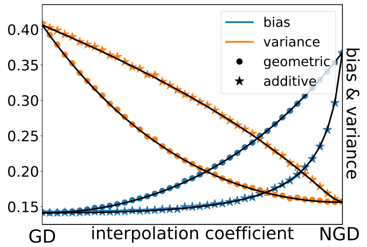


Figure 5: Bias-variance tradeoff (SNR=32/5). As we additively or geometrically interpolate from GD to NGD (left to right), the stationary bias (blue) increases and the variance (orange) decreases.

In other words, as the signal-to-noise ratio (SNR) decreases (i.e. more label noise), one can increase α, which makes the update closer to NGD, to improve generalization, and vice versa[2] (small α entails GD-like update). This intuition is supported by Figure 5 and 13(c): for certain SNR, interpolating between $\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}$ and $\boldsymbol{\Sigma_\theta}$ can lead to lower stationary risk than both GD and NGD.

Remark 3 *Two aforementioned interpolation schemes summarize common choices in practice: additive interpolation (ii) corresponds to the damping to stably invert the Fisher, whereas geometric interpolation (iii) includes the square-root scaling in adaptive gradient methods [26, 44].*

3.2. The Role of Early Stopping

We previously considered the stationary solution of the unregularized objective. It is known that the bias-variance tradeoff can also be controlled by either explicit or algorithmic regularization. We briefly comment on the effect of early stopping, starting from the monotonicity of the variance term.

Proposition 5 *For all $\boldsymbol{P}$ satisfying (A2), the variance $V(\boldsymbol{\theta_P}(t))$ monotonically increases with t.*

The proposition confirms that early stopping reduces overfitting. Variance reduction can benefit GD in its comparison with NGD, which achieves the lowest stationary variance: Figure 2 and 16 show that GD can be advantageous under early stopping even if NGD has lower stationary risk.

On the other hand, early stopping may not always improve the well-specified bias. While a complete analysis is difficult due to the non-monotonicity of the bias term (see Appendix A), we speculate that previous observations on the stationary bias also translate to early stopping. As a concrete example, we consider well-specified settings in which either GD or NGD achieves the optimal stationary bias, and demonstrate that such optimality is also preserved under early stopping:

Proposition 6 *Given (A3) and denote the optimal early stopping bias as $B^{\mathrm{opt}}(\boldsymbol{\theta}) = \inf_{t\geq 0} B(\boldsymbol{\theta}(t))$. When $\boldsymbol{\Sigma_\theta} = \boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}$, $B^{\mathrm{opt}}(\boldsymbol{\theta_P}) \geq B^{\mathrm{opt}}(\boldsymbol{\theta}_{\boldsymbol{F}^{-1}})$. Whereas when $\boldsymbol{\Sigma_\theta} = \boldsymbol{I}_d$, $B^{\mathrm{opt}}(\boldsymbol{\theta}_{\boldsymbol{F}^{-1}}) \geq B^{\mathrm{opt}}(\boldsymbol{\theta_I})$.*

Figure 16 illustrates that the observed trend in stationary bias (well-specified) is indeed preserved under optimal early stopping: GD or NGD achieves lower early stopping bias under isotropic or misaligned teacher model, respectively. We leave the precise characterization as future work.

[2] In Appendix D.5 we empirically verify the monotonicity bias over all $\alpha \in [0,1]$ beyond the proposition.

3.3. Fast Decay of Population Risk

Our previous analysis suggests that certain preconditioners can achieve lower population risk, but does not address which method decreases the risk more efficiently. Knowing that preconditioning may accelerate optimization, one natural question to ask is, is this speedup also present for generalization under fixed dataset? We provide an affirmative answer in a slightly different model: we study least squares regression in the RKHS, and show that a preconditioned update that interpolates between GD and NGD achieves the minimax optimal rate in much fewer steps than GD.

We provide a brief outline and defer the detailed setup to Appendix D. Let $\mathcal{H}$ be an RKHS included in $L_2(P_X)$ equipped with a bounded kernel function k, and $K_{\boldsymbol{x}} \in \mathcal{H}$ be the Riesz representation of the kernel function. Define S as the canonical operator from $\mathcal{H}$ to $L_2(P_X)$, and write $\Sigma = S^*S$ and $L = SS^*$. Given a teacher model f^*, we assume the following:

- **(A4) Source Condition:** $\exists r \in (0, \infty)$, $M > 0$ s.t. $f^* = L^r h^*$ for $h^* \in L_2(P_X)$ and $\|f^*\|_\infty \leq M$.
- **(A5) Capacity Condition:** $\exists s > 1$ s.t. $\operatorname{tr}\left(\Sigma^{1/s}\right) < \infty$ and $2r + s^{-1} > 1$.
- **(A6) Regularity of RKHS:** $\exists \mu \in [s^{-1}, 1]$, $C_\mu > 0$ s.t. $\sup_{\boldsymbol{x} \in \operatorname{supp}(P_X)} \left\|\Sigma^{1/2-1/\mu} K_{\boldsymbol{x}}\right\|_{\mathcal{H}} \leq C_\mu$.

Note that in the source condition (A4), the coefficient r controls the complexity of the teacher model and relates to the notions of model misalignment in Section 2: large r indicates a smoother teacher model which is "easier" to learn, and vice versa [74]. Given training points $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, we consider the following preconditioned update on the student model $f_t \in \mathcal{H}$:

$$f_t = f_{t-1} - \eta(\Sigma + \alpha I)^{-1}(\hat{\Sigma} f_{t-1} - \hat{S}^* Y), \quad f_0 = 0, \tag{3.1}$$

where $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n K_{\boldsymbol{x}_i} \otimes K_{\boldsymbol{x}_i}$, $\hat{S}^*Y = \frac{1}{n}\sum_{i=1}^n y_i K_{\boldsymbol{x}_i}$. In this setting, the population Fisher corresponds to covariance operator Σ, and thus (3.1) can be interpreted as *additive* interpolation between GD and NGD: update with large α behaves like GD, and small α like NGD. The following theorem shows that with appropriate α, preconditioning leads to faster decay of population risk.

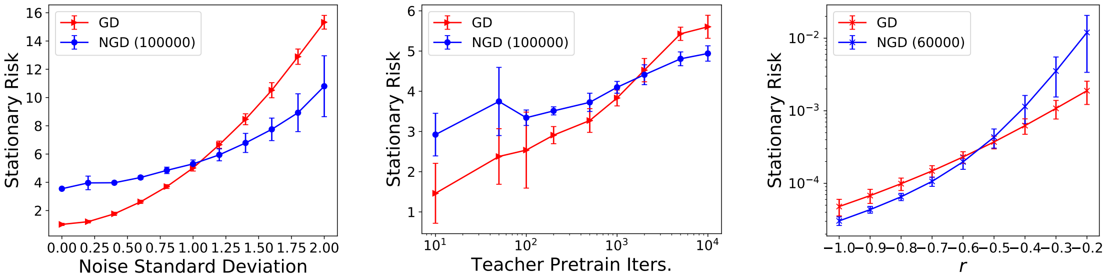
Theorem 7 (Informal) *Define the population risk $R(f_t) = \|Sf_t - f^*\|_{L_2(P_X)}^2$. Given (A4-6), if $r \geq 1/2$ or $\mu \leq 2r$, then preconditioned update (3.1) with $\alpha = n^{-\frac{2s}{2rs+1}}$ achieves minimax optimal rate $R(f_t) = \tilde{O}\left(n^{-\frac{2rs}{2rs+1}}\right)$ in $t = \Theta(\log n)$ steps, whereas GD requires $t = \Theta\left(n^{\frac{2rs}{2rs+1}}\right)$ steps.*

We remark that the optimal interpolation coefficient α and stopping time t are chosen to balance the bias $B(t)$ and variance $V(t)$. Note that α depends on the teacher model in the following way: for $n > 1$, α decreases as r becomes smaller, which corresponds to non-smooth and "difficult" f^*, and vice versa. This agrees with our previous observation that NGD is advantageous when the teacher model is difficult to learn. We defer empirical verification of this result to Appendix C.

4. Neural Network Experiments

Protocol. We compare the generalization performance of GD and NGD in neural network settings and illustrate the influence of the following factors: (i) label noise; (ii) misspecification; (iii) signal misalignment. In Appendix A we also verify the advantage of interpolating between GD and NGD.

To create a student-teacher setup, we split the training set into two halves, one of which (*pretrain* split) along with the original labels is used to pretrain the teacher, and the other (*distill* split) along with the teacher's labels is used to distill [35] the student. We normalize the teacher's labels



(a) label noise (MNIST). (b) misspecification (CIFAR-10). (c) misalignment (MNIST).

Figure 6: Neural net experiments: in parentheses are amount of unlabeled data for estimating the Fisher.

(logits) [9] before potentially adding label noise, and fit the student by minimizing the L2 loss. We implement NGD using Hessian-free optimization [56], and use 100k unlabeled data to estimate the population Fisher. We report the test error when the training error is below 0.2% of its initial value as a proxy for the stationary risk. We defer detailed setup to Appendix E.

4.1. Empirical Findings

Label Noise. We pretrain the teacher on the pretrain split and use 1024 examples from the distill split to fit the student. Both the student and teacher are two-layer ReLU nets with 80 hidden units. We corrupt the labels with isotropic Gaussian noise. Figure 6(a) shows that as the noise level increases, the stationary risk of GD worsening faster, which aligns with our observation in Figure 3.

Misspecification. We use a ResNet-20 teacher and the same 80-neuron two-layer ReLU student. We control the misspecification by varying amount of pretraining of the teacher. Intuitively, large teacher models that are trained longer should be more complex and thus likely to be outside of functions that the two-layer student can represent (i.e. more misspecified). Indeed, Figure 6(b) shows that NGD eventually achieves better generalization as the teacher is trained for more steps. In Appendix A we discuss a heuristic measure of misspecification that supports our construction.

Misalignment. We set the student and teacher to be the same two-layer ReLU network. We construct the teacher by perturbing the student's initialization, the direction of which is given by $\boldsymbol{F}^r$, where $\boldsymbol{F}$ is the student's Fisher and $r \in [-1, 0]$. As r approaches -1, the important parameters of the teacher (i.e. larger update directions) becomes misaligned with the student's Hessian, and thus learning is more "difficult". While this analogy is rather superficial, Figure 6(c) shows that as r becomes smaller (more misaligned), NGD begins to generalize better in terms of stationary risk.

5. Discussion and Conclusion

We analyzed the generalization of preconditioned gradient descent in overparameterized least squares regression (with emphasis on NGD). We identified three factors that affect the relative generalization performance, and determined the corresponding optimal $\boldsymbol{P}$. We also provided justification for common algorithmic choices by discussing the bias-variance tradeoff. Note that our current setup is limited to time-invariant preconditioners, which does not cover many adaptive gradient methods; understanding these optimizers in similar setting would be an interesting future direction. Another important problem is to further characterize the interplay between preconditioning and explicit (e.g. weight decay) or algorithmic regularization (e.g. large step size).

References

[1] Alnur Ali, J Zico Kolter, and Ryan J Tibshirani. A continuous-time view of early stopping for least squares. In *International Conference on Artificial Intelligence and Statistics*, volume 22, 2019.

[2] Shun-ichi Amari. Neural learning in structured parameter spaces-natural riemannian gradient. In *Advances in neural information processing systems*, pages 127–133, 1997.

[3] Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2): 251–276, 1998.

[4] Shun-Ichi Amari, Hyeyoung Park, and Kenji Fukumizu. Adaptive method of realizing natural gradient learning for multilayer perceptrons. *Neural computation*, 12(6):1399–1409, 2000.

[5] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, pages 7411–7422, 2019.

[6] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *arXiv preprint arXiv:1901.08584*, 2019.

[7] Navid Azizan and Babak Hassibi. Stochastic gradient/mirror descent: Minimax optimality and implicit regularization. *arXiv preprint arXiv:1806.00952*, 2018.

[8] Navid Azizan, Sahin Lale, and Babak Hassibi. Stochastic mirror descent on overparameterized nonlinear models: Convergence, implicit regularization, and generalization. *arXiv preprint arXiv:1906.03830*, 2019.

[9] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In *Advances in neural information processing systems*, pages 2654–2662, 2014.

[10] Jimmy Ba, Murat A. Erdogdu, Taiji Suzuki, Denny Wu, and Tianzong Zhang. Generalization of two-layer neural networks: An asymptotic viewpoint. *International Conference on Learning Representations*, 2020.

[11] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *arXiv preprint arXiv:1906.11300*, 2019.

[12] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine learning and the bias-variance trade-off. *arXiv preprint arXiv:1812.11118*, 2018.

[13] Mikhail Belkin, Daniel J Hsu, and Partha Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. In *Advances in neural information processing systems*, pages 2300–2311, 2018.

[14] Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? *arXiv preprint arXiv:1806.09471*, 2018.

[15] Alberto Bietti and Julien Mairal. On the inductive bias of neural tangent kernels. In *Advances in Neural Information Processing Systems*, pages 12873–12884, 2019.

[16] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[17] Tianle Cai, Ruiqi Gao, Jikai Hou, Siyu Chen, Dong Wang, Di He, Zhihua Zhang, and Liwei Wang. A gram-gauss-newton method learning overparameterized deep neural networks for regression problems. *arXiv preprint arXiv:1905.11675*, 2019.

[18] Yuan Cao, Zhiying Fang, Yue Wu, Ding-Xuan Zhou, and Quanquan Gu. Towards understanding the spectral bias of deep learning. *arXiv preprint arXiv:1912.01198*, 2019.

[19] Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

[20] Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. *arXiv preprint arXiv:2002.04486*, 2020.

[21] Stéphane d'Ascoli, Maria Refinetti, Giulio Biroli, and Florent Krzakala. Double trouble in double descent: Bias and variance (s) in the lazy regime. *arXiv preprint arXiv:2003.01054*, 2020.

[22] Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis. A model of double descent for high-dimensional binary linear classification. *arXiv preprint arXiv:1911.05822*, 2019.

[23] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. *arXiv preprint arXiv:1703.04933*, 2017.

[24] Edgar Dobriban, Stefan Wager, et al. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

[25] Bin Dong, Jikai Hou, Yiping Lu, and Zhihua Zhang. Distillation $\approx$ early stopping? harvesting dark knowledge utilizing anisotropic information retrieval for overparameterized neural network. *arXiv preprint arXiv:1910.01255*, 2019.

[26] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(Jul):2121–2159, 2011.

[27] Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Generalisation error in learning with random features and the hidden manifold model. *arXiv preprint arXiv:2002.09339*, 2020.

[28] Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in deep linear neural networks. *arXiv preprint arXiv:1904.13262*, 2019.

[29] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.

[30] Roger Grosse and James Martens. A kronecker-factored approximate fisher matrix for convolution layers. In *International Conference on Machine Learning*, pages 573–582, 2016.

[31] Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, pages 6151–6159, 2017.

[32] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. *arXiv preprint arXiv:1802.08246*, 2018.

[33] Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, pages 9461–9471, 2018.

[34] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

[35] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[36] Takashi Ishida, Ikko Yamane, Tomoya Sakai, Gang Niu, and Masashi Sugiyama. Do we need zero training loss after achieving zero training error? *arXiv preprint arXiv:2002.08709*, 2020.

[37] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

[38] Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. *arXiv preprint arXiv:1810.02032*, 2018.

[39] Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory*, pages 1772–1798, 2019.

[40] Ryo Karakida, Shotaro Akaho, and Shun-ichi Amari. Universal statistics of fisher information in deep neural networks: Mean field approach. *arXiv preprint arXiv:1806.01316*, 2018.

[41] Noureddine El Karoui. Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv preprint arXiv:1311.2445*, 2013.

[42] Nitish Shirish Keskar and Richard Socher. Improving generalization performance by switching from adam to sgd. *arXiv preprint arXiv:1712.07628*, 2017.

[43] 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.

[44] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[45] Frederik Kunstner, Lukas Balles, and Philipp Hennig. Limitations of the empirical fisher approximation. *arXiv preprint arXiv:1905.12558*, 2019.

[46] Olivier Ledoit and Sandrine Péché. Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 151(1-2):233–264, 2011.

[47] Daniel Levy and John C Duchi. Necessary and sufficient geometries for gradient methods. In *Advances in Neural Information Processing Systems*, pages 11491–11501, 2019.

[48] Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. *arXiv preprint arXiv:1903.11680*, 2019.

[49] Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. *arXiv preprint arXiv:1712.09203*, 2017.

[50] Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel" ridgeless" regression can generalize. *arXiv preprint arXiv:1808.00387*, 2018.

[51] Zhenyu Liao and Romain Couillet. The dynamics of learning: a random matrix approach. *arXiv preprint arXiv:1805.11917*, 2018.

[52] Junhong Lin and Lorenzo Rosasco. Optimal rates for multi-pass stochastic gradient methods. *The Journal of Machine Learning Research*, 18(1):3375–3421, 2017.

[53] Cosme Louart, Zhenyu Liao, Romain Couillet, et al. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.

[54] Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019.

[55] Gaétan Marceau-Caron and Yann Ollivier. Practical riemannian neural networks. *arXiv preprint arXiv:1602.08007*, 2016.

[56] James Martens. Deep learning via hessian-free optimization. In *ICML*, volume 27, pages 735–742, 2010.

[57] James Martens. New insights and perspectives on the natural gradient method. *arXiv preprint arXiv:1412.1193*, 2014.

[58] James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pages 2408–2417, 2015.

[59] James Martens and Ilya Sutskever. Training deep and recurrent networks with hessian-free optimization. In *Neural networks: Tricks of the trade*, pages 479–535. Springer, 2012.

[60] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

[61] Stanislav Minsker. On some extensions of Bernstein's inequality for self-adjoint operators. *Statistics & Probability Letters*, 127:111–119, 2017.

[62] Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

[63] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *arXiv preprint arXiv:1912.02292*, 2019.

[64] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[65] Yann Ollivier. Riemannian metrics for neural networks i: feedforward networks. *Information and Inference: A Journal of the IMA*, 4(2):108–153, 2015.

[66] Razvan Pascanu and Yoshua Bengio. Revisiting natural gradient for deep networks. *arXiv preprint arXiv:1301.3584*, 2013.

[67] Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. In *Advances in Neural Information Processing Systems*, pages 8114–8124, 2018.

[68] Qian Qian and Xiaoyuan Qian. The implicit bias of adagrad on separable data. In *Advances in Neural Information Processing Systems*, pages 7759–7767, 2019.

[69] Noam Razin and Nadav Cohen. Implicit regularization in deep learning may not be explainable by norms. *arXiv preprint arXiv:2005.06398*, 2020.

[70] Francisco Rubio and Xavier Mestre. Spectral convergence for a general class of random matrices. *Statistics & probability letters*, 81(5):592–602, 2011.

[71] Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, pages 3215–3225, 2017.

[72] Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

[73] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

[74] Ingo Steinwart, Don R Hush, and Clint Scovel. Optimal rates for regularized least squares regression. In *COLT*, pages 79–93, 2009.

[75] Lili Su and Pengkun Yang. On learning over-parameterized neural networks: A functional approximation perspective. In *Advances in Neural Information Processing Systems*, pages 2637–2646, 2019.

[76] Arun Suggala, Adarsh Prasad, and Pradeep K Ravikumar. Connecting optimization and regularization paths. In *Advances in Neural Information Processing Systems*, pages 10608–10619, 2018.

[77] Valentin Thomas, Fabian Pedregosa, Bart Merriënboer, Pierre-Antoine Manzagol, Yoshua Bengio, and Nicolas Le Roux. On the interplay between noise and curvature and its effect on optimization and generalization. In *International Conference on Artificial Intelligence and Statistics*, pages 3503–3513, 2020.

[78] Sharan Vaswani, Reza Babanezhad, Jose Gallego, Aaron Mishkin, Simon Lacoste-Julien, and Nicolas Le Roux. To each optimizer a norm, to each norm its generalization. *arXiv preprint arXiv:2006.06821*, 2020.

