On-Average Stability of Multipass Preconditioned SGD and Effective Dimension

Simon Vary SIMON.VARY@STATS.OX.AC.UK
Department of Statistics, University of Oxford

Tyler Farghly FARGHLY@STATS.OX.AC.UK
Department of Statistics, University of Oxford

Ilja Kuzborskij ILJAK@GOOGLE.COM
Google DeepMind

Patrick Rebeschini PATRICK.REBESCHINI@STATS.OX.AC.UK
Department of Statistics, University of Oxford

Editors: Steve Hanneke and Tor Lattimore

Abstract

We study trade-offs between the population risk curvature, geometry of the noise, and preconditioning on the generalisation ability of the multipass Preconditioned Stochastic Gradient Descent (PSGD). Many practical optimisation heuristics implicitly navigate this trade-off in different ways — for instance, some aim to whiten gradient noise, while others aim to align updates with expected loss curvature. When the geometry of the population risk curvature and the geometry of the gradient noise do not match, an aggressive choice that improves one aspect can amplify instability along the other, leading to suboptimal statistical behavior. In this paper we employ *on-average algorithmic stability* to connect generalisation of PSGD to the *effective dimension* that depends on these sources of curvature. While existing techniques for on-average stability of SGD are limited to a single pass, as first contribution we develop a new on-average stability analysis for multipass SGD that handles the correlations induced by data reuse. This allows us to derive excess risk bounds that depend on the effective dimension. In particular, we show that an improperly chosen preconditioner can yield suboptimal effective dimension dependence in both optimisation and generalisation. Finally, we complement our upper bounds with matching, instance-dependent lower bounds.

Keywords: Algorithmic stability, generalization bounds, preconditioning

1. Introduction

Training of machine learning models is usually posed as a minimisation of the population risk. In particular, given a data distribution Q supported on the example space $\mathcal{Z}$, the goal is to minimise the *population risk* f. The population risk and its empirical counterpart are defined as,

$$f(x) = \mathbb{E}_{z\sim \mathrm{Q}}[\ell(x,z)]\,, \qquad f_S(x) = \frac{1}{n}\sum_{i=1}^{n}\ell(x,z_i)\,,$$

respectively, where ℓ is a smooth loss function parameterized by x and evaluated on an example z. In the standard setting, where the data distribution is unknown and we have access only to a finite training set $S = \{z_1, \ldots, z_n\} \subset \mathcal{Z}$ of n-samples drawn i.i.d. from Q, we instead minimise

the *empirical risk* f_S. Given the solution $\hat{x}$ returned by an algorithm, its generalization ability is captured by the *excess risk*,

$$\mathbb{E}[\delta f(\hat{x})] \qquad \text{where} \qquad \delta f(x) = f(x) - \inf_{x \in \mathcal{X}} f(x) \, .$$

In this work, we focus on preconditioned SGD, meaning that empirical risk is minimised iteratively by observing gradients on individual examples drawn uniformly from the training set:

$$x_{t+1} = x_t - \eta_t \, P \, \nabla \ell(x_t, z_{i_t}) \, , \quad i_t \sim \text{Unif}(\{1, \dots, n\}) \, , \quad t = 0, 1, 2, \dots \tag{1}$$

where η_t is the step size, P is a Positive Definite (PD) *preconditioning* matrix and $z_{i_t} \in S$ are sampled randomly from S uniformly with replacement. Note that the update in eq. (1) is often not limited to a single pass over the training set. Hence, in the present work, we consider Preconditioned Stochastic Gradient Descent (PSGD) in the *multipass* regime. Since gradients are random variables, this randomised procedure is inevitably affected by the noise in stochastic gradients. In this work we pay close attention to the geometry of gradient covariance considering the *gradient covariance matrix* $\Sigma \succeq \text{Var}_z(\nabla \ell(x, z))$.[1]

At this point, we highlight that the learning problem is governed by three sources of curvature: the Hessian of the population risk $\nabla^2 f \equiv \nabla^2 f(\hat{x})$ for some minimiser $\hat{x}$, the gradient covariance matrix Σ, and the preconditioner P which is chosen by the practitioner. The goal of this paper is to understand, in the *finite-sample nonasymptotic setting*, how the excess risk of PSGD depends on the interaction between $\nabla^2 f