[79] Neha S Wadia, Daniel Duckworth, Samuel S Schoenholz, Ethan Dyer, and Jascha Sohl-Dickstein. Whitening and second order optimization both destroy information about the dataset, and can make generalization impossible. *arXiv preprint arXiv:2008.07545*, 2020.

[80] Francis Williams, Matthew Trager, Claudio Silva, Daniele Panozzo, Denis Zorin, and Joan Bruna. Gradient dynamics of shallow univariate relu networks. *arXiv preprint arXiv:1906.07842*, 2019.

[81] Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems*, pages 4148–4158, 2017.

[82] Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. *arXiv preprint arXiv:2002.09277*, 2020.

[83] 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.

[84] Bo Xie, Yingyu Liang, and Le Song. Diverse neural network learns true target functions. *arXiv preprint arXiv:1611.03131*, 2016.

[85] Ji Xu and Daniel Hsu. How many variables should be entered in a principal component regression equation? *arXiv preprint arXiv:1906.01139*, 2019.

[86] Peng Xu, Fred Roosta, and Michael W Mahoney. Second-order optimization for non-convex machine learning: An empirical study. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, pages 199–207. SIAM, 2020.

[87] Guodong Zhang, Chaoqi Wang, Bowen Xu, and Roger Grosse. Three mechanisms of weight decay regularization. *arXiv preprint arXiv:1810.12281*, 2018.

[88] Guodong Zhang, Lala Li, Zachary Nado, James Martens, Sushant Sachdeva, George Dahl, Chris Shallue, and Roger B Grosse. Which algorithmic choices matter at which batch sizes? insights from a noisy quadratic model. In *Advances in Neural Information Processing Systems*, pages 8194–8205, 2019.

[89] Guodong Zhang, James Martens, and Roger B Grosse. Fast convergence of natural gradient descent for over-parameterized neural networks. In *Advances in Neural Information Processing Systems*, pages 8080–8091, 2019.

Appendix A. Discussion on Additional Results

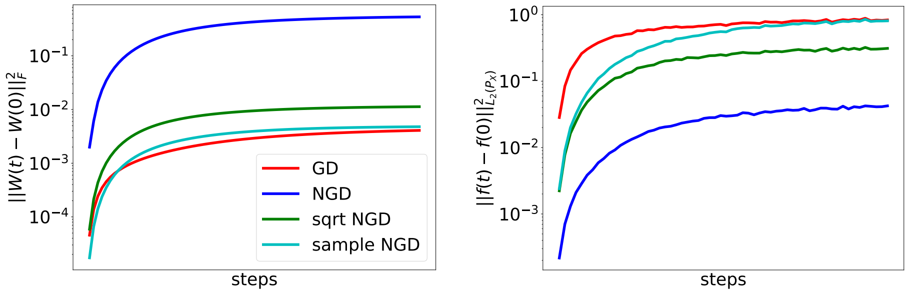
A.1. Implicit Bias of GD vs. NGD

It is known that gradient descent is the steepest descent with respect to the ℓ_2 norm, i.e. the update direction is constructed to decrease the loss under small changes in the parameters measured by the ℓ_2 norm [32]. Following this analogy, NGD is the steepest descent with respect to the KL divergence on the predictive distributions [57]; this can be interpreted as a proximal update which penalizes how much the predictions change on the data distribution.

Intuitively, the above discussion suggests GD tend to find solution that is close to the initialization in the Euclidean distance between parameters, whereas NGD prefers solution close to the initialization in terms of the function outputs on P_X. This observation turns out to be exact in the case of ridgeless interpolant under the squared loss, as remarked in Section 2. Moreover, Figure 1 and 7 confirms the same trend in neural network optimization. In particular, we observe that

- GD results in small changes in the parameters, and NGD results in small changes in the function.
- preconditioning with the pseudo-inverse of the sample Fisher, i.e., $\boldsymbol{P} = (\boldsymbol{J}^\top \boldsymbol{J})^\dagger$, leads to implicit bias similar to that of GD, but not NGD with the population Fisher.
- interpolating between GD and NGD ($\boldsymbol{P} = \boldsymbol{F}^{-1/2}$) results in properties in between GD and NGD.

Remark 4 *The small change in the function output is the essential reason that NGD performs well under noisy labels: NGD seeks to interpolate training data by changing the function only "locally", so that memorizing noisy labels has small impact on the "global" shape of the learned function.*



(a) Difference in Parameters.

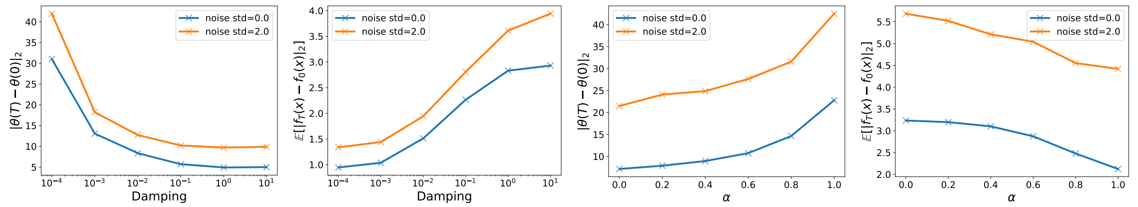
(b) Difference in Function Values.

Figure 7: Illustration of implicit bias of GD and NGD. We set $n = 100$, $d = 50$, and regress a two-layer ReLU network with 50 hidden units towards a teacher model of the same architecture on Gaussian input. The x-axis is rescaled for each optimizer such that the final training error is below 10^{-3}. GD finds solution with small changes in the parameters, whereas NGD finds solution with small changes in the function. Note that the sample Fisher (cyan) has implicit bias similar to GD and does not resemble NGD (population Fisher).

We note that the above observation also implies that wide neural networks trained with NGD (population Fisher) is less likely to stay in the kernel regime: the distance traveled from initialization can be large (see Figure 7(a)) and thus the Taylor expansion around the initialization is no longer accurate. In other words, the analogy between wide neural net and its linearized kernel model (which we partially employed in Section 4) may not be valid in models trained with NGD[3].

[3] Note that this gap is only present when the population Fisher is used; previous works have shown NTK-type global convergence for sample Fisher-related update [17, 89].

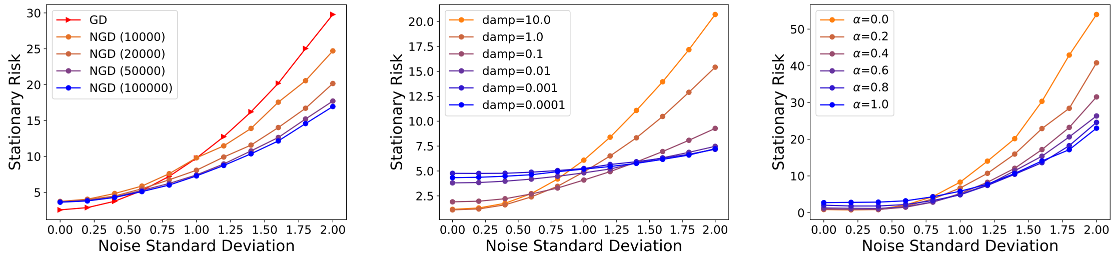
Implicit Bias of Interpolating Preconditioners. We also expect that as we interpolate from GD to NGD, the distance traveled by the parameter space would gradually increase, and distance traveled in the function space would decrease. Figure 8 demonstrate that this is indeed the case for linear model as well as neural network: we use the same two-layer MLP setup on MNIST as in Section 4. Note that updates that are closer to GD result in smaller change in the parameters, whereas ones close to NGD lead to smaller change in the function outputs.



(a) additive interp.; difference in parameters. (b) additive interp.; difference in functions. (c) geometric interp.; difference in parameters. (d) geometric interp.; difference in functions.

Figure 8: Illustration of the implicit bias of preconditioned gradient descent that interpolates between GD and NGD on MNIST. As the update becomes more similar to NGD (smaller damping or larger α), the distance traveled in the parameter space increases, where as the distance traveled on the output space decreases.

A.2. Interpolating between Preconditioners



(a) interpolation between sample and population Fisher (CIFAR-10). (b) additive interpolation between GD and NGD (MNIST). (c) geometric interpolation between GD and NGD (MNIST).

Figure 9: (a) numbers in parentheses indicate the amount of unlabeled data used in estimating the Fisher $\boldsymbol{F}$; we expect the estimated Fisher to be closer to the sample Fisher when the number of unlabeled data is small. (a) additive interpolation $\boldsymbol{P}=(\boldsymbol{F}+\alpha\boldsymbol{I}_d)^{-1}$; larger damping parameter yields update closer to GD. (b) geometric interpolation $\boldsymbol{P}=\boldsymbol{F}^{-\alpha}$; larger α parameter yields update closer to that of NGD (blue). We use the singular value decomposition to compute the minus α power of the Fisher (CG is not applicable).

We empirically validate our observations in Section 2 and 3 on the difference between the sample Fisher and population Fisher, and the potential benefit of interpolating between GD and NGD, in neural network experiments. Figure 9(a) shows that as we decrease the number of unlabeled data in estimating the Fisher, which renders the preconditioner closer to the sample Fisher, the stationary risk becomes more akin to that of GD, especially in the large noise setting. This agrees with our remark on sample vs. population Fisher in Section 2 and Appendix A.

Figure 9(b)(c) confirms the finding in Section 3.1 that interpolating preconditioners provides bias-variance tradeoff also holds in neural network settings: We optimize two-layer MLP student model with preconditioned update that interpolates between GD and NGD either additively ($\boldsymbol{P} = (\boldsymbol{F}+\alpha\boldsymbol{I}_d)^{-1}$) or geometrically ($\boldsymbol{P} = \boldsymbol{F}^{-\alpha}$). We interpret the left end of the x-axis to correspond to a bias-dominant regime (due to the same architecture of two-layer MLP for the student and teacher), and the right end to correspond to the variance-dominant regime (due to the added label noise). Observe that at a certain SNR, a preconditioner that interpolates between GD and NGD achieves lower stationary risk.

A.3. Non-monotonicity of the Bias Term

Many previous works on the high-dimensional characterization of linear regression assumed a random effects model with an isotropic prior on the true parameters [24, 34, 85], which may not be realistic. As an example of the limitation of this assumption, we note that when $\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \boldsymbol{I}_d$, it can be shown that the expected bias $B(\hat{\boldsymbol{\theta}}(t))$ monotonically decreases through time (see proof of Proposition 6 for details). In contrast, when the target parameters do not follow an isotropic prior, the bias of GD can exhibit non-monotonicity, which gives rise to the surprising "epoch-wise double descent" phenomenon also observed in deep learning [36, 63].

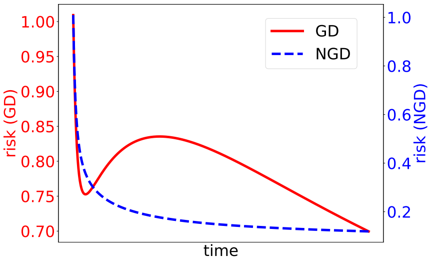


Figure 10: Epoch-wise double descent. Note that non-monotonicity of the bias term is present in GD but not NGD.

We empirically demonstrate this non-monotonicity when the model is close to the interpolation threshold in Figure 10. We set eigenvalues of $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ to be two equally-weighted point masses with $\kappa_X = 32$, $\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}$ and $\gamma = 16/15$. Note that the GD trajectory (red) exhibits non-monotonicity in the bias term, whereas for NGD the bias is monotonically decreasing through time (which we confirm in the proof of Proposition 6). We remark that this mechanism of epoch-wise double descent may not be related to the empirical findings in deep neural networks (the robustness of which is also unknown), in which it is typically speculated that the variance term exhibits non-monotonicity.

A.4. Heuristic Measure of Misspecification

To quantify the level of misspecification, we compute the quantity $\sqrt{\boldsymbol{y}^\top \boldsymbol{K}^{-1}\boldsymbol{y}/n}$, which relates to generalization of wide neural networks in the kernel regime. This quantity can be interpreted as a proxy for measuring how much signal and noise are distributed along the eigendirections of the NTK (see [18, 25, 48, 75] for detailed discussion). Roughly speaking, large value implies that the problem is difficult to learn by GD, and vice versa. The quantity also relates to the HSIC [29] between the NTK features and the true labels.

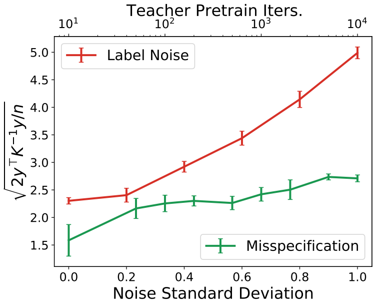


Figure 11: Neural net experiment (CIFAR): $\sqrt{\boldsymbol{y}\boldsymbol{K}^{-1}\boldsymbol{y}}$ vs. label noise or pretrained iterations of the teacher.

Here we give a heuristic argument on how $\sqrt{\boldsymbol{y}^\top \boldsymbol{K}^{-1}\boldsymbol{y}/n}$ relates to label noise and misspecification in our setup. For the ridgeless regression model considered in Section 2, if we write the label as $y_i = f^*(\boldsymbol{x}_i) + f^c(\boldsymbol{x}_i) + \varepsilon_i$, where $f^*(\boldsymbol{x}) = \boldsymbol{x}^\top\boldsymbol{\theta}^*$, f^c is the misspecified component, and ε_i is the label noise, we have the following non-rigorous calculation:

$$\mathbb{E}\Big[\boldsymbol{y}^\top \boldsymbol{K}^{-1}\boldsymbol{y}\Big] = \mathbb{E}\Big[(f^*(\boldsymbol{X}) + f^c(\boldsymbol{X}) + \boldsymbol{\varepsilon})^\top (\boldsymbol{X}\boldsymbol{X}^\top)^{-1}(f^*(\boldsymbol{X}) + f^c(\boldsymbol{X}) + \boldsymbol{\varepsilon})\Big]$$
$$\overset{(i)}{\approx} \mathrm{tr}\Big(\boldsymbol{\theta}^*\boldsymbol{\theta}^{*\top}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}\Big) + (\sigma^2 + \sigma_c^2)\mathrm{tr}\Big((\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\Big), \tag{A.1}$$

where we heuristically replaced the misspecified component with i.i.d. noise of the same variance σ_c^2 (as argued in Section 2). The first term of (A.1) resembles an RKHS norm of the target $\boldsymbol{\theta}^*$, whereas the second term is small when the data is well-conditioned or when the level of label noise σ and misspecification σ_c^2 is small (note that these are conditions under which GD achieves good generalization by Theorem 1 and Proposition 3). We expect similar trends for neural networks close to the kernel regime. This provides a non-rigorous explanation of the trend we observed in Figure 11: $\sqrt{\boldsymbol{y}^\top \boldsymbol{K}^{-1}\boldsymbol{y}/n}$ becomes larger as we increase the level of label noise and model misspecification (i.e. number of pretrain steps of the teacher model).

Appendix B. Background and Related Works

Natural Gradient Descent. NGD is a second-order optimization method originally proposed in [2]. Consider a data distribution $p(\boldsymbol{x})$ on the space $\mathcal{X}$, a function $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{Z}$ parameterized by $\boldsymbol{\theta}$, and a loss function $L(\boldsymbol{X}, f_{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^n l(y_i, f_{\boldsymbol{\theta}}(\boldsymbol{x}_i))$, where $l : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}$. Also suppose a probability distribution $p(y|\boldsymbol{z}) = p(y|f_{\boldsymbol{\theta}}(\boldsymbol{x}))$ is defined on the space of labels as part of the model. Then, the natural gradient is the direction of steepest ascent in the Fisher information norm given by $\tilde{\nabla}_\theta L(\boldsymbol{X}, f_{\boldsymbol{\theta}}) = \boldsymbol{F}^{-1}\nabla_\theta L(\boldsymbol{X}, f_{\boldsymbol{\theta}})$, where

$$\boldsymbol{F} = \mathbb{E}[\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{x}, y|\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{x}, y|\boldsymbol{\theta})^\top] = -\mathbb{E}[\nabla^2_{\boldsymbol{\theta}} \log p(\boldsymbol{x}, y|\boldsymbol{\theta})] \tag{B.1}$$

is the *Fisher information matrix*, or simply the (population) Fisher. Note the expectations in (B.1) are under the joint distribution of the model $p(\boldsymbol{x}, y|\boldsymbol{\theta}) = p(\boldsymbol{x})p(y|f_{\boldsymbol{\theta}}(\boldsymbol{x}))$. In the literature, the Fisher is sometimes defined under the empirical data distribution, i.e. based on a finite set of training examples $\{\boldsymbol{x}_i\}_{i=1}^n$ [4]. We instead refer to this quantity as the *sample Fisher*, the properties of which influence optimization and have been studied in various works [40, 45, 77]. Note that in linear and kernel regression (unregularized) under the squared loss, sample Fisher-based preconditioned updates give the same stationary solution as GD (see [89] and Section 2), whereas population Fisher-based update may not.

While the population Fisher is typically difficult to obtain, extra unlabeled data can be used in its estimation, which empirically improves generalization under appropriately damping [66]. Moreover, under structural assumptions, estimating the Fisher with parametric approaches can be more sample-efficient [30, 55, 58, 65], and thus closing the gap between the sample and population Fisher.

When the per-instance loss l is the negative log-probability of an exponential family, the sample Fisher coincides with the *generalized Gauss-Newton matrix* [57]. In least squares regression, which is the focus of this work, the quantity also coincides with the Hessian due to the linear prediction function. Therefore, we take NGD as a representative example of preconditioned update, and we expect our findings to also translate to other second-order methods (not including adaptive gradient methods) applied to regression problems.

Implicit Regularization in Optimization. In overparameterized linear models, GD finds the minimum ℓ_2 norm solution under many loss functions. For the more general mirror descent, the implicit bias is determined by the Bregman divergence of the update [7, 8, 33, 76]. Under the exponential or logistic loss, recent works demonstrated that GD finds the max-margin direction in various models [20, 38, 39, 54, 73]. The implicit bias of Adagrad has been analyzed under similar setting [68]. The implicit regularization of the optimizer often relates to the model architecture; examples include matrix factorization [5, 28, 31, 72] and various types of neural network [33, 49, 80, 82]. For neural networks in the kernel regime [37], the implicit bias of GD relates to properties of the limiting neural tangent kernel (NTK) [6, 15, 84]. We also note that the implicit bias of GD is not always explained by the minimum norm property [69].

Asymptotics of Interpolating Estimators. In Section 2 we analyze overparameterized estimators that interpolate the training data. Recent works have shown that interpolation may not lead to overfitting [11, 13, 14, 50], and the optimal risk may be achieved under no regularization and extreme overparameterization [12, 85]. The asymptotic risk of overparameterized models has been characterized in various settings, such as linear regression [24, 34, 41], random features regression [21, 27, 60], max-margin classification [22, 62], and certain neural networks [10, 53]. Our analysis is based on results in random matrix theory developed in [46, 70]. Similar tools can also be used to study the gradient descent dynamics of linear regression [1, 51].

Analysis of Preconditioned Gradient Descent. While Wilson et al. [81] outlined one example under fixed training data where GD generalizes better than adaptive methods, in the online learning setting, for which optimization speed relates to generalization, several works have shown the advantage of preconditioning [47, 88]. In addition, global convergence and generalization guarantees were derived for the sample Fisher-based update in neural networks in the kernel regime [17, 89]. Lastly, the generalization of different optimizers also connects to "sharpness" of the solution [23, 43], and it has been argued that second-order updates tend to find sharper minima [83].

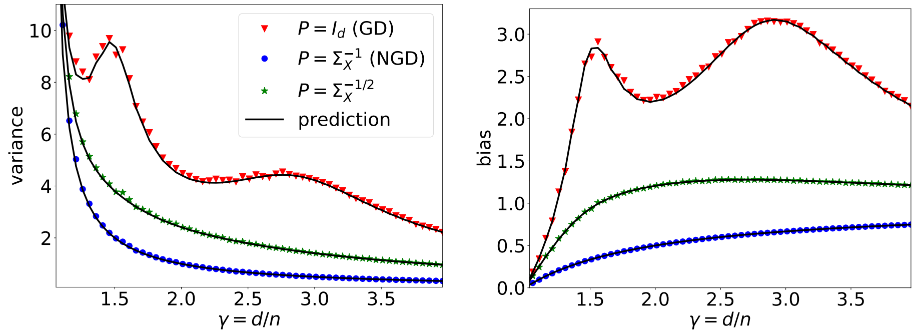
We note that two concurrent works also discussed the generalization performance of preconditioned updates. Wadia et al. [79] connected second-order methods with data whitening in linear models, and qualitatively showed that whitening (thus second-order update) harms generalization in certain cases. Vaswani et al. [78] analyzed the complexity of the maximum $\boldsymbol{P}$-margin solution in linear classification problems. We emphasize that instead of *upper bounding* the risk (e.g. Rademacher complexity), which may not decide the optimal $\boldsymbol{P}$ (for generalization), we compute the *exact risk* for least squares regression, which allows us to precisely compare different preconditioners.

Appendix C. Additional Figures

C.1. Ridgeless Regression

Non-monotonicity of the Risk. Under our generalized (anisotropic) assumption on the covariance of the features and the target, both the bias and the variance term can exhibit non-monotonicity w.r.t. the overparameterization level $\gamma > 1$: in Figure 12 we observe two peaks in the bias term and three peaks in the variance term. In contrast, it is known that when $\boldsymbol{\Sigma}_{\boldsymbol{X}} = \boldsymbol{I}_d$, both the bias and variance are *monotonic* for when $\gamma > 1$.

Additional Figures for Section 2 and 3. We include additional figures on (a) well-specified bias when $\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \boldsymbol{I}_d$ (GD is optimal); (b) misspecified bias under unobserved features (predicted by

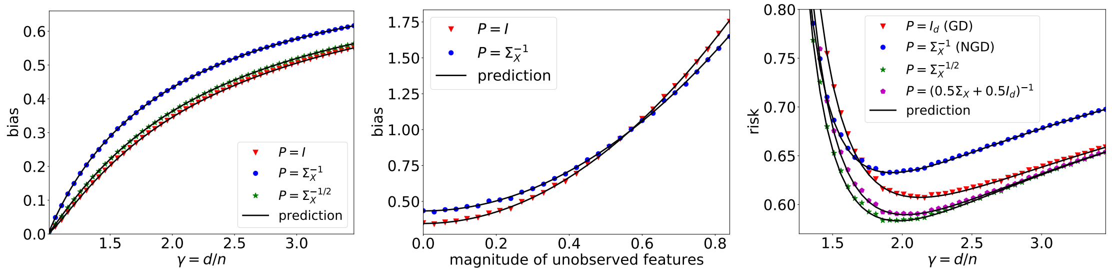


(a) variance.

(b) bias (well-specified).

Figure 12: Illustration of the "multiple-descent" curve of the risk for $\gamma > 1$. We take $n = 300$, eigenvalues of $\boldsymbol{\Sigma_X}$ as three equally-spaced point masses with $\kappa_X = 5000$ and $\|\boldsymbol{\Sigma_X}\|_F^2 = d$, and $\boldsymbol{\Sigma_\theta} = \boldsymbol{\Sigma_X}^{-1}$ (misaligned). Note that for GD, both the bias and the variance are highly non-monotonic for $\gamma > 1$.

Proposition 3); (c) bias-variance tradeoff by interpolating between preconditioners (SNR=5). Observe that in all cases the experimental values match the theoretical predictions.

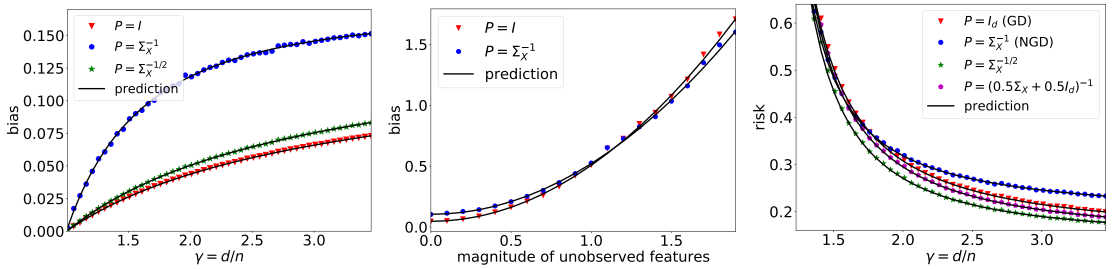


(a) well-specified bias (aligned).

(b) misspecified bias (unobserved features).

(c) bias-variance tradeoff.

Figure 13: We set eigenvalues of $\boldsymbol{\Sigma_X}$ as a uniform distribution with $\kappa_X = 20$ and $\|\boldsymbol{\Sigma_X}\|_F^2 = d$.



(a) well-specified bias (aligned).

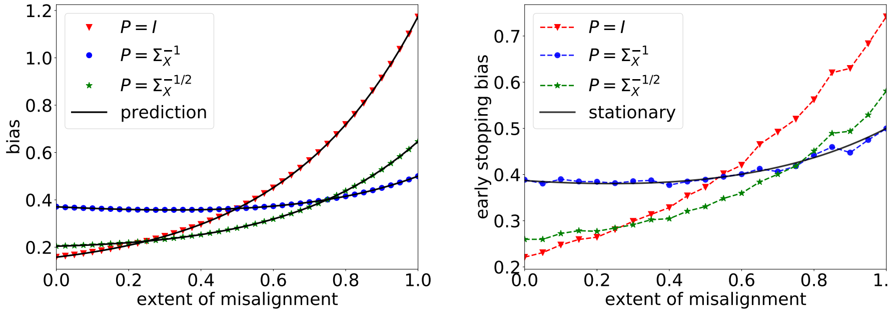
(b) misspecified bias (unobserved features).

(c) bias-variance tradeoff.

Figure 14: We construct eigenvalues of $\boldsymbol{\Sigma_X}$ with a polynomial decay: $\lambda_i(\boldsymbol{\Sigma_X}) = i^{-1}$ and then rescale the eigenvalues such that $\kappa_X = 500$ and $\|\boldsymbol{\Sigma_X}\|_F^2 = d$.

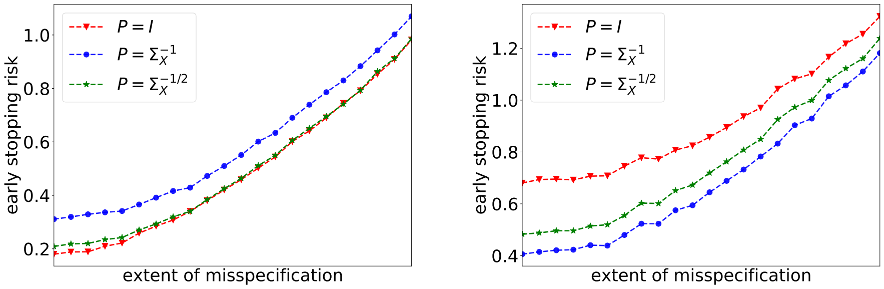
Early Stopping Risk. Figure 15 compares the stationary risk with the optimal early stopping risk under varying misalignment level. To increase the extent of misalignment, we set $\boldsymbol{\Sigma_\theta} = \boldsymbol{\Sigma_X}^{-\alpha}$ and vary α from 0 to 1: larger α entails more "misaligned" teacher, and vice versa. Note that as the problem becomes more misaligned, NGD achieves lower stationary and early stopping risk.

Figure 16 reports the optimal early stopping risk under misspecification (same trend can be obtained when the x-axis is label noise). In contrast to the stationary risk (Figure 4), GD can be advantageous under early stopping even with large extent of misspecification (for isotropic teacher). This aligns with our finding in Section 3.2 that early stopping reduces the variance and the misspecified bias.



(a) stationary risk. (b) optimal early stopping risk.

Figure 15: Well-specified bias against different extent of "alignment". We set $n = 300$, eigenvalues of $\boldsymbol{\Sigma_X}$ as two point masses with $\kappa_X = 20$, and take $\boldsymbol{\Sigma_\theta} = \boldsymbol{\Sigma_X^{-\alpha}}$ and vary α from 0 to 1. (a) GD achieves lower bias when $\boldsymbol{\Sigma_\theta}$ is isotropic, whereas NGD dominates when $\boldsymbol{\Sigma_X} = \boldsymbol{\Sigma_\theta^{-1}}$; $\boldsymbol{P} = \boldsymbol{\Sigma_X^{-1/2}}$ (interpolates between GD and NGD) is advantageous in between. (b) optimal early stopping bias follows similar trend as stationary bias.



(a) optimal early stopping risk (aligned & misspecified). (b) optimal early stopping risk (misaligned & misspecified).

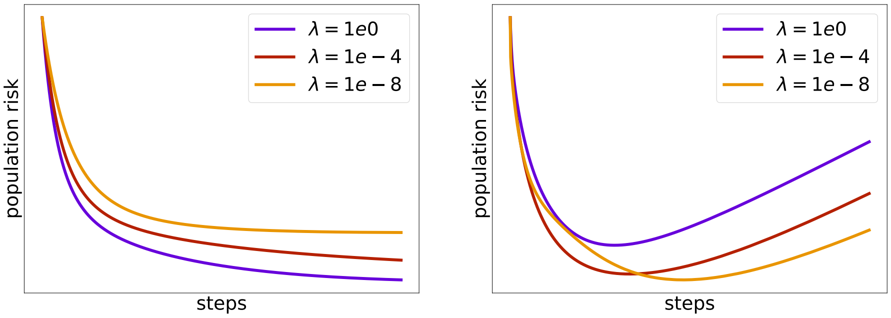
Figure 16: Optimal early stopping risk vs. increasing model misspecification. We follow the same setup as Figure 3(c). (a) $\boldsymbol{\Sigma_\theta} = \boldsymbol{I}_d$ (favors GD); unlike Figure 3(c), GD has lower early stopping risk even under large extent of misspecification. (b) $\boldsymbol{\Sigma_\theta} = \boldsymbol{\Sigma_X^{-1}}$ (favors NGD); NGD is also advantageous under early stopping.

C.2. RKHS Regression

We simulate the optimization in the coordinates of RKHS via a finite-dimensional approximation (using extra unlabeled data). In particular, we consider the teacher model in the form of $f^*(\boldsymbol{x}) = \sum_{i=1}^N h_i \mu_i^r \phi_i(\boldsymbol{x})$ for square summable $\{h_i\}_{i=1}^N$, in which r controls the "difficulty" of the learning problem. We find $\{\mu_i\}_{i=1}^N$ and $\{\phi_i\}_{i=1}^N$ by solving the eigenfunction problem for some kernel k.

The student model takes the form of $f(\boldsymbol{x}) = \sum_{i=1}^{N} \frac{a_i}{\sqrt{\mu_i}} \phi_i(\boldsymbol{x})$ and we optimize the coefficients $\{a_i\}_{i=1}^{N}$ via the preconditioned update (3.1). We set $n = 1000$, $d = 5$, $N = 2500$ and consider the inverse multiquadratic (IMQ) kernel: $k(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{\sqrt{1+\|\boldsymbol{x}-\boldsymbol{y}\|_2^2}}$.

Recall that Theorem 7 suggests that for small r, i.e. "difficult" problem, the damping coefficient λ would need to be small (which makes the update NGD-like), and vice versa. This result is (qualitatively) supported by Figure 17, from which we can see that small λ is beneficial when r is small, and vice versa. We remark that this observed trend is rather fragile and sensitive to various hyperparameters, and we leave a comprehensive characterization of this observation as future work.



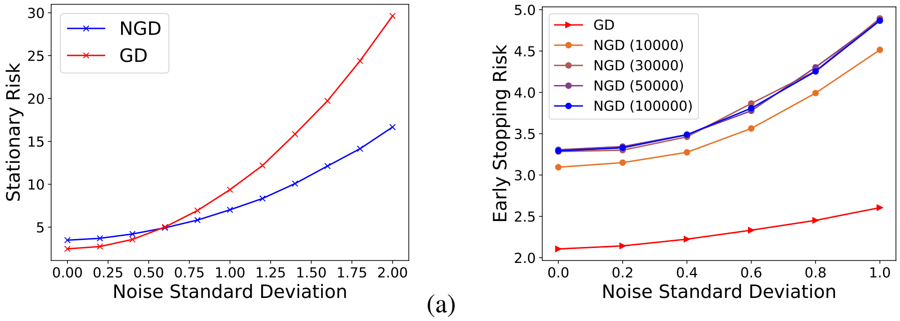
(a) $r = 3/4$.

(b) $r = 1/4$.

Figure 17: Population risk of the preconditioned update in RKHS that interpolates between GD and NGD. We use the IMQ kernel and set $n = 1000$, $d = 5$, $N = 2500$, $\sigma^2 = 5 \times 10^{-4}$. The x-axis has been rescaled for each curve and thus convergence speed is not directly comparable. Note that (a) large λ (i.e., GD-like update) is beneficial when r is large, and (b) small λ (i.e., NGD-like update) is beneficial when r is small.

C.3. Neural Networks

Label Noise. In Figure 18, (a) we observe the same phenomenon on CIFAR-10 that NGD generalizes better as more label noise is added to the training data, and vice versa. Figure 18 (b) shows that in all cases with varying amounts of label noise, the early stopping risk is however worse than that of GD. This agrees with the observation in Section 3 and Figure 16(a) that early stopping can potentially favor GD due to the reduced variance.



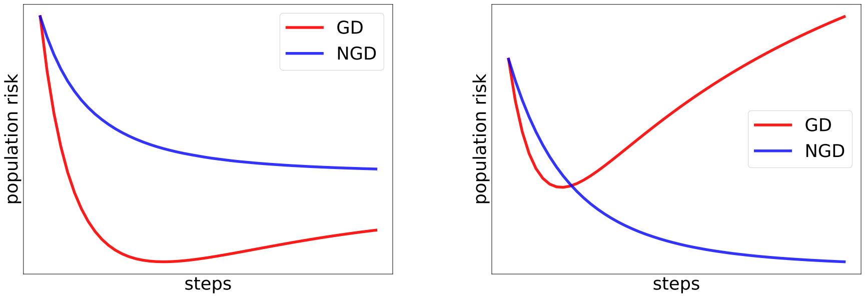
(a)

stationary risk.

(b) optimal early stopping risk.

Figure 18: Additional label noise experiment on CIFAR-10.

Misalignment. We illustrate the finding in Proposition 6 and Figure 15(b) in neural networks under synthetic data: we consider 50-dimensional Gaussian input, and both the teacher and the student model are two-layer ReLU networks with 50 hidden units. We construct the teacher by perturbing the initialization of the student as described in Section 4. Figure 19 shows that as r approaches -1 (more "misaligned"), NGD achieves lower early stopping risk (b), whereas GD dominates the early stopping risk in less misaligned setting (a). We note that this phenomenon is difficult to observe in practical neural network training on real-world data, which may be partially due to the fragility of the analogy between neural nets and linear models, especially under NGD (discussed in Appendix A).



(a) $r = -1/2$.

(b) $r = -3/4$.

Figure 19: Population risk of two-layer neural networks in the misalignment setup (noiseless) with synthetic Gaussian data. We set $n = 200$, $d = 50$, the damping coefficient $\lambda = 10^{-6}$, and both the student and the teacher are two-layer ReLU networks with 50 hidden units. The x-axis and the learning rate have been rescaled for each curve. When r is sufficiently small, NGD achieves lower early stopping risk, and vice versa.

Appendix D. Proofs and Derivations

D.1. Missing Derivations in Section 2

Gradient Flow of Preconditioned Updates. Given positive definite $\boldsymbol{P}$ and $\gamma > 1$, it is clear that the gradient flow solution at time t can be written as

$$\boldsymbol{\theta}_{\boldsymbol{P}}(t) = \boldsymbol{P}\boldsymbol{X}^\top \left[\boldsymbol{I}_n - \exp\left(-\frac{t}{n}\boldsymbol{X}\boldsymbol{P}\boldsymbol{X}^\top\right)\right](\boldsymbol{X}\boldsymbol{P}\boldsymbol{X}^\top)^{-1}\boldsymbol{y}.$$

Taking $t \to \infty$ yields the stationary solution $\hat{\boldsymbol{\theta}}_{\boldsymbol{P}} = \boldsymbol{P}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{P}\boldsymbol{X}^\top)^{-1}\boldsymbol{y}$. We remark that the damped inverse of the sample Fisher $\boldsymbol{P} = (\boldsymbol{X}\boldsymbol{X}^\top + \lambda\boldsymbol{I}_d)^{-1}$ leads to the same minimum-norm solution as GD $\hat{\boldsymbol{\theta}}_{\boldsymbol{I}} = \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y}$ since $\boldsymbol{P}\boldsymbol{X}^\top$ and $\boldsymbol{X}$ share the same eigenvectors. On the other hand, when $\boldsymbol{P}$ is the pseudo-inverse of the sample Fisher $(\boldsymbol{X}\boldsymbol{X}^\top)^\dagger$ which is not full-rank, the trajectory can be obtained via the variation of constants formula:

$$\boldsymbol{\theta}(t) = \left[\frac{t}{n}\sum_{k=0}^{\infty}\frac{1}{(k+1)!}\left(-\frac{t}{n}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}\right)^k\right]\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y},$$

for which taking the large t limit also yields the minimum-norm solution $\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y}$.

Minimum $\|\boldsymbol{\theta}\|_{\boldsymbol{P}^{-1}}$ Norm Interpolant. For positive definite $\boldsymbol{P}$ and the corresponding stationary solution $\hat{\boldsymbol{\theta}}_{\boldsymbol{P}} = \boldsymbol{P}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{P}\boldsymbol{X}^\top)^{-1}\boldsymbol{y}$, note that given any other interpolant $\hat{\boldsymbol{\theta}}'$, we have $(\hat{\boldsymbol{\theta}}_{\boldsymbol{P}} - \hat{\boldsymbol{\theta}}')\boldsymbol{P}^{-1}\hat{\boldsymbol{\theta}}_{\boldsymbol{P}} = 0$ because both $\hat{\boldsymbol{\theta}}_{\boldsymbol{P}}$ and $\hat{\boldsymbol{\theta}}'$ achieves zero empirical risk. Therefore, $\|\hat{\boldsymbol{\theta}}'\|^2_{\boldsymbol{P}^{-1}} -$

$\|\hat{\boldsymbol{\theta}}_{\boldsymbol{P}}\|^2_{\boldsymbol{P}^{-1}} = \|\hat{\boldsymbol{\theta}}' - \hat{\boldsymbol{\theta}}_{\boldsymbol{P}}\|^2_{\boldsymbol{P}^{-1}} \geq 0$. This confirms that $\hat{\boldsymbol{\theta}}_{\boldsymbol{P}}$ is the unique minimum $\|\boldsymbol{\theta}\|_{\boldsymbol{P}^{-1}}$ norm solution.

D.2. Proof of Theorem 1

Proof By the definition of the variance term and the stationary $\hat{\boldsymbol{\theta}}$,

$$V(\hat{\boldsymbol{\theta}}) = \mathrm{tr}\Big(\mathrm{Cov}(\hat{\boldsymbol{\theta}})\boldsymbol{\Sigma}_{\boldsymbol{X}}\Big) = \sigma^2 \mathrm{tr}\Big(\boldsymbol{P}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{P}\boldsymbol{X}^\top)^{-2}\boldsymbol{X}\boldsymbol{P}\boldsymbol{\Sigma}_{\boldsymbol{X}}\Big).$$

Write $\bar{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{P}^{1/2}$. Similarly, we define $\boldsymbol{\Sigma}_{\boldsymbol{X}\boldsymbol{P}} = \boldsymbol{P}^{1/2}\boldsymbol{\Sigma}_{\boldsymbol{X}}\boldsymbol{P}^{1/2}$. The equation above thus simplifies to

$$V(\hat{\boldsymbol{\theta}}_{\boldsymbol{P}}) = \sigma^2 \mathrm{tr}\Big(\bar{\boldsymbol{X}}^\top(\bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^\top)^{-2}\bar{\boldsymbol{X}}\boldsymbol{\Sigma}_{\boldsymbol{X}\boldsymbol{P}}\Big).$$

The analytic expression of the variance term follows from a direct application of Hastie et al. [34, Thorem 4], in which conditions on the population covariance are satisfied by (A2).

Taking the derivative of $m(-\lambda)$ yields

$$m'(-\lambda) = \left(\frac{1}{m^2(-\lambda)} - \gamma\int\frac{\tau^2}{(1+\tau m(-\lambda))^2}\mathrm{d}\boldsymbol{H}_{\boldsymbol{X}\boldsymbol{P}}(\tau)\right)^{-1}.$$

Plugging the quantity into the expression of the variance (omitting the scaling σ^2 and constant shift),

$$\frac{m'(-\lambda)}{m^2(-\lambda)} = \left(1 - \gamma m^2(-\lambda)\int\frac{\tau^2}{(1+\tau m(-\lambda))^2}\mathrm{d}\boldsymbol{H}_{\boldsymbol{X}\boldsymbol{P}}(\tau)\right)^{-1}.$$

From the monotonicity of $\frac{x}{1+x}$ on $x > 0$ or the Jensen's inequality we know that

$$1 - \gamma\int\left(\frac{\tau m(-\lambda)}{1+\tau m(-\lambda)}\right)^2\mathrm{d}\boldsymbol{H}_{\boldsymbol{X}\boldsymbol{P}}(\tau) \leq 1 - \gamma\left(\int\frac{\tau m(-\lambda)}{1+\tau m(-\lambda)}\mathrm{d}\boldsymbol{H}_{\boldsymbol{X}\boldsymbol{P}}(\tau)\right)^2.$$

Taking $\lambda \to 0$ and omitting the scalar σ^2, the RHS evaluates to $1 - 1/\gamma$. We thus arrive at the lower bound $V \geq (\gamma-1)^{-1}$. Note that the equality is only achieved when $\boldsymbol{H}_{\boldsymbol{X}\boldsymbol{P}}$ is a point mass, i.e. $\boldsymbol{P} = \boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}$. In other words, the minimum variance is achieved by NGD. As a verification, the variance of the NGD solution $\hat{\boldsymbol{\theta}}_{\boldsymbol{F}^{-1}}$ agrees with the calculation in Hastie et al. [34, A.3]. ■

D.3. Proof of Theorem 2

Proof By the definition of the bias term (note that $\boldsymbol{\Sigma}_{\boldsymbol{X}}$, $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$, $\boldsymbol{P}$ are all positive semi-definite),

$$\begin{aligned}
B(\hat{\boldsymbol{\theta}}_{\boldsymbol{P}}) &= \mathbb{E}_{\boldsymbol{\theta}^*}\left[\left\|\boldsymbol{P}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{P}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}\boldsymbol{\theta}_* - \boldsymbol{\theta}^*\right\|^2_{\boldsymbol{\Sigma}_{\boldsymbol{X}}}\right] \\
&= \frac{1}{d}\mathrm{tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\Big(\boldsymbol{I}_d - \boldsymbol{P}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{P}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}\Big)^\top\boldsymbol{\Sigma}_{\boldsymbol{X}}\Big(\boldsymbol{I}_d - \boldsymbol{P}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{P}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}\Big)\right) \\
&\overset{(i)}{=} \frac{1}{d}\mathrm{tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}}\Big(\boldsymbol{I}_d - \bar{\boldsymbol{X}}^\top(\bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^\top)^{-1}\bar{\boldsymbol{X}}\Big)^\top\boldsymbol{\Sigma}_{\boldsymbol{X}\boldsymbol{P}}\Big(\boldsymbol{I}_d - \bar{\boldsymbol{X}}^\top(\bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^\top)^{-1}\bar{\boldsymbol{X}}\Big)\right)
\end{aligned}$$

$$\begin{aligned}
&\overset{(ii)}{=} \lim_{\lambda\to 0_+} \frac{\lambda^2}{d}\mathrm{tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}}\left(\frac{1}{n}\bar{\boldsymbol{X}}^\top\bar{\boldsymbol{X}}+\lambda\boldsymbol{I}_d\right)^{-1}\boldsymbol{\Sigma}_{\boldsymbol{X}\boldsymbol{P}}\left(\frac{1}{n}\bar{\boldsymbol{X}}^\top\bar{\boldsymbol{X}}+\lambda\boldsymbol{I}_d\right)^{-1}\right)\\
&\overset{(iii)}{=} \lim_{\lambda\to 0_+} \frac{\lambda^2}{d}\mathrm{tr}\left(\left(\frac{1}{n}\hat{\boldsymbol{X}}^\top\hat{\boldsymbol{X}}+\lambda\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}}^{-1}\right)^{-2}\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}}^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{X}\boldsymbol{P}}\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}}^{-1/2}\right),
\end{aligned}$$

where we utilized (A3) and defined $\bar{\boldsymbol{X}}=\boldsymbol{X}\boldsymbol{P}^{1/2}$, $\boldsymbol{\Sigma}_{\boldsymbol{X}\boldsymbol{P}}=\boldsymbol{P}^{1/2}\boldsymbol{\Sigma}_{\boldsymbol{X}}\boldsymbol{P}^{1/2}$, $\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}}=\boldsymbol{P}^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{P}^{-1/2}$ in (i), applied the equality $(\boldsymbol{A}\boldsymbol{A}^\top)^\dagger\boldsymbol{A}=\lim_{\lambda\to 0}(\boldsymbol{A}^\top\boldsymbol{A}+\lambda\boldsymbol{I})^{-1}\boldsymbol{A}$ in (ii), and defined $\hat{\boldsymbol{X}}=\boldsymbol{X}\boldsymbol{P}^{1/2}\boldsymbol{\Sigma}_\theta^{-1/2}$ in (iii). To proceed, we first assume that $\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}}$ is invertible (i.e. $\lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}})$ is bounded away from 0) and observe the following relation via a leave-one-out argument similar to that in [85],

$$\frac{1}{d}\mathrm{tr}\left(\frac{1}{n}\hat{\boldsymbol{X}}^\top\hat{\boldsymbol{X}}\left(\frac{1}{n}\hat{\boldsymbol{X}}^\top\hat{\boldsymbol{X}}+\lambda\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}}^{-1}\right)^{-2}\right) \tag{D.1}$$

$$\begin{aligned}
&\overset{(i)}{=}\frac{1}{d}\sum_{i=1}^{n}\frac{\frac{1}{n}\hat{\boldsymbol{x}}_i^\top\left(\frac{1}{n}\hat{\boldsymbol{X}}^\top\hat{\boldsymbol{X}}+\lambda\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}}^{-1}\right)_{\neg i}^{-2}\hat{\boldsymbol{x}}_i}{\left(1+\frac{1}{n}\hat{\boldsymbol{x}}_i^\top\left(\frac{1}{n}\hat{\boldsymbol{X}}^\top\hat{\boldsymbol{X}}+\lambda\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}}^{-1}\right)_{\neg i}^{-1}\hat{\boldsymbol{x}}_i\right)^2}\\
&\underset{p}{\overset{(ii)}{\to}}\frac{\frac{1}{d}\mathrm{tr}\left(\left(\frac{1}{n}\hat{\boldsymbol{X}}^\top\hat{\boldsymbol{X}}+\lambda\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}}^{-1}\right)^{-2}\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}}^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{X}\boldsymbol{P}}\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}}^{-1/2}\right)}{\left(1+\frac{1}{n}\mathrm{tr}\left(\left(\frac{1}{n}\bar{\boldsymbol{X}}^\top\bar{\boldsymbol{X}}+\lambda\boldsymbol{I}_d\right)^{-1}\boldsymbol{\Sigma}_{\boldsymbol{X}\boldsymbol{P}}\right)\right)^2},
\end{aligned}\tag{D.2}$$

where (i) is due to the Woodbury identity and we defined $\left(\frac{1}{n}\hat{\boldsymbol{X}}^\top\hat{\boldsymbol{X}}+\lambda\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}}^{-1}\right)_{\neg i}=\frac{1}{n}\hat{\boldsymbol{X}}^\top\hat{\boldsymbol{X}}-\frac{1}{n}\hat{\boldsymbol{x}}_i\hat{\boldsymbol{x}}_i^\top+\lambda\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}}^{-1}$ which is independent to $\hat{\boldsymbol{x}}_i$ (see Xu and Hsu [85, Eq. 58] for details), and in (ii) we used (A3), the convergence to trace [46, Lemma 2.1] and its stability under low-rank perturbation (e.g., see Ledoit and Péché [46, Eq. 18]) which we elaborate below. In particular, denote $\hat{\boldsymbol{\Sigma}}=\frac{1}{n}\hat{\boldsymbol{X}}^\top\hat{\boldsymbol{X}}+\lambda\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}}^{-1}$, for the denominator we have

$$\begin{aligned}
&\sup_i\left|\frac{\lambda}{n}\mathrm{tr}\left(\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}}^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{X}\boldsymbol{P}}\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}}^{-1/2}\right)-\frac{\lambda}{n}\mathrm{tr}\left(\hat{\boldsymbol{\Sigma}}_{\neg i}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}}^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{X}\boldsymbol{P}}\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}}^{-1/2}\right)\right|\\
\le&\frac{\lambda}{n}\left\|\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}}^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{X}\boldsymbol{P}}\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}}^{-1/2}\right\|_2\sup_i\left|\mathrm{tr}\left(\hat{\boldsymbol{\Sigma}}^{-1}\left(\hat{\boldsymbol{\Sigma}}-\hat{\boldsymbol{\Sigma}}_{\neg i}\right)\hat{\boldsymbol{\Sigma}}_{\neg i}^{-1}\right)\right|\\
\le&\frac{\lambda}{n}\left\|\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}}^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{X}\boldsymbol{P}}\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}}^{-1/2}\right\|_2\left\|\hat{\boldsymbol{\Sigma}}^{-1}\right\|_2\sup_i\left\|\hat{\boldsymbol{\Sigma}}_{\neg i}^{-1}\right\|_2\mathrm{tr}\left(\hat{\boldsymbol{\Sigma}}-\hat{\boldsymbol{\Sigma}}_{\neg i}\right)\overset{(i)}{\to}O_p\left(\frac{1}{n}\right),
\end{aligned}$$

where (i) is due to the definition of $\hat{\boldsymbol{\Sigma}}_{\neg i}$ and (A1)(A3). The result on the numerator can be obtained via a similar calculation, the details of which we omit.

Note that the denominator can be evaluated by previous results (e.g. Dobriban et al. [24, Theorem 2.1]) as follows,

$$\frac{1}{n}\mathrm{tr}\left(\left(\frac{1}{n}\bar{\boldsymbol{X}}^\top\bar{\boldsymbol{X}}+\lambda\boldsymbol{I}_d\right)^{-1}\boldsymbol{\Sigma}_{\boldsymbol{X}\boldsymbol{P}}\right)\overset{a.s.}{\to}\frac{1}{\lambda m(-\lambda)}-1. \tag{D.3}$$

On the other hand, following the same derivation as Dobriban et al. [24], Hastie et al. [34], (D.1) can be decomposed as

$$
\begin{aligned}
&\frac{1}{d}\operatorname{tr}\left(\frac{1}{n}\hat{\boldsymbol{X}}^\top\hat{\boldsymbol{X}}\left(\frac{1}{n}\hat{\boldsymbol{X}}^\top\hat{\boldsymbol{X}}+\lambda\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}}^{-1}\right)^{-2}\right)\\
=&\frac{1}{d}\operatorname{tr}\left(\left(\frac{1}{n}\bar{\boldsymbol{X}}^\top\bar{\boldsymbol{X}}+\lambda\boldsymbol{I}_d\right)^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}}\right)-\frac{\lambda}{d}\operatorname{tr}\left(\left(\frac{1}{n}\bar{\boldsymbol{X}}^\top\bar{\boldsymbol{X}}+\lambda\boldsymbol{I}_d\right)^{-2}\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}}\right)\\
=&\frac{1}{d}\operatorname{tr}\left(\left(\frac{1}{n}\bar{\boldsymbol{X}}^\top\bar{\boldsymbol{X}}+\lambda\boldsymbol{I}_d\right)^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}}\right)+\frac{\lambda}{d}\frac{\mathrm{d}}{\mathrm{d}\lambda}\operatorname{tr}\left(\left(\frac{1}{n}\bar{\boldsymbol{X}}^\top\bar{\boldsymbol{X}}+\lambda\boldsymbol{I}_d\right)^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}}\right). \qquad \text{(D.4)}
\end{aligned}
$$

We employ Rubio and Mestre [70, Theorem 1] to characterize (D.4). In particular, For any deterministic sequence of matrices $\boldsymbol{\Theta}_n \in \mathbb{R}^{d\times d}$ with finite trace norm, as $n, d \to \infty$ we have

$$
\operatorname{tr}\left(\boldsymbol{\Theta}_n\left(\frac{1}{n}\bar{\boldsymbol{X}}^\top\bar{\boldsymbol{X}}-z\boldsymbol{I}_d\right)^{-1}-\boldsymbol{\Theta}_n(c_n(z)\boldsymbol{\Sigma}_{\boldsymbol{XP}}-z\boldsymbol{I}_d)^{-1}\right)\overset{a.s.}{\to}0,
$$

in which $c_n(z) \to -zm(z)$ for $z \in \mathbb{C}\backslash\mathbb{R}^+$ and $m(z)$ is defined in Theorem 1 due to the dominated convergence theorem. By (A3) we are allowed to take $\boldsymbol{\Theta}_n = \frac{1}{d}\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}}$. Thus we have

$$
\begin{aligned}
\frac{\lambda}{d}\operatorname{tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}}\left(\frac{1}{n}\bar{\boldsymbol{X}}^\top\bar{\boldsymbol{X}}+\lambda\boldsymbol{I}_d\right)^{-1}\right)\to&\frac{\lambda}{d}\operatorname{tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}}(\lambda m(-\lambda)\boldsymbol{\Sigma}_{\boldsymbol{XP}}+\lambda\boldsymbol{I}_d)^{-1}\right)\\
&\overset{(i)}{=}\mathbb{E}\left[\frac{\upsilon_x\upsilon_\theta\upsilon_{xp}^{-1}}{1+m(-\lambda)\upsilon_{xp}}\right], \quad \forall\lambda > -c_l, \qquad \text{(D.5)}
\end{aligned}
$$

in which (i) is due to (A3), the fact that the LHS is almost surely bounded for $\lambda > -c_l$, where c_l is the lowest non-zero eigenvalue of $\frac{1}{n}\bar{\boldsymbol{X}}^\top\bar{\boldsymbol{X}}$, and the application of the dominated convergence theorem. Differentiating (D.5) (note that the derivative is also bounded on $\lambda > -c_l$) yields

$$
\frac{\lambda}{d}\frac{\mathrm{d}}{\mathrm{d}\lambda}\operatorname{tr}\left(\left(\frac{1}{n}\bar{\boldsymbol{X}}^\top\bar{\boldsymbol{X}}+\lambda\boldsymbol{I}_d\right)^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}}\right)\to\mathbb{E}\left[\frac{\upsilon_x\upsilon_\theta\upsilon_{xp}^{-1}}{\lambda(1+m(-\lambda)\upsilon_{xp})}-\frac{m'(-\lambda)\upsilon_x\upsilon_\theta}{(1+m(-\lambda)\upsilon_{xp})^2}\right] \text{(D.6)}
$$

Note that the numerator of (D.2) is the quantity of interest. Combining (D.1) (D.2) (D.3) (D.4) (D.5) (D.6) and taking $\lambda \to 0$ yields the formula of the bias term. Finally, when $\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}}$ is not invertible, observe that if we increment all eigenvalues by some $\epsilon > 0$ to ensure invertibility $\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}}^\epsilon = \boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}} + \epsilon\boldsymbol{I}_d$, the bias term is bounded and also decreasing w.r.t. ϵ. Thus by the dominated convergence theorem we take $\epsilon \to 0$ and obtain the desired result. We remark that similar (but less general) characterization can also be derived based on Ledoit and Péché [46, Theorem 1.2] when the eigenvalues of $\boldsymbol{\Sigma}_{\boldsymbol{XP}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}}$ exhibit certain relations.

To show that $\boldsymbol{P} = \boldsymbol{U}\operatorname{diag}(\boldsymbol{e}_\theta)\boldsymbol{U}^\top$ achieves the lowest bias, first note that under the definition of random variables in (A3), our claimed optimal preconditioner is equivalent to $\upsilon_{xp} \overset{a.s.}{=} \upsilon_x\upsilon_\theta$. We therefore define an interpolation $\upsilon_\alpha = \alpha\upsilon_x\upsilon_\theta + (1-\alpha)\bar{\upsilon}$ for some $\bar{\upsilon}$ and write the corresponding Stieltjes transform as $m_\alpha(-\lambda)$ and the bias term as B_α. We aim to show that $\operatorname{argmin}_{\alpha\in[0,1]} B_\alpha = 1$.

For notational convenience define $g_\alpha \triangleq m_\alpha(0)\upsilon_x\upsilon_\theta$ and $h_\alpha \triangleq m_\alpha(0)\upsilon_\alpha$. One can check that

$$B_\alpha = \mathbb{E}\left[\frac{\upsilon_x\upsilon_\theta}{(1+h_\alpha)^2}\right]\mathbb{E}\left[\frac{h_\alpha}{(1+h_\alpha)^2}\right]^{-1}; \quad \frac{\mathrm{d}m_\alpha(-\lambda)}{\mathrm{d}\alpha}\Big|_{\lambda\to 0} = \frac{m_\alpha(0)\mathbb{E}\left[\frac{h_\alpha-g_\alpha}{(1+h_\alpha)^2}\right]}{(1-\alpha)\mathbb{E}\left[\frac{h_\alpha}{(1+h_\alpha)^2}\right]}.$$

We now verify that the derivative of B_α w.r.t. α is non-positive for $\alpha \in [0,1]$. A standard simplification of the derivative yields

$$\begin{aligned}\frac{\mathrm{d}B_\alpha}{\mathrm{d}\alpha} \propto &-2\mathbb{E}\left[\frac{(g_\alpha-h_\alpha)^2}{(1+h_\alpha)^3}\right]\left(\mathbb{E}\left[\frac{h_\alpha}{(1+h_\alpha)^2}\right]\right)^2 - 2\left(\mathbb{E}\left[\frac{g_\alpha-h_\alpha}{(1+h_\alpha)^2}\right]\right)^2\mathbb{E}\left[\frac{h_\alpha^2}{(1+h_\alpha)^3}\right]\\ &+4\mathbb{E}\left[\frac{h_\alpha(g_\alpha-h_\alpha)}{(1+h_\alpha)^3}\right]\mathbb{E}\left[\frac{g_\alpha-h_\alpha}{(1+h_\alpha)^2}\right]\mathbb{E}\left[\frac{h_\alpha}{(1+h_\alpha)^2}\right]\\ \overset{(i)}{\leq} &-4\sqrt{\mathbb{E}\left[\frac{(g_\alpha-h_\alpha)^2}{(1+h_\alpha)^3}\right]\mathbb{E}\left[\frac{h_\alpha^2}{(1+h_\alpha)^3}\right]\left(\mathbb{E}\left[\frac{g_\alpha-h_\alpha}{(1+h_\alpha)^2}\right]\right)^2\left(\mathbb{E}\left[\frac{h_\alpha}{(1+h_\alpha)^2}\right]\right)^2}\\ &+4\mathbb{E}\left[\frac{h_\alpha(g_\alpha-h_\alpha)}{(1+h_\alpha)^3}\right]\mathbb{E}\left[\frac{g_\alpha-h_\alpha}{(1+h_\alpha)^2}\right]\mathbb{E}\left[\frac{h_\alpha}{(1+h_\alpha)^2}\right] \overset{(ii)}{\leq} 0,\end{aligned}$$

where (i) is due to AM-GM and (ii) due to Cauchy-Schwarz on the first term. Note that the two equalities hold when $g_\alpha = h_\alpha$, from which one can easily deduce that the optimum is achieved when $\upsilon_{xp} \overset{a.s.}{=} \upsilon_x\upsilon_\theta$, and thus we know that the proposed $\boldsymbol{P}$ is the optimal preconditioner that is codiagonazable with $\boldsymbol{\Sigma_X}$. ■

D.4. Proof of Proposition 3

Proof Via calculation similar to Hastie et al. [34, Section 5], the bias can be decomposed as

$$\begin{aligned}\mathbb{E}\left[B(\hat{\boldsymbol{\theta}}_{\boldsymbol{P}})\right] =&\mathbb{E}_{\boldsymbol{x},\hat{\boldsymbol{x}},\boldsymbol{\theta}^*,\boldsymbol{\theta}^c}\left[\left(\boldsymbol{x}^\top\boldsymbol{P}\boldsymbol{X}^\top\left(\boldsymbol{X}\boldsymbol{P}\boldsymbol{X}^\top\right)^{-1}(\boldsymbol{X}\boldsymbol{\theta}^*+\boldsymbol{X}^c\boldsymbol{\theta}^c)-(\boldsymbol{x}^\top\boldsymbol{\theta}^*+\hat{\boldsymbol{x}}^\top\boldsymbol{\theta}^c)\right)^2\right]\\ \overset{(i)}{=}&\mathbb{E}_{\boldsymbol{x},\boldsymbol{\theta}^*}\left[\left(\boldsymbol{x}^\top\boldsymbol{P}\boldsymbol{X}^\top\left(\boldsymbol{X}\boldsymbol{P}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{X}\boldsymbol{\theta}^*-\boldsymbol{x}^\top\boldsymbol{\theta}^*\right)^2\right]+\mathbb{E}_{\boldsymbol{x}^c,\boldsymbol{\theta}^x}\left[(\hat{\boldsymbol{x}}^\top\boldsymbol{\theta}^c)^2\right]\\ &+\mathbb{E}_{\boldsymbol{x},\boldsymbol{\theta}^c}\left[\left(\boldsymbol{x}^\top\boldsymbol{P}\boldsymbol{X}^\top\left(\boldsymbol{X}\boldsymbol{P}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{X}^c\boldsymbol{\theta}^c\boldsymbol{\theta}^{c\top}\boldsymbol{X}^{c\top}\left(\boldsymbol{X}\boldsymbol{P}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{X}\boldsymbol{P}\boldsymbol{x}\right)^2\right]\\ \overset{(ii)}{\to}&B_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{P}})+\frac{1}{d^c}\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{X}}^c\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^c)(1+V(\hat{\boldsymbol{\theta}}_{\boldsymbol{P}})),\end{aligned}$$

where we used the independence of $\boldsymbol{x},\hat{\boldsymbol{x}}$ and $\boldsymbol{\theta}^*,\boldsymbol{\theta}^c$ in (i), and (A1-3) as well as the definition of the well-specified bias $B_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{P}})$ and variance $V(\hat{\boldsymbol{\theta}}_{\boldsymbol{P}})$ in (ii). ■

D.5. Proof of Proposition 4

Proof We first outline a more general setup where $\boldsymbol{P}_\alpha = f(\boldsymbol{\Sigma_X};\alpha)$ for continuous and differentiable function of α and f applied to eigenvalues of $\boldsymbol{\Sigma_x}$. For any interval $\mathcal{I} \subseteq [0,1]$, we claim

(a) Suppose all four functions $\frac{1}{xf(x;\alpha)}$, $f(x;\alpha)$, $\frac{\partial f(x;\alpha)}{\partial \alpha}/f(x;\alpha)$ and $x\frac{\partial f(x;\alpha)}{\partial \alpha}$ are decreasing functions of x on the support of v_x for all $\alpha \in \mathcal{I}$. In addition, $\frac{\partial f(x;\alpha)}{\partial \alpha} \geq 0$ on the support of v_x for all $\alpha \in \mathcal{I}$. Then the stationary bias is an increasing function of α on $\mathcal{I}$.

(b) For all $\alpha \in \mathcal{I}$, suppose $xf(x;\alpha)$ is a monotonic function of x on the support of v_x and $\frac{\partial f(x;\alpha)}{\partial \alpha}/f(x;\alpha)$ is a decreasing function of x on the support of v_x. Then the stationary variance is a decreasing function of α on $\mathcal{I}$.

Let us verify the three choices of $\boldsymbol{P}_\alpha$ in Proposition 4 one by one.

- When $\boldsymbol{P}_\alpha = (1-\alpha)\boldsymbol{I}_d + \alpha(\boldsymbol{\Sigma}_{\boldsymbol{X}})^{-1}$, the corresponding $f(x;\alpha)$ is $(1-\alpha)+\alpha x$. It is clear that it satisfies all conditions in (a) and (b) for all $\alpha \in [0,1]$. Hence, the stationary variance is a decreasing function and the stationary bias is an increasing function of $\alpha \in [0,1]$.
- When $\boldsymbol{P}_\alpha = (\boldsymbol{\Sigma}_{\boldsymbol{X}})^{-\alpha}$, the corresponding $f(x;\alpha)$ is $x^{-\alpha}$. It is clear that it satisfies all conditions in (a) and (b) for all $\alpha \in [0,1]$ except for the condition that $x\frac{\partial f(x;\alpha)}{\partial \alpha} = -x^{1-\alpha}\ln x$ is a decreasing function of x. Note that $x\frac{\partial f(x;\alpha)}{\partial \alpha} = -x^{1-\alpha}\ln x$ is a decreasing function of x on the support of v_x only for $\alpha \geq \frac{\ln(\kappa)-1}{\ln(\kappa)}$ where $\kappa = \sup v_x/\inf v_x$. Hence, the stationary variance is a decreasing function of $\alpha \in [0,1]$ and the stationary bias is an increasing function of $\alpha \in [\max(0,\frac{\ln(\kappa)-1}{\ln(\kappa)}),1]$.
- When $\boldsymbol{P}_\alpha = (\alpha\boldsymbol{\Sigma}_{\boldsymbol{X}}+(1-\alpha)\boldsymbol{I}_d)^{-1}$, the corresponding $f(x;\alpha)$ is $1/(\alpha x+(1-\alpha))$. It is clear that it satisfies all conditions in (a) and (b) for all $\alpha \in [0,1]$ except for the condition that $x\frac{\partial f(x;\alpha)}{\partial \alpha} = \frac{x(1-x)}{(\alpha x+(1-\alpha))^2}$ is a decreasing function of x. Note that $x\frac{\partial f(x;\alpha)}{\partial \alpha} = \frac{x(1-x)}{(\alpha x+(1-\alpha))^2}$ is a decreasing function of x on the support of v_x only for $\alpha \geq \frac{\kappa-2}{\kappa-1}$. Hence, the stationary variance is a decreasing function of $\alpha \in [0,1]$ and the stationary bias is an increasing function of $\alpha \in [\max(0,\frac{\kappa-2}{\kappa-1}),1]$.

To show (a) and (b), note that under the conditions on $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ assumed in Proposition 4, the stationary bias $B(\hat{\boldsymbol{\theta}}_{\boldsymbol{P}_\alpha})$ and the stationary variance $V(\hat{\boldsymbol{\theta}}_{\boldsymbol{P}_\alpha})$ can be simplified to

$$B(\hat{\boldsymbol{\theta}}_{\boldsymbol{P}_\alpha}) = \frac{m'_\alpha(0)}{m^2_\alpha(0)}\mathbb{E}\frac{v_x}{(1+v_xf(v_x;\alpha)m_\alpha(0))^2} \quad \text{and} \quad V(\hat{\boldsymbol{\theta}}_{\boldsymbol{P}_\alpha}) = \sigma^2\cdot\left(\frac{m'_\alpha(0)}{m^2_\alpha(0)}-1\right),$$

where $m_\alpha(z)$ and $m'_\alpha(z)$ satisfy

$$1 = -zm_\alpha(z) + \gamma\mathbb{E}\frac{v_xf(v_x;\alpha)m_\alpha(z)}{1+v_xf(v_x;\alpha)m_\alpha(z)} \tag{D.7}$$

$$\frac{m'_\alpha(z)}{m^2_\alpha(z)} = \frac{1}{1-\gamma\mathbb{E}\left(\frac{f(v_x;\alpha)m_\alpha(z)}{1+f(v_x;\alpha)m_\alpha(z)}\right)^2}. \tag{D.8}$$

For notation convenience, let $f_\alpha := v_xf(v_x;\alpha)$. From (D.8), we have the following equivalent expressions.

$$B(\hat{\boldsymbol{\theta}}_{\boldsymbol{P}_\alpha}) = \frac{\mathbb{E}\frac{v_x}{(1+f_\alpha m_\alpha(0))^2}}{1-\gamma\mathbb{E}\left(\frac{f_\alpha m_\alpha(0)}{1+f_\alpha m_\alpha(0)}\right)^2}, \tag{D.9}$$

$$V(\hat{\boldsymbol{\theta}}_{\boldsymbol{P}_\alpha}) = \sigma^2\left(\frac{1}{1-\gamma\mathbb{E}\left(\frac{f_\alpha m_\alpha(0)}{1+f_\alpha m_\alpha(0)}\right)^2}-1\right). \tag{D.10}$$

We first show that (b) holds. Note that from (D.10), we have

$$\frac{\partial V(\hat{\boldsymbol{\theta}}_{\boldsymbol{P}_\alpha})}{\partial \alpha} = \gamma\sigma^2 \left(\frac{1}{1-\gamma\mathbb{E}\left(\frac{f_\alpha m_\alpha(0)}{1+f_\alpha m_\alpha(0)}\right)^2}\right)^2 \mathbb{E}\left[\frac{2f_\alpha m_\alpha(0)}{(1+f_\alpha m_\alpha(0))^3}\left(f_\alpha \frac{\partial m_\alpha(z)}{\partial\alpha}\Big|_{z=0} + \frac{\partial f_\alpha}{\partial \alpha} m_\alpha(0)\right)\right]. \tag{D.11}$$

To calculate $\frac{\partial m_\alpha(z)}{\partial \alpha}\Big|_{z=0}$, we take derivatives with respect to α on both sides of (D.7),

$$0 = \gamma\mathbb{E}\left[\frac{1}{(1+f_\alpha m_\alpha(0))^2}\cdot\left(f_\alpha \frac{\partial m_\alpha(z)}{\partial\alpha}\Big|_{z=0} + \frac{\partial f_\alpha}{\partial \alpha} m_\alpha(0)\right)\right]. \tag{D.12}$$

Therefore, plugging (D.12) into (D.11) yields

$$\begin{aligned}\frac{\partial V(\hat{\boldsymbol{\theta}}_{\boldsymbol{P}_\alpha})}{\partial \alpha} =& 2\gamma\sigma^2 \left(\frac{m_\alpha(0)}{1-\gamma\mathbb{E}\left(\frac{f_\alpha m_\alpha(0)}{1+f_\alpha m_\alpha(0)}\right)^2}\right)^2 \left(\mathbb{E}\frac{f_\alpha}{(1+f_\alpha m_\alpha(0))^2}\right)^{-1} \\ & \times\left(\mathbb{E}\frac{f_\alpha\frac{\partial f_\alpha}{\partial\alpha}}{(1+f_\alpha m_\alpha(0))^3}\mathbb{E}\frac{f_\alpha}{(1+f_\alpha m_\alpha(0))^2} - \mathbb{E}\frac{f_\alpha^2}{(1+f_\alpha m_\alpha(0))^3}\mathbb{E}\frac{\frac{\partial f_\alpha}{\partial\alpha}}{(1+f_\alpha m_\alpha(0))^2}\right)\end{aligned}$$

Thus showing $V(\hat{\boldsymbol{\theta}}_{\boldsymbol{P}_\alpha})$ is a decreasing function of α is equivalent to showing that

$$\mathbb{E}\frac{f_\alpha^2}{(1+f_\alpha m_\alpha(0))^3}\mathbb{E}\frac{\frac{\partial f_\alpha}{\partial\alpha}}{(1+f_\alpha m_\alpha(0))^2} \geq \mathbb{E}\frac{f_\alpha\frac{\partial f_\alpha}{\partial\alpha}}{(1+f_\alpha m_\alpha(0))^3}\mathbb{E}\frac{f_\alpha}{(1+f_\alpha m_\alpha(0))^2}. \tag{D.13}$$

Let μ_x be the probability measure of v_x. We define a new measure $\tilde{\mu}_x = \frac{f_\alpha \mu_x}{(1+f_\alpha m_\alpha(0))^2}$, and let $\tilde{v}_x$ follow the new measure. Since $\frac{\partial f(x;\alpha)}{\partial\alpha}/f(x;\alpha)$ is a decreasing function of x and $xf(x;\alpha)$ is a monotonic function of x,

$$\mathbb{E}\frac{\tilde{v}_x f(\tilde{v}_x;\alpha)}{1+\tilde{v}_x f(\tilde{v}_x;\alpha)m_\alpha(0)}\mathbb{E}\frac{\frac{\partial \tilde{v}_x f(\tilde{v}_x;\alpha)}{\partial\alpha}}{\tilde{v}_x f(\tilde{v}_x;\alpha)} \geq \mathbb{E}\frac{\frac{\partial \tilde{v}_x f(\tilde{v}_x;\alpha)}{\partial\alpha}}{1+\tilde{v}_x f(\tilde{v}_x;\alpha)m_\alpha(0)}.$$

Changing $\tilde{v}_x$ back to v_x, we arrive at (D.13) and thus (b).

For the bias term $B(\hat{\boldsymbol{\theta}}_{\boldsymbol{P}_\alpha})$, note that from (D.7) and (D.9), we have

$$\begin{aligned}\frac{\partial B(\hat{\boldsymbol{\theta}}_{\boldsymbol{P}_\alpha})}{\partial \alpha} =& \frac{1}{\gamma}\left(\frac{1}{\gamma} - \mathbb{E}\left(\frac{f_\alpha m_\alpha(0)}{1+f_\alpha m_\alpha(0)}\right)^2\right)^{-2} \\ & \times\left(-\mathbb{E}\left[2\frac{v_x}{(1+f_\alpha m_\alpha(0))^3}\cdot\left(f_\alpha \frac{\partial m_\alpha(z)}{\partial\alpha}\Big|_{z=0} + \frac{\partial f_\alpha}{\partial \alpha} m_\alpha(0)\right)\right]\mathbb{E}\frac{f_\alpha m_\alpha(0)}{(1+f_\alpha m_\alpha(0))^2}\right. \\ & \left. + \mathbb{E}\frac{v_x}{(1+f_\alpha m_\alpha(0))^2}\mathbb{E}\left[2\frac{f_\alpha m_\alpha(0)}{(1+f_\alpha m_\alpha(0))^3}\cdot\left(f_\alpha \frac{\partial m_\alpha(z)}{\partial\alpha}\Big|_{z=0} + \frac{\partial f_\alpha}{\partial \alpha} m_\alpha(0)\right)\right]\right)\end{aligned} \tag{D.14}$$

Similarly, we combine (D.12) and (D.14) and simplify the expression. To verify $B(\hat{\boldsymbol{\theta}}_{\boldsymbol{P}_\alpha})$ is an increasing function of α, we need to show that

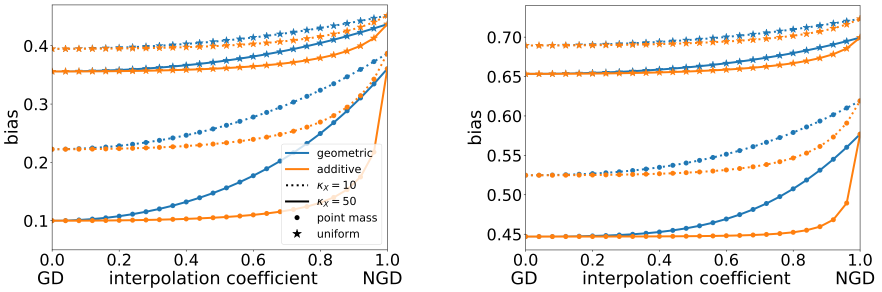
$$0 \leq \left(\mathbb{E}\frac{v_x f_\alpha m_\alpha(0)}{(1+f_\alpha m_\alpha(0))^3}\mathbb{E}\frac{\frac{\partial f_\alpha}{\partial\alpha}}{(1+f_\alpha m_\alpha(0))^2} - \mathbb{E}\frac{v_x\frac{\partial f_\alpha}{\partial\alpha}}{(1+f_\alpha m_\alpha(0))^3}\mathbb{E}\frac{f_\alpha m_\alpha(0)}{(1+f_\alpha m_\alpha(0))^2}\right)\mathbb{E}\frac{f_\alpha m_\alpha(0)}{(1+f_\alpha m_\alpha(0))^2}$$

$$-\mathbb{E}\frac{v_x}{(1+f_\alpha m_\alpha(0))^2}\left(\mathbb{E}\frac{(f_\alpha m_\alpha(0))^2}{(1+f_\alpha m_\alpha(0))^3}\mathbb{E}\frac{\frac{\partial f_\alpha}{\partial\alpha}}{(1+f_\alpha m_\alpha(0))^2}-\mathbb{E}\frac{f_\alpha m_\alpha(0)\frac{\partial f_\alpha}{\partial\alpha}}{(1+f_\alpha m_\alpha(0))^3}\mathbb{E}\frac{f_\alpha m_\alpha(0)}{(1+f_\alpha m_\alpha(0))^2}\right), \tag{D.15}$$

Let $h_\alpha \triangleq f_\alpha m_\alpha(0) = v_x f(v_x;\alpha)m_\alpha(0)$ and $g_\alpha \triangleq \frac{\partial f_\alpha}{\partial\alpha} = v_x\frac{\partial f(v_x;\alpha)}{\partial\alpha}$. Then (D.15) can be further simplified to the following equation

$$\begin{aligned}
0 \le\ & \underbrace{\mathbb{E}\frac{v_x h_\alpha}{(1+h_\alpha)^3}\mathbb{E}\frac{g_\alpha}{(1+h_\alpha)^3}\mathbb{E}\frac{h_\alpha}{(1+h_\alpha)^3}-\mathbb{E}\frac{v_x}{(1+h_\alpha)^3}\mathbb{E}\frac{g_\alpha}{(1+h_\alpha)^3}\mathbb{E}\frac{h_\alpha^2}{(1+h_\alpha)^3}}_{\text{part 1}}\\
&+\underbrace{\mathbb{E}\frac{v_x}{(1+h_\alpha)^3}\mathbb{E}\frac{g_\alpha h_\alpha}{(1+h_\alpha)^3}\mathbb{E}\frac{h_\alpha}{(1+h_\alpha)^3}-\mathbb{E}\frac{v_x g_\alpha}{(1+h_\alpha)^3}\mathbb{E}\frac{h_\alpha}{(1+h_\alpha)^3}\mathbb{E}\frac{h_\alpha}{(1+h_\alpha)^3}}_{\text{part 2}}\\
&+\underbrace{2\mathbb{E}\frac{v_x h_\alpha}{(1+h_\alpha)^3}\mathbb{E}\frac{g_\alpha h_\alpha}{(1+h_\alpha)^3}\mathbb{E}\frac{h_\alpha}{(1+h_\alpha)^3}-2\mathbb{E}\frac{v_x g_\alpha}{(1+h_\alpha)^3}\mathbb{E}\frac{h_\alpha^2}{(1+h_\alpha)^3}\mathbb{E}\frac{h_\alpha}{(1+h_\alpha)^3}}_{\text{part 3}}\\
&+\underbrace{\mathbb{E}\frac{v_x h_\alpha}{(1+h_\alpha)^3}\mathbb{E}\frac{g_\alpha h_\alpha}{(1+h_\alpha)^3}\mathbb{E}\frac{h_\alpha^2}{(1+h_\alpha)^3}-\mathbb{E}\frac{v_x g_\alpha}{(1+h_\alpha)^3}\mathbb{E}\frac{h_\alpha^2}{(1+h_\alpha)^3}\mathbb{E}\frac{h_\alpha^2}{(1+h_\alpha)^3}}_{\text{part 4}}.
\end{aligned} \tag{D.16}$$

Note that under condition of (a), we know that both h_α and v_x/h_α are increasing functions of v_x; and both g_α/h_α and g_α are decreasing functions of v_x. Hence, with calculation similar to (D.13), we know part 1,2,3,4 in (D.16) are all non-negative, and therefore (D.16) holds. ∎



(a) $\gamma = 2$.

(b) $\gamma = 5$.

Figure 20: Illustration of the monotonicity of the bias term under $\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \boldsymbol{I}_d$. We consider two distributions of eigenvalues for $\boldsymbol{\Sigma}_{\boldsymbol{X}}$: two equally weighted point masses (circle) and a uniform distribution (star), and vary the condition number κ_X and overparameterization level γ. In all cases the bias in monotone in $\alpha \in [0, 1]$.

Remark 5 *The above characterization provides sufficient but not necessary conditions for the monotonicity of the bias term. In general, the expression of the bias is rather opaque, and determining the sign of its derivative can be tedious, except for certain special cases (e.g. $\gamma = 2$ and the eigenvalues of $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ are two equally weighted point masses, for which m_α has a simple form and one may analytically check the monotonicity). We conjecture that the bias is monotone for $\alpha \in [0, 1]$ for a much wider class of $\boldsymbol{\Sigma}_{\boldsymbol{X}}$, as shown in Figure 20.*

D.6. Proof of Proposition 5

Proof Taking the derivative of $V(\boldsymbol{\theta}_{\boldsymbol{P}}(t))$ w.r.t. time yields (omitting the scalar σ^2),

$$\frac{\mathrm{d}V(\boldsymbol{\theta}_{\boldsymbol{P}}(t))}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\left\|\boldsymbol{\Sigma}_{\boldsymbol{X}}^{1/2}\boldsymbol{P}\boldsymbol{X}^\top\left(\boldsymbol{I}_n - \exp\left(-\frac{t}{n}\boldsymbol{X}\boldsymbol{P}\boldsymbol{X}^\top\right)\right)\left(\boldsymbol{X}\boldsymbol{P}\boldsymbol{X}^\top\right)^{-1}\right\|_F^2$$

$$\overset{(i)}{=}\frac{1}{n}\mathrm{tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{XP}}\underbrace{\bar{\boldsymbol{X}}^\top\boldsymbol{S}_{\boldsymbol{P}}\exp\left(-\frac{t}{n}\boldsymbol{S}_{\boldsymbol{P}}\right)\boldsymbol{S}_{\boldsymbol{P}}^{-2}\left(\boldsymbol{I}_n - \exp\left(-\frac{t}{n}\boldsymbol{S}_{\boldsymbol{P}}\right)\right)\bar{\boldsymbol{X}}}_{p.s.d.}\right)\overset{(ii)}{>}0,$$

where we defined $\bar{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{P}^{1/2}$ and $\boldsymbol{S}_{\boldsymbol{P}} = \boldsymbol{X}\boldsymbol{P}\boldsymbol{X}^\top$ in (i), and (ii) is due to (A2-3) the inequality $\mathrm{tr}(\boldsymbol{A}\boldsymbol{B}) \geq \lambda_{\min}(\boldsymbol{A})\mathrm{tr}(\boldsymbol{B})$ for positive semi-definite $\boldsymbol{A}$ and $\boldsymbol{B}$. ■

D.7. Proof of Proposition 6

Proof Recall the definition of the bias (well-specified) of $\hat{\boldsymbol{\theta}}_{\boldsymbol{P}}(t)$,

$$B(\boldsymbol{\theta}_{\boldsymbol{P}}(t)) \overset{(i)}{=}\frac{1}{d}\mathrm{tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\left(\boldsymbol{I}_d - \boldsymbol{P}\boldsymbol{X}^\top\boldsymbol{W}_{\boldsymbol{P}}(t)\boldsymbol{S}_{\boldsymbol{P}}^{-1}\boldsymbol{X}\right)^\top\boldsymbol{\Sigma}_{\boldsymbol{X}}\left(\boldsymbol{I}_d - \boldsymbol{P}\boldsymbol{X}^\top\boldsymbol{W}_{\boldsymbol{P}}(t)\boldsymbol{S}_{\boldsymbol{P}}^{-1}\boldsymbol{X}\right)\right)$$

$$\overset{(ii)}{=}\frac{1}{d}\mathrm{tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}}\left(\boldsymbol{I}_d - \bar{\boldsymbol{X}}^\top\boldsymbol{W}_{\boldsymbol{P}}(t)\boldsymbol{S}_{\boldsymbol{P}}^{-1}\bar{\boldsymbol{X}}\right)^\top\boldsymbol{\Sigma}_{\boldsymbol{XP}}\left(\boldsymbol{I}_d - \bar{\boldsymbol{X}}^\top\boldsymbol{W}_{\boldsymbol{P}}(t)\boldsymbol{S}_{\boldsymbol{P}}^{-1}\bar{\boldsymbol{X}}\right)\right)$$

$$\overset{(iii)}{\geq}\frac{1}{d}\mathrm{tr}\left(\left(\boldsymbol{\Sigma}_{\boldsymbol{XP}}^{1/2}\left(\boldsymbol{I}_d - \bar{\boldsymbol{X}}^\top\boldsymbol{W}_{\boldsymbol{P}}(t)\boldsymbol{S}_{\boldsymbol{P}}^{-1}\bar{\boldsymbol{X}}\right)\boldsymbol{\Sigma}_{\boldsymbol{\theta}/\boldsymbol{P}}^{1/2}\right)^2\right), \tag{D.17}$$

where we defined $\boldsymbol{S}_{\boldsymbol{P}} = \boldsymbol{X}\boldsymbol{P}\boldsymbol{X}^\top$, $\boldsymbol{W}_{\boldsymbol{P}}(t) = \boldsymbol{I}_n - \exp\left(-\frac{t}{n}\boldsymbol{S}_{\boldsymbol{P}}\right)$ in (i), $\bar{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{P}^{1/2}$ in (ii), and (iii) is due to the inequality $\mathrm{tr}\left(\boldsymbol{A}^\top\boldsymbol{A}\right) \geq \mathrm{tr}\left(\boldsymbol{A}^2\right)$.

When $\boldsymbol{\Sigma}_{\boldsymbol{X}} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}$, i.e. NGD achieves lowest stationary bias, (D.17) simplifies to

$$B(\boldsymbol{\theta}_{\boldsymbol{P}}(t)) \geq \frac{1}{d}\mathrm{tr}\left(\left(\boldsymbol{I}_d - \bar{\boldsymbol{X}}^\top\boldsymbol{W}_{\boldsymbol{P}}(t)\boldsymbol{S}_{\boldsymbol{P}}^{-1}\bar{\boldsymbol{X}}\right)^2\right) = \left(1 - \frac{1}{\gamma}\right) + \frac{1}{d}\sum_{i=1}^n \exp\left(-\frac{t}{n}\bar{\lambda}_i\right)^2, \tag{D.18}$$

where $\bar{\lambda}$ is the eigenvalue of $\boldsymbol{S}_{\boldsymbol{P}}$. On the other hand, since $\boldsymbol{F} = \boldsymbol{\Sigma}_{\boldsymbol{X}}$, for the NGD iterate $\hat{\boldsymbol{\theta}}_{\boldsymbol{F}^{-1}}(t)$ we have

$$B(\boldsymbol{\theta}_{\boldsymbol{F}^{-1}}(t)) = \frac{1}{d}\mathrm{tr}\left(\left(\boldsymbol{I}_d - \hat{\boldsymbol{X}}^\top\boldsymbol{W}_{\boldsymbol{F}^{-1}}(t)\boldsymbol{S}_{\boldsymbol{F}^{-1}}^{-1}\hat{\boldsymbol{X}}\right)^2\right) = \left(1 - \frac{1}{\gamma}\right) + \frac{1}{d}\sum_{i=1}^n \exp\left(-\frac{t}{n}\hat{\lambda}_i\right)^2 \tag{D.19}$$

where $\hat{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1/2}$ and $\bar{\lambda}$ is the eigenvalue of $\boldsymbol{S}_{\boldsymbol{F}^{-1}} = \hat{\boldsymbol{X}}\hat{\boldsymbol{X}}^\top$. Comparing (D.18)(D.19), we see that given $\hat{\boldsymbol{\theta}}_{\boldsymbol{P}}(t)$ at a fixed t, if we run NGD for time $T > \frac{\bar{\lambda}_{\max}}{\hat{\lambda}_{\min}}t$ (note that $T/t = O(1)$ by (A2-3)), then we have $B(\boldsymbol{\theta}_{\boldsymbol{P}}(t)) \geq B(\boldsymbol{\theta}_{\boldsymbol{F}^{-1}}(T))$ for any $\boldsymbol{P}$ satisfying (A3). This thus implies that $B^{\mathrm{opt}}(\boldsymbol{\theta}_{\boldsymbol{P}}) \geq B^{\mathrm{opt}}(\boldsymbol{\theta}_{\boldsymbol{F}^{-1}})$.

On the other hand, when $\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \boldsymbol{I}_d$, we can show that the bias term of GD is monotonically decreasing through time by taking its derivative,

$$\frac{\mathrm{d}}{\mathrm{d}t}B(\boldsymbol{\theta}_{\boldsymbol{I}}(t)) = \frac{1}{d}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{tr}\left(\left(\boldsymbol{I}_d - \boldsymbol{X}^\top\boldsymbol{W}_{\boldsymbol{I}}(t)\boldsymbol{S}_{\boldsymbol{I}}^{-1}\boldsymbol{X}\right)^\top\boldsymbol{\Sigma}_{\boldsymbol{X}}\left(\boldsymbol{I}_d - \boldsymbol{X}^\top\boldsymbol{W}_{\boldsymbol{I}}(t)\boldsymbol{S}_{\boldsymbol{I}}^{-1}\boldsymbol{X}\right)\right)$$

$$= -\frac{1}{nd}\mathrm{tr}\left(\boldsymbol{\Sigma_X}\underbrace{\boldsymbol{X}^\top\boldsymbol{S}\exp\left(-\frac{t}{n}\boldsymbol{S}\right)\boldsymbol{S}^{-1}\boldsymbol{X}\left(\boldsymbol{I}_d-\boldsymbol{X}^\top\boldsymbol{W_I}(t)\boldsymbol{S_I}^{-1}\boldsymbol{X}\right)}_{p.s.d.}\right) < 0. \tag{D.20}$$

Similarly, one can verify that the expected bias of NGD is monotonically decreasing for all choices of $\boldsymbol{\Sigma_X}$ and $\boldsymbol{\Sigma_\theta}$ satisfying (A2-4),

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{tr}\left(\boldsymbol{\Sigma_\theta}\left(\boldsymbol{I}_d-\boldsymbol{F}^{-1}\boldsymbol{X}^\top\boldsymbol{W}_{\boldsymbol{F}^{-1}}(t)\boldsymbol{S}_{\boldsymbol{F}^{-1}}^{-1}\boldsymbol{X}\right)^\top\boldsymbol{\Sigma_X}\left(\boldsymbol{I}_d-\boldsymbol{F}^{-1}\boldsymbol{X}^\top\boldsymbol{W}_{\boldsymbol{F}^{-1}}(t)\boldsymbol{S}_{\boldsymbol{F}^{-1}}^{-1}\boldsymbol{X}\right)\right)\\ =&\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{tr}\left(\boldsymbol{\Sigma_{X\theta}}\left(\boldsymbol{I}_d-\hat{\boldsymbol{X}}^\top\boldsymbol{W}_{\boldsymbol{F}^{-1}}(t)\boldsymbol{S}_{\boldsymbol{F}^{-1}}^{-1}\hat{\boldsymbol{X}}\right)^\top\left(\boldsymbol{I}_d-\hat{\boldsymbol{X}}^\top\boldsymbol{W}_{\boldsymbol{F}^{-1}}(t)\boldsymbol{S}_{\boldsymbol{F}^{-1}}^{-1}\hat{\boldsymbol{X}}\right)\right)\overset{(i)}{<}0,\end{aligned}$$

where (i) follows from calculation similar to (D.20). Since the expected bias is decreasing through time for both GD and NGD when $\boldsymbol{\Sigma_\theta}=\boldsymbol{I}_d$, and from Theorem 2 we know that $B(\hat{\boldsymbol{\theta}}_{\boldsymbol{I}})\le B(\hat{\boldsymbol{\theta}}_{\boldsymbol{F}^{-1}})$, we conclude that $B^{\mathrm{opt}}(\boldsymbol{\theta}_{\boldsymbol{I}})\le B^{\mathrm{opt}}(\boldsymbol{\theta}_{\boldsymbol{F}^{-1}})$. ■

D.8. Proof of Theorem 7

D.8.1. SETUP AND RESULT

We restate the setting and assumptions. $\mathcal{H}$ is an RKHS included in $L_2(P_X)$ equipped with a bounded kernel function k. $K_{\boldsymbol{x}}\in\mathcal{H}$ is the Riesz representation of the kernel function $k(x,\cdot)$, that is, $k(\boldsymbol{x},\boldsymbol{y})=\langle K_{\boldsymbol{x}},K_{\boldsymbol{y}}\rangle_{\mathcal{H}}$. S is the canonical embedding operator from $\mathcal{H}$ to $L_2(P_X)$. We write $\Sigma=S^*S:\mathcal{H}\to\mathcal{H}$ and $L=SS^*$. Note that the boundedness of the kernel gives $\|Sf\|_{L_2(P_X)}\le\sup_{\boldsymbol{x}}|f(\boldsymbol{x})|=\sup_{\boldsymbol{x}}|\langle K_{\boldsymbol{x}},f\rangle|\le\|K_{\boldsymbol{x}}\|_{\mathcal{H}}\|f\|_{\mathcal{H}}\le\|f\|_{\mathcal{H}}$. Hence we know $\|\Sigma\|\le1$ and $\|L\|\le1$. Our analysis will be made under the following assumptions.

- There exist $r\in(0,\infty)$ and $M>0$ such that $f^*=L^rh^*$ for some $h^*\in L_2(P_X)$ and $\|f^*\|_\infty\le M$.
- There exists $s>1$ s.t. $\mathrm{tr}\left(\Sigma^{1/s}\right)<\infty$ and $2r+s^{-1}>1$.
- There exist $\mu\in[s^{-1},1]$ and $C_\mu>0$ such that $\sup_{\boldsymbol{x}\in\mathrm{supp}(P_X)}\left\|\Sigma^{1/2-1/\mu}K_{\boldsymbol{x}}\right\|_{\mathcal{H}}\le C_\mu$.
- $\sup_{\boldsymbol{x}\in\mathrm{supp}(P_X)}k(\boldsymbol{x},\boldsymbol{x})\le1$.

We remark that (A5-6) are standard assumptions in the literature that provide capacity control of the RKHS [19, 67]. It is also worth noting that previous works mostly consider $r\ge1/2$ which implies $f^*\in\mathcal{H}$.

The training data is generated as $y_i=f^*(\boldsymbol{x}_i)+\varepsilon_i$, where ε_i is an i.i.d. noise satisfying $|\varepsilon_i|\le\sigma$ almost surely. Let $\boldsymbol{y}\in\mathbb{R}^n$ be the label vector. We identify $\mathbb{R}^n$ with $L_2(P_n)$ and define

$$\hat{\Sigma}=\frac{1}{n}\sum_{i=1}^nK_{\boldsymbol{x}_i}\otimes K_{\boldsymbol{x}_i}:\mathcal{H}\to\mathcal{H},\quad \hat{S}^*Y=\frac{1}{n}\sum_{i=1}^nY_iK_{\boldsymbol{x}_i},\ (Y\in L_2(P_n)).$$

We consider the following preconditioned update on $f_t\in\mathcal{H}$:

$$f_t=f_{t-1}-\eta(\Sigma+\lambda I)^{-1}(\hat{\Sigma}f_{t-1}-\hat{S}^*Y),\quad f_0=0.$$

We aim to show the following theorem:

Theorem 8 *Given the assumptions above, if the sample size* n *is sufficiently large so that* $1/(n\lambda) \ll 1$*, then for* $\eta < \|\Sigma\|$ *with* $\eta t \geq 1$ *and* $0 < \delta < 1$ *and* $0 < \lambda < 1$*, it holds that*

$$\|Sf_t - f^*\|^2_{L_2(P_X)} \leq C(B(t) + V(t)),$$

with probability $1 - 3\delta$*, where* C *is a constant and*

$$B(t) := \exp(-\eta t) \vee \left(\frac{\lambda}{\eta t}\right)^{2r},$$

$$V(t) := V_1(t) + (1 + \eta t)\left(\frac{\lambda^{-1}B(t) + \sigma^2 \mathrm{tr}\left(\Sigma^{\frac{1}{s}}\right)\lambda^{-\frac{1}{s}}}{n} + \frac{\lambda^{-1}(\sigma + M + (1 + t\eta)\lambda^{-(\frac{1}{2}-r)_+})^2}{n^2}\right)\log(1/\delta)^2,$$

in which

$$V_1(t) := \left[\exp(-\eta t) \vee \left(\frac{\lambda}{\eta t}\right)^{2r} + (t\eta)^2\left(\frac{\beta'(1 \vee \lambda^{2r-\mu})\mathrm{tr}\left(\Sigma^{\frac{1}{s}}\right)\lambda^{-\frac{1}{s}}}{n} + \frac{\beta'^2(1 + \lambda^{-\mu}(1 \vee \lambda^{2r-\mu})}{n^2}\right)\right](1 + t\eta)^2,$$

for $\beta' = \log\left(\frac{28C_\mu^2(2^{2r-\mu} \vee \lambda^{-\mu+2r})\mathrm{tr}\left(\Sigma^{1/s}\right)\lambda^{-1/s}}{\delta}\right)$*. When* $r \geq 1/2$*, if we set* $\lambda = n^{-\frac{s}{2rs+1}} =: \lambda^*$ *and* $t = \Theta(\log(n))$*, then the overall convergence rate becomes*

$$\|Sg_t - f^*\|^2_{L_2(P_X)} = \widetilde{O}_p\left(n^{-\frac{2rs}{2rs+1}}\right),$$

which is the minimax optimal rate ($\widetilde{O}_p(\cdot)$ *hides a poly-*$\log(n)$ *factor). On the other hand, when* $r < 1/2$*, the bound is also* $\widetilde{O}_p\left(n^{-\frac{2rs}{2rs+1}}\right)$ *except the term* $V_1(t)$*. In this case, if* $2r \geq \mu$ *holds additionally, we have* $V_t(t) = \widetilde{O}_p\left(n^{-\frac{2rs}{2rs+1}}\right)$*, which again recovers the optimal rate.*

Note that if the GD (with iterates $\tilde{f}_t$) is employed, from previous work [52] we know that the *bias term* $\left(\frac{\lambda}{\eta t}\right)^{2r}$ is replaced by $\left(\frac{1}{\eta t}\right)^{2r}$, and therefore the upper bound translates to

$$\|S\tilde{f}_t - f^*\|^2_{L_2(P_X)} \leq C\left\{(\eta t)^{-2r} + \frac{1}{n}\left(\mathrm{tr}\left(\Sigma^{1/s}\right)(\eta t)^{1/s} + \frac{\eta t}{n}\right)\left(\sigma^2 + \left(\frac{1}{\eta t}\right)^{2r} + \frac{M^2 + (\eta t)^{-(2r-1)}}{n}\right)\right\},$$

with high probability. In other words, by the condition $\eta = O(1)$, we need $t = \Theta(n^{\frac{2rs}{2rs+1}})$ steps to sufficiently diminish the bias term. In contrast, the preconditioned update that interpolates between GD and NGD (3.1) only require $t = O(\log(n))$ steps to make the bias term negligible. This is because the NGD amplifies the high frequency component and rapidly captures the detailed "shape" of the target function f^*.

D.8.2. PROOF OF MAIN RESULT

Proof We follow the proof strategy of [52]. First we define a reference optimization problem with iterates $\bar{f}_t$ that directly minimize the population risk:

$$\bar{f}_t = \bar{f}_{t-1} - \eta(\Sigma + \lambda I)^{-1}(\Sigma\bar{f}_{t-1} - S^*f^*), \quad \bar{f}_0 = 0.$$

Note that $\mathbb{E}[f_t] = \bar{f}_t$. In addition, we define the degrees of freedom and its related quantity as

$$\mathcal{N}_\infty(\lambda) := \mathbb{E}_{\boldsymbol{x}}[\langle K_{\boldsymbol{x}}, \Sigma_\lambda^{-1} K_{\boldsymbol{x}}\rangle_{\mathcal{H}}] = \operatorname{tr}\big(\Sigma\Sigma_\lambda^{-1}\big), \quad \mathcal{F}_\infty(\lambda) := \sup_{\boldsymbol{x}\in\operatorname{supp}(P_X)} \|\Sigma_\lambda^{-1/2} K_{\boldsymbol{x}}\|_{\mathcal{H}}^2.$$

We can see that the risk admits the following bias-variance decomposition

$$\|Sf_t - f^*\|_{L_2(P_X)}^2 \le 2(\underbrace{\|Sf_t - S\bar{f}_t\|_{L_2(P_X)}^2}_{V(t),\ \text{variance}} + \underbrace{\|\bar{f}_t - f^*\|_{L_2(P_X)}^2}_{B(t),\ \text{bias}}).$$

We upper bound the bias and variance separately.

Bounding the bias term $B(t)$: Note that by the update rule (3.1), it holds that

$$\begin{aligned} & S\bar{f}_t - f^* = S\bar{f}_{t-1} - f^* - \eta S(\Sigma + \lambda I)^{-1}(\Sigma\bar{f}_{t-1} - S^* f^*) \\ \Leftrightarrow\ & S\bar{f}_t - f^* = (I - \eta S(\Sigma + \lambda I)^{-1} S^*)(S\bar{f}_{t-1} - f^*). \end{aligned}$$

Therefore, unrolling the recursion gives $S\bar{f}_t - f^* = (I - \eta S(\Sigma + \lambda I)^{-1} S^*)^t (S\bar{f}_0 - f^*) = (I - \eta S(\Sigma + \lambda I)^{-1} S^*)^t (-f^*) = -(I - \eta S(\Sigma + \lambda I)^{-1} S^*)^t L^r h^*$. Write the spectral decomposition of L as $L = \sum_{j=1}^\infty \sigma_j \phi_j \phi_j^*$ for $\phi_j \in L_2(P_X)$ for $\sigma_j \ge 0$. We have $\|(I - \eta S(\Sigma + \lambda I)^{-1} S^*)^t L^r h^*\|_{L_2(P_X)} = \sum_{j=1}^\infty (1 - \eta\frac{\sigma_j}{\sigma_j + \lambda})^{2t} \sigma_j^{2r} h_j^2$, where $h = \sum_{j=1}^\infty h_j \phi_j$. We then apply Lemma 9 to obtain

$$B(t) \le \exp(-\eta t) \sum_{j:\sigma_j \ge \lambda} h_j^2 + \left(\frac{2r}{e}\frac{\lambda}{\eta t}\right)^{2r} \sum_{j:\sigma_j < \lambda} h_j^2 \le C\left[\exp(-\eta t) \vee \left(\frac{\lambda}{\eta t}\right)^{2r}\right] \|h^*\|_{L_2(P_X)}^2,$$

where C is a constant depending only on r.

Bounding the variance term $V(t)$: We now handle the variance term $V(t)$. For notational convenience, we write $A_\lambda := A + \lambda I$ for a linear operator A from a Hilbert space H to H. By the definition of f_t, we know

$$\begin{aligned} f_t &= (I - \eta(\Sigma + \lambda I)^{-1}\hat{\Sigma}) f_{t-1} + \eta(\Sigma + \lambda I)^{-1} \hat{S}^* Y \\ &= \sum_{j=0}^{t-1} (I - \eta(\Sigma + \lambda I)^{-1}\hat{\Sigma})^j \eta (\Sigma + \lambda I)^{-1} \hat{S}^* Y \\ &= \Sigma_\lambda^{-1/2} \eta \left[\sum_{j=0}^{t-1} (I - \eta \Sigma_\lambda^{-1/2} \hat{\Sigma} \Sigma_\lambda^{-1/2})^j\right] \Sigma_\lambda^{-1/2} \hat{S}^* Y =: \Sigma_\lambda^{-1/2} G_t \Sigma_\lambda^{-1/2} \hat{S}^* Y, \end{aligned}$$

where we defined $G_t := \eta\left[\sum_{j=0}^{t-1} (I - \eta \Sigma_\lambda^{-1/2} \hat{\Sigma} \Sigma_\lambda^{-1/2})^j\right]$. Accordingly, we decompose $V(t)$ as

$$\begin{aligned} \|Sf_t - S\bar{f}_t\|_{L_2(P_X)}^2 \le & 2(\underbrace{\|S(f_t - \Sigma_\lambda^{-1/2} G_t \Sigma_\lambda^{-1/2} \hat{\Sigma} \bar{f}_t)\|_{L_2(P_X)}^2}_{(a)} \\ & + \underbrace{\|S(\Sigma_\lambda^{-1/2} G_t \Sigma_\lambda^{-1/2} \hat{\Sigma} \bar{f}_t - \bar{f}_t)\|_{L_2(P_X)}^2}_{(b)}). \end{aligned}$$

We bound (a) and (b) separately.

Step 1. Bounding (a)**.** Decompose (a) as

$$\begin{aligned}&\|S(f_t-\Sigma_\lambda^{-1/2}G_t\Sigma_\lambda^{-1/2}\hat{\Sigma}\bar{f}_t)\|_{L_2(P_X)}^2=\|S\Sigma_\lambda^{-1/2}G_t\Sigma_\lambda^{-1/2}(\hat{S}^*Y-\hat{\Sigma}\bar{f}_t)\|_{L_2(P_X)}^2\\ \le&\|S\Sigma_\lambda^{-1/2}\|^2\|G_t\Sigma_\lambda^{-1/2}\hat{\Sigma}_\lambda\Sigma_\lambda^{-1/2}\|^2\|\Sigma_\lambda^{1/2}\hat{\Sigma}_\lambda^{-1}\Sigma_\lambda^{1/2}\|^2\|\Sigma_\lambda^{-1/2}(\hat{S}^*Y-\hat{\Sigma}\bar{f}_t)\|_{\mathcal{H}}^2.\end{aligned}$$

We bound the terms in the RHS individually.

(i) $\|S\Sigma_\lambda^{-1/2}\|^2=\|\Sigma_\lambda^{-1/2}\Sigma\Sigma_\lambda^{-1/2}\|\le 1$.

(ii) Note that $\Sigma_\lambda^{-1/2}\hat{\Sigma}_\lambda\Sigma_\lambda^{-1/2}=I-\Sigma_\lambda^{-1/2}(\Sigma-\hat{\Sigma})\Sigma_\lambda^{-1/2}\succeq(1-\|\Sigma_\lambda^{-1/2}(\Sigma-\hat{\Sigma})\Sigma_\lambda^{-1/2}\|)I$. Proposition 6 of [71] and its proof implies that for $\lambda\le\|\Sigma\|$ and $0<\delta<1$, it holds that

$$\|\Sigma_\lambda^{-1/2}(\Sigma-\hat{\Sigma})\Sigma_\lambda^{-1/2}\|\le\sqrt{\frac{2\beta\mathcal{F}_\infty(\lambda)}{n}}+\frac{2\beta(1+\mathcal{F}_\infty(\lambda))}{3n}=:\Xi_n, \tag{D.21}$$

with probability $1-\delta$, where $\beta=\log\left(\frac{4\mathrm{tr}(\Sigma\Sigma_\lambda^{-1})}{\delta}\right)=\log\left(\frac{4\mathcal{N}_\infty(\lambda)}{\delta}\right)$. By Lemma 12, $\beta\le\log\left(\frac{4\mathrm{tr}(\Sigma^{1/s})\lambda^{-1/s}}{\delta}\right)$ and $\mathcal{F}_\infty(\lambda)\le\lambda^{-1}$. Therefore, if $\lambda=o(n^{-1}\log(n))$ and $\lambda=\Omega(n^{-1/s})$, the RHS can be smaller than $1/2$ for sufficiently large n, i.e. $\Xi_n=O(\sqrt{\log(n)/(n\lambda)})\le 1/2$. In this case we have,

$$\Sigma_\lambda^{-1/2}\hat{\Sigma}_\lambda\Sigma_\lambda^{-1/2}\succeq\frac{1}{2}I.$$

We denote this event as $\mathcal{E}_1$.

(iii) Note that

$$\begin{aligned}G_t\Sigma_\lambda^{-1/2}\hat{\Sigma}_\lambda\Sigma_\lambda^{-1/2}&=\eta\left[\sum_{j=0}^{t-1}(I-\eta\Sigma_\lambda^{-1/2}\hat{\Sigma}\Sigma_\lambda^{-1/2})^j\right]\Sigma_\lambda^{-1/2}\hat{\Sigma}_\lambda\Sigma_\lambda^{-1/2}\\&=\eta\left[\sum_{j=0}^{t-1}(I-\eta\Sigma_\lambda^{-1/2}\hat{\Sigma}\Sigma_\lambda^{-1/2})^j\right](\Sigma_\lambda^{-1/2}\hat{\Sigma}\Sigma_\lambda^{-1/2}+\lambda\Sigma_\lambda^{-1}).\end{aligned}$$

Thus, by Lemma 10 we have

$$\begin{aligned}&\|G_t\Sigma_\lambda^{-1/2}\hat{\Sigma}_\lambda\Sigma_\lambda^{-1/2}\|\\ \le&\underbrace{\left\|\eta\left[\sum_{j=0}^{t-1}(I-\eta\Sigma_\lambda^{-1/2}\hat{\Sigma}\Sigma_\lambda^{-1/2})^j\right]\Sigma_\lambda^{-1/2}\hat{\Sigma}\Sigma_\lambda^{-1/2}\right\|}_{\le 1\ \text{(due to Lemma 10)}}+\left\|\eta\left[\sum_{j=0}^{t-1}(I-\eta\Sigma_\lambda^{-1/2}\hat{\Sigma}\Sigma_\lambda^{-1/2})^j\right]\lambda\Sigma_\lambda^{-1}\right\|\\ \le&1+\eta\sum_{j=0}^{t-1}\|(I-\eta\Sigma_\lambda^{-1/2}\hat{\Sigma}\Sigma_\lambda^{-1/2})^j\|\|\lambda\Sigma_\lambda^{-1}\|\le 1+\eta t.\end{aligned}$$

(iv) Note that

$$\|\Sigma_\lambda^{-1/2}(\hat{S}^*Y-\hat{\Sigma}\bar{f}_t)\|_{\mathcal{H}}^2\le 2(\|\Sigma_\lambda^{-1/2}[(\hat{S}^*Y-\hat{\Sigma}\bar{f}_t)-(S^*f^*-\Sigma\bar{f}_t)]\|_{\mathcal{H}}^2+\|\Sigma_\lambda^{-1/2}(S^*f^*-\Sigma\bar{f}_t)\|_{\mathcal{H}}^2).$$

First we bound the first term of the RHS. Let $\xi_i = \Sigma_\lambda^{-1/2}[K_{\boldsymbol{x}_i} y_i - K_{\boldsymbol{x}_i}\bar{f}_t(\boldsymbol{x}_i) - (S^* f^* - \Sigma\bar{f}_t)]$. Then, $\{\xi_i\}_{i=1}^n$ is an i.i.d. sequence of zero-centered random variables taking value in $\mathcal{H}$ and thus we have

$$\|\Sigma_\lambda^{-1/2}[(\hat{S}^* Y - \hat{\Sigma}\bar{f}_t) - (S^* f^* - \Sigma\bar{f}_t)]\|_{\mathcal{H}}^2 = \left\|\frac{1}{n}\sum_{i=1}^n \xi_i\right\|_{\mathcal{H}}^2.$$

The RHS can be bounded by using Bernstein's inequality in Hilbert space [19]. To apply the inequality, we need to bound the variance and sup-norm of the random variable. The variance can be bounded as

$$\begin{aligned}
\mathbb{E}[\|\xi_i\|_{\mathcal{H}}^2] &\le \mathbb{E}_{(\boldsymbol{x},y)}\left[\|\Sigma_\lambda^{-1/2}(K_{\boldsymbol{x}}(f^*(\boldsymbol{x}) - \bar{f}_t(\boldsymbol{x})) + K_\xi \epsilon)\|_{\mathcal{H}}^2\right] \\
&\le 2\Big\{\mathbb{E}_{(\boldsymbol{x},y)}\left[\|\Sigma_\lambda^{-1/2}(K_{\boldsymbol{x}}(f^*(\boldsymbol{x}) - \bar{f}_t(x))\|_{\mathcal{H}}^2 + \|\Sigma_\lambda^{-1/2}(K_{\boldsymbol{x}}\epsilon)\|_{\mathcal{H}}^2\right]\Big\} \\
&\le 2\left\{\sup_{\boldsymbol{x}\in\mathrm{supp}(P_X)} \|\Sigma_\lambda^{-1/2}K_{\boldsymbol{x}}\|^2 \|f^* - S\bar{f}_t\|_{L_2(P_X)}^2 + \sigma^2 \mathrm{tr}\big(\Sigma_\lambda^{-1}\Sigma\big)\right\} \\
&\le 2\big\{\mathcal{F}_\infty(\lambda)B(t) + \sigma^2\mathrm{tr}\big(\Sigma_\lambda^{-1}\Sigma\big)\big\} \\
&\le 2\big\{\lambda^{-1}B(t) + \sigma^2\mathrm{tr}\big(\Sigma_\lambda^{-1}\Sigma\big)\big\},
\end{aligned}$$

The sup-norm can be bounded as follows. Observe that $\|\bar{f}_t\|_\infty \le \|\bar{f}_t\|_{\mathcal{H}}$, and thus by Lemma 11,

$$\begin{aligned}
\|\xi_i\|_{\mathcal{H}} &\le 2 \sup_{\boldsymbol{x}\in\mathrm{supp}(P_X)} \|\Sigma_\lambda^{-1/2}K_{\boldsymbol{x}}\|_{\mathcal{H}}(\sigma + \|f^*\|_\infty + \|\bar{f}_t\|_\infty) \\
&\lesssim \mathcal{F}_\infty^{1/2}(\lambda)(\sigma + M + (1+t\eta)\lambda^{-(1/2-r)_+}) \\
&\lesssim \lambda^{-1/2}(\sigma + M + (1+t\eta)\lambda^{-(1/2-r)_+}).
\end{aligned}$$

Therefore, for $0<\delta<1$, Bernstein's inequality (see Proposition 2 of [19]) yields that

$$\left\|\frac{1}{n}\sum_{i=1}^n \xi_i\right\|_{\mathcal{H}}^2 \le C\left(\sqrt{\frac{\lambda^{-1}B(t) + \sigma^2\mathrm{tr}\big(\Sigma_\lambda^{-1}\Sigma\big)}{n}} + \frac{\lambda^{-1/2}(\sigma + M + (1+t\eta)\lambda^{-(1/2-r)_+})}{n}\right)^2 \log(1/\delta)^2$$

with probability $1-\delta$ where C is a universal constant. We define this event as $\mathcal{E}_2$.

For the second term $\|\Sigma_\lambda^{-1/2}(S^* f^* - \Sigma\bar{f}_t)\|_{\mathcal{H}}^2$ we have

$$\|\Sigma_\lambda^{-1/2}(S^* f^* - \Sigma\bar{f}_t)\|_{\mathcal{H}}^2 \le \|\Sigma_\lambda^{1/2}(f^* - Sf_t)\|_{\mathcal{H}}^2 = \|f^* - S\bar{f}_t\|_{L_2(P_X)}^2 \le B(t).$$

Combining these evaluations, on the event $\mathcal{E}_2$ where $P(\mathcal{E}_2) \ge 1-\delta$ for $0<\delta<1$ we have

$$\begin{aligned}
&\|\Sigma_\lambda^{-1/2}(\hat{S}^* Y - \hat{\Sigma}\bar{f}_t)\|_{\mathcal{H}}^2 \\
&\overset{(i)}{\le} C\left(\sqrt{\frac{\lambda^{-1}B(t) + \sigma^2\mathrm{tr}\big(\Sigma_\lambda^{-1}\Sigma\big)}{n}} + \frac{\lambda^{-1/2}(\sigma + M + (1+t\eta)\lambda^{-(1/2-r)_+})}{n}\right)^2 \log(1/\delta)^2 + B(t).
\end{aligned}$$

where we used Lemma 12 in (i).

Step 2. Bounding (b)**.** On the event $\mathcal{E}_1$, the term (b) can be evaluated as

$$\begin{aligned}
&\|S(\Sigma_\lambda^{-1/2}G_t\Sigma_\lambda^{-1/2}\hat{\Sigma}\bar{f}_t - \bar{f}_t)\|_{L_2(P_X)}^2\\
\le&\|\Sigma^{1/2}(\Sigma_\lambda^{-1/2}G_t\Sigma_\lambda^{-1/2}\hat{\Sigma}\bar{f}_t - \bar{f}_t)\|_{\mathcal{H}}^2\\
\le&\|\Sigma^{1/2}\Sigma_\lambda^{-1/2}(G_t\Sigma_\lambda^{-1/2}\hat{\Sigma}\Sigma_\lambda^{-1/2} - I)\Sigma_\lambda^{1/2}\bar{f}_t\|_{\mathcal{H}}^2\\
\le&\|\Sigma^{1/2}\Sigma_\lambda^{-1/2}\|\|(G_t\Sigma_\lambda^{-1/2}\hat{\Sigma}\Sigma_\lambda^{-1/2} - I)\Sigma_\lambda^{1/2}\bar{f}_t\|_{\mathcal{H}}^2\\
\le&\|(G_t\Sigma_\lambda^{-1/2}\hat{\Sigma}\Sigma_\lambda^{-1/2} - I)\Sigma_\lambda^{1/2}\bar{f}_t\|_{\mathcal{H}}^2. \qquad \text{(D.22)}
\end{aligned}$$

where we used Lemma 11 in the last inequality. The term $\|(G_t\Sigma_\lambda^{-1/2}\hat{\Sigma}\Sigma_\lambda^{-1/2} - I)\Sigma_\lambda^{1/2}f_t\|_{\mathcal{H}}$ can be bounded as follows. First, note that

$$\begin{aligned}
(G_t\Sigma_\lambda^{-1/2}\hat{\Sigma}\Sigma_\lambda^{-1/2} - I)\Sigma_\lambda^{1/2} &= \left\{\eta\left[\sum_{j=0}^{t-1}(I-\eta\Sigma_\lambda^{-1/2}\hat{\Sigma}\Sigma_\lambda^{-1/2})^j\right]\Sigma_\lambda^{-1/2}\hat{\Sigma}\Sigma_\lambda^{-1/2} - I\right\}\Sigma_\lambda^{1/2}\\
&= (I-\eta\Sigma_\lambda^{-1/2}\hat{\Sigma}\Sigma_\lambda^{-1/2})^t\Sigma_\lambda^{1/2}.
\end{aligned}$$

Therefore, the RHS of (D.22) can be further bounded by

$$\begin{aligned}
&\|(I-\eta\Sigma_\lambda^{-1/2}\hat{\Sigma}\Sigma_\lambda^{-1/2})^t\Sigma_\lambda^{1/2}\bar{f}_t\|_{\mathcal{H}}\\
=&\|(I-\eta\Sigma_\lambda^{-1/2}\Sigma\Sigma_\lambda^{-1/2} + \eta\Sigma_\lambda^{-1/2}(\Sigma-\hat{\Sigma})\Sigma_\lambda^{-1/2})^t\Sigma_\lambda^{1/2}\bar{f}_t\|_{\mathcal{H}}\\
=&\|\sum_{k=0}^{t-1}(I-\eta\Sigma_\lambda^{-1/2}\hat{\Sigma}\Sigma_\lambda^{-1/2})^k(\eta\Sigma_\lambda^{-1/2}(\Sigma-\hat{\Sigma})\Sigma_\lambda^{-1/2})(I-\eta\Sigma_\lambda^{-1}\Sigma)^{t-k-1}\Sigma_\lambda^{1/2}\bar{f}_t - (I-\eta\Sigma_\lambda^{-1}\Sigma)^t\Sigma_\lambda^{1/2}\bar{f}_t\|_{\mathcal{H}}\\
\overset{(i)}{\le}&\|(I-\eta\Sigma_\lambda^{-1}\Sigma)^t\Sigma_\lambda^{1/2}\bar{f}_t\|_{\mathcal{H}}\\
&+\eta\sum_{k=0}^{t-1}\|(I-\eta\Sigma_\lambda^{-1/2}\hat{\Sigma}\Sigma_\lambda^{-1/2})^k\Sigma_\lambda^{-1/2}(\Sigma-\hat{\Sigma})\Sigma_\lambda^{-1/2+r}(I-\eta\Sigma_\lambda^{-1}\Sigma)^{t-k-1}\Sigma_\lambda^{1/2-r}\bar{f}_t\|_{\mathcal{H}}\\
\le&\|(I-\eta\Sigma_\lambda^{-1}\Sigma)^t\Sigma_\lambda^{1/2}\bar{f}_t\|_{\mathcal{H}} + t\eta\|\Sigma_\lambda^{-1/2}(\Sigma-\hat{\Sigma})\Sigma_\lambda^{-1/2+r}\|\|\Sigma_\lambda^{1/2-r}\bar{f}_t\|_{\mathcal{H}}\\
=&\|(I-\eta\Sigma_\lambda^{-1}\Sigma)^t\Sigma_\lambda^r\|\|\Sigma_\lambda^{1/2-r}\bar{f}_t\|_{\mathcal{H}} + t\eta\|\Sigma_\lambda^{-1/2}(\Sigma-\hat{\Sigma})\Sigma_\lambda^{-1/2+r}\|\|\Sigma_\lambda^{1/2-r}\bar{f}_t\|_{\mathcal{H}}\\
\lesssim&\|(I-\eta\Sigma_\lambda^{-1}\Sigma)^t\Sigma_\lambda^r\| + t\eta\|\Sigma_\lambda^{-1/2}(\Sigma-\hat{\Sigma})\Sigma_\lambda^{-1/2+r}\|(1+t\eta)\|h^*\|_{L_2(P_X)}, \qquad \text{(D.23)}
\end{aligned}$$

where (i) is due to exchangeability of Σ_λ and Σ. By Lemma 9, for the RHS we have

$$\|(I-\eta\Sigma_\lambda^{-1}\Sigma)^t\Sigma_\lambda^r\| \le \exp(-\eta t/2) \vee \left(\frac{1}{e}\frac{\lambda}{\eta t}\right)^r.$$

Next, as in the (D.21), by applying the Bernstein inequality for asymmetric operators (Corollary 3.1 of [61] with the argument in its Section 3.2), it holds that

$$\begin{aligned}
&\|\Sigma_\lambda^{-1/2}(\Sigma-\hat{\Sigma})\Sigma_\lambda^{-1/2+r}\|\\
\le&C'\left(\sqrt{\frac{\beta' C_\mu^2(2^{2r-\mu}\vee\lambda^{2r-\mu})\mathcal{N}_\infty(\lambda)}{n}} + \frac{\beta'((1+\lambda)^r + C_\mu^2\lambda^{-\mu/2}(2^{2r-\mu}\vee\lambda^{r-\mu/2})}{n}\right) =: \Xi_n',
\end{aligned}$$

with probability $1-\delta$, where C' is a universal constant and $\beta' \le \log\left(\frac{28C_\mu^2(2^{2r-\mu}\vee\lambda^{-\mu+2r})\mathrm{tr}\left(\Sigma^{1/s}\right)\lambda^{-1/s}}{\delta}\right)$. We also used the following bounds on the sup-norm and the second order moments:

$$\begin{aligned}
(\text{sup-norm})\quad & \|\Sigma_\lambda^{-1/2}(K_{\boldsymbol{x}}K_{\boldsymbol{x}}^* - \Sigma)\Sigma_\lambda^{-1/2+r}\| \\
& \le \|\Sigma_\lambda^{-1/2}K_{\boldsymbol{x}}K_{\boldsymbol{x}}^*\Sigma_\lambda^{-1/2+r}\| + \|\Sigma_\lambda^r\| \\
& \le \|\Sigma_\lambda^{-\mu/2}\Sigma_\lambda^{\mu/2-1/2}K_{\boldsymbol{x}}K_{\boldsymbol{x}}^*\Sigma_\lambda^{-1/2+\mu/2}\Sigma_\lambda^{r-\mu/2}\| + \|\Sigma_\lambda^r\| \\
& \le C_\mu^2\lambda^{-\mu/2}(2^{r-\mu/2}\vee\lambda^{r-\mu/2}) + (1+\lambda)^r \quad (\text{a.s.}), \\
(\text{2nd order moment 1})\quad & \|\mathbb{E}_{\boldsymbol{x}}[\Sigma_\lambda^{-1/2}(K_{\boldsymbol{x}}K_{\boldsymbol{x}}^* - \Sigma)\Sigma_\lambda^{-1+2r}(K_{\boldsymbol{x}}K_{\boldsymbol{x}}^* - \Sigma)\Sigma_\lambda^{-1/2}]\| \\
& \le \|\Sigma_\lambda^{-1/2}\Sigma\Sigma_\lambda^{-1/2}\| \sup_{\boldsymbol{x}\in\mathrm{supp}(P_X)}[K_{\boldsymbol{x}}^*\Sigma_\lambda^{-1/2+\mu/2}\Sigma_\lambda^{-\mu+2r}\Sigma_\lambda^{-1/2+\mu/2}K_{\boldsymbol{x}}] \\
& \le C_\mu^2(2^{2r-\mu}\vee\lambda^{2r-\mu}), \\
(\text{2nd order moment 2})\quad & \|\mathbb{E}_{\boldsymbol{x}}[\Sigma_\lambda^{-1/2+r}(K_{\boldsymbol{x}}K_{\boldsymbol{x}}^* - \Sigma)\Sigma_\lambda^{-1/2}\Sigma_\lambda^{-1/2}(K_{\boldsymbol{x}}K_{\boldsymbol{x}}^* - \Sigma)\Sigma_\lambda^{-1/2+r}]\| \\
& \le \|\mathbb{E}_{\boldsymbol{x}}[\Sigma_\lambda^{-1/2+r}K_{\boldsymbol{x}}K_{\boldsymbol{x}}^*\Sigma_\lambda^{-1}K_{\boldsymbol{x}}K_{\boldsymbol{x}}^*\Sigma_\lambda^{-1/2+r}]\| \\
& \le C_\mu^2(2^{2r-\mu}\vee\lambda^{2r-\mu})\mathbb{E}_{\boldsymbol{x}}[K_{\boldsymbol{x}}^*\Sigma_\lambda^{-1}K_{\boldsymbol{x}}] \\
& = C_\mu^2(2^{2r-\mu}\vee\lambda^{2r-\mu})\mathrm{tr}\left(\Sigma\Sigma_\lambda^{-1}\right) \\
& = C_\mu^2(2^{2r-\mu}\vee\lambda^{2r-\mu})\mathcal{N}_\infty(\lambda).
\end{aligned}$$

We define this event as $\mathcal{E}_3$. Therefore, the RHS of (D.23) can be further bounded by

$$\begin{aligned}
&[\|(I-\eta\Sigma_\lambda^{-1}\Sigma)^t\Sigma_\lambda^r\| + Ct\eta\|\Sigma_\lambda^{-1/2}(\Sigma-\hat{\Sigma})\Sigma_\lambda^{-1/2+r}\|](1+t\eta)\|h^*\|_{L_2(P_X)} \\
\le&\left[\exp(-\eta t/2)\vee\left(\frac{1}{e}\frac{\lambda}{\eta t}\right)^r + t\eta\Xi_n'\right](1+t\eta)\|h^*\|_{L_2(P_X)}.
\end{aligned}$$

Finally, note that when $\lambda=\lambda^*$ and $2r\ge\mu$,

$$\Xi_n'^2 = \tilde{O}\left(\frac{\lambda^{*2r-\mu-1/s}}{n} + \frac{\lambda^{*2(r-\mu)}}{n^2}\right) \le \tilde{O}(n^{-\frac{s(4r-\mu)}{2rs+1}} + n^{-\frac{s(4r-2\mu)+2}{2rs+1}}) \le \tilde{O}(n^{-\frac{2rs}{2rs+1}}).$$

Step 3. Combining the calculations in Step 1 and 2 leads to the desired result. ■

D.8.3. AUXILIARY LEMMAS

Lemma 9 *For $t\in\mathbb{N}$, $0<\eta<1$, $0<\sigma\le1$ and $0\le\lambda$, it holds that*

$$\left(1-\eta\frac{\sigma}{\sigma+\lambda}\right)^t\sigma^r \le \begin{cases}\exp(-\eta t/2) & (\sigma\ge\lambda)\\ \left(\frac{2r}{e}\frac{\lambda}{\eta t}\right)^r & (\sigma<\lambda)\end{cases}.$$

Proof When $\sigma\ge\lambda$, we have

$$\left(1-\eta\frac{\sigma}{\sigma+\lambda}\right)^t\sigma^r \le \left(1-\eta\frac{\sigma}{2\sigma}\right)^t\sigma^r = (1-\eta/2)^t\sigma^r \le \exp(-t\eta/2)\sigma^r \le \exp(-t\eta/2)$$

due to $\sigma \leq 1$. On the other hand, note that

$$\begin{aligned}\left(1-\eta\frac{\sigma}{\sigma+\lambda}\right)^t\sigma^r &\leq \exp\left(-\eta t\frac{\sigma}{\sigma+\lambda}\right)\times\left(\frac{\sigma\eta t}{\sigma+\lambda}\right)^r\left(\frac{\sigma+\lambda}{\eta t}\right)^r\\ &\leq \sup_{x>0}\exp(-x)x^r\left(\frac{\sigma+\lambda}{\eta t}\right)^r \leq \left(\frac{(\sigma+\lambda)r}{\eta te}\right)^r,\end{aligned}$$

where we used $\sup_{x>0}\exp(-x)x^r = (r/e)^r$. ■

Lemma 10 *For $t = \mathbb{N}$, $0 < \eta$ and $0 \leq \sigma$ such that $\eta\sigma < 1$, it holds that $\eta\sum_{j=0}^{t-1}(1-\eta\sigma)^j\sigma \leq 1$.*

Proof If $\sigma = 0$, then the statement is obvious. Assume that $\sigma > 0$, then

$$\sum_{j=0}^{t-1}(1-\eta\sigma)^j\sigma = \frac{1-(1-\eta\sigma)^t}{1-(1-\eta\sigma)}\sigma = \frac{1}{\eta}[1-(1-\eta\sigma)^t] \leq \eta^{-1}.$$

This yields the desired claim. ■

Lemma 11 *Under (A5-7), for any $0 < \lambda < 1$ and $q \leq r$, it holds that*

$$\|\Sigma_\lambda^{-s}\bar{f}_t\|_{\mathcal{H}} \lesssim (1+\lambda^{-(1/2+(q-r))_+}+\lambda t\eta\lambda^{-(3/2+(q-r))_+})\|h^*\|_{L_2(P_X)}.$$

Proof Recall that

$$\bar{f}_t = (I-\eta(\Sigma+\lambda I)^{-1}\Sigma)\bar{f}_{t-1}+\eta(\Sigma+\lambda I)^{-1}S^*f^* = \sum_{j=0}^{t-1}(I-\eta(\Sigma+\lambda I)^{-1}\Sigma)^j\eta(\Sigma+\lambda I)^{-1}S^*f^*.$$

Therefore, we obtain the following

$$\begin{aligned}&\|\Sigma_\lambda^{-q}\bar{f}_t\|_{\mathcal{H}} = \eta\|\sum_{j=0}^{t-1}(I-\eta\Sigma_\lambda^{-1}\Sigma)^j\Sigma_\lambda^{-1-q}S^*L^rh^*\|_{\mathcal{H}}\\
=&\eta\|\sum_{j=0}^{t-1}(I-\eta\Sigma_\lambda^{-1}\Sigma)^j\Sigma_\lambda^{-1}(\Sigma+\lambda I)\Sigma_\lambda^{-q-1}S^*L^rh^*\|_{\mathcal{H}}\\
\leq&\eta\|\sum_{j=0}^{t-1}(I-\eta\Sigma_\lambda^{-1}\Sigma)^j\Sigma_\lambda^{-1}\Sigma\Sigma_\lambda^{-q-1}S^*L^rh^*\|_{\mathcal{H}}+\lambda\eta\|\sum_{j=0}^{t-1}(I-\eta\Sigma_\lambda^{-1}\Sigma)^j\Sigma_\lambda^{-1}\Sigma_\lambda^{-q-1}S^*L^rh^*\|_{\mathcal{H}}\\
\leq&\eta\|\sum_{j=0}^{t-1}(I-\eta\Sigma_\lambda^{-1}\Sigma)^j\Sigma_\lambda^{-1}\Sigma\|\|\Sigma_\lambda^{-q-1}S^*L^rh^*\|_{\mathcal{H}}+\lambda\eta\|\sum_{j=0}^{t-1}(I-\eta\Sigma_\lambda^{-1}\Sigma)^j\Sigma_\lambda^{-1}\Sigma_\lambda^{-q-1}S^*L^rh^*\|_{\mathcal{H}}\\
\leq&\|\Sigma_\lambda^{-q-1}S^*L^rh^*\|_{\mathcal{H}}+\lambda t\eta\|\Sigma_\lambda^{-1}\Sigma_\lambda^{-q-1}S^*L^rh^*\|_{\mathcal{H}}\\
\leq&\|S^*L_\lambda^{-q-1+r}h^*\|_{\mathcal{H}}+\lambda t\eta\|S^*L_\lambda^{-q-2+r}h^*\|_{\mathcal{H}}\\
\leq&\sqrt{\langle h^*, L_\lambda^{-q-1+r}SS^*L_\lambda^{-q-1+r}h^*\rangle_{L_2(P_X)}}+\lambda t\eta\sqrt{\langle h^*, L_\lambda^{-q-2+r}SS^*L_\lambda^{-q-2+r}h^*\rangle_{L_2(P_X)}}\end{aligned}$$

$$
\begin{aligned}
=&\sqrt{\langle h^*, L_\lambda^{-q-1+r} L L_\lambda^{-q-1+r} h^*\rangle_{L_2(P_X)}} + \lambda t\eta \sqrt{\langle h^*, L_\lambda^{-q-2+r} L L_\lambda^{-q-2+r} h^*\rangle_{L_2(P_X)}} \\
\leq&(\lambda^{-1/2-(q-r)} + \lambda t\eta\lambda^{-3/2-(q-r)})\|h^*\|_{L_2(P_X)} \leq (1+t\eta)\lambda^{-1/2-(q-r)}\|h^*\|_{L_2(P_X)}.
\end{aligned}
$$

■

Lemma 12 *Under (A5-7) and for* $\lambda \in (0,1)$*, it holds that* $\mathcal{N}_\infty(\lambda) \leq \operatorname{tr}(\Sigma^{1/s})\lambda^{-1/s}$*, and* $\mathcal{F}_\infty(\lambda) \leq 1/\lambda$*.*

Proof For the first inequality, we have

$$
\begin{aligned}
\mathcal{N}_\infty(\lambda) =&\operatorname{tr}(\Sigma\Sigma_\lambda^{-1}) = \operatorname{tr}\Big(\Sigma^{1/s}\Sigma^{1-1/s}\Sigma_\lambda^{-(1-1/s)}\Sigma_\lambda^{-1/s}\Big) \\
\leq&\operatorname{tr}\Big(\Sigma^{1/s}\Sigma^{1-1/s}\Sigma_\lambda^{-(1-1/s)}\Big)\lambda^{-1/s} \leq \operatorname{tr}\Big(\Sigma^{1/s}\Big)\lambda^{-1/s}.
\end{aligned}
$$

As for the second inequality, note that

$$
\mathcal{F}_\infty(\lambda) = \sup_{\boldsymbol{x}}\langle K_{\boldsymbol{x}}, \Sigma_\lambda^{-1}K_{\boldsymbol{x}}\rangle_{\mathcal{H}} \leq \sup_{\boldsymbol{x}} \lambda^{-1}\langle K_{\boldsymbol{x}}, K_{\boldsymbol{x}}\rangle_{\mathcal{H}} \leq \lambda^{-1}\sup_{\boldsymbol{x}} k(\boldsymbol{x},\boldsymbol{x}) \leq \lambda^{-1}.
$$

■

Appendix E. Experiment Setup

E.1. Processing the Datasets

To obtain extra unlabeled data to estimate the Fisher, we zero pad pixels on the boarders of each image before randomly cropping; a random horizontal flip is also applied for CIFAR10 images. We preprocess all images by dividing pixel values by 255 before centering them to be located within $[-0.5, 0.5]$ with the subtraction by $1/2$. For experiments on CIFAR10, we downsample the original images using a max pooling layer with kernel size 2 and stride 2.

E.2. Setup and Implementation for Optimizers

In all settings, GD uses a learning rate of 0.01 that is exponentially decayed every 1k updates with the parameter value 0.999. For NGD, we use a fixed learning rate of 0.03. Since inverting a parameter-by-parameter-sized Fisher estimate per iteration would be costly, we adopt the Hessian free approach [56] which computes approximate matrix-inverse-vector products using the conjugate gradient (CG) method [16, 64]. For each approximate inversion, we run CG for 200 iterations starting from the solution returned by the previous CG run. The precise number of CG iterations and the initialization heuristic roughly follow [59]. For the first run of CG, we initialize the vector from a standard Gaussian, and run CG for 5k iterations. To ensure invertibility, we apply a very small amount of damping (0.00001) in most scenarios.

E.3. Other Details

For experiments in the label noise and misspecification sections, we pretrain the teacher using the Adam optimizer [44] with its default hyperparameters and a learning rate of 0.001.

For experiments in the misalignment section, we downsample all images twice using max pooling with kernel size 2 and stride 2. Moreover, only for experiments in this section, we implement natural gradient descent by exactly computing the Fisher on a large batch of unlabeled data and inverting the matrix by calling PyTorch's `torch.inverse` before right multiplying the gradient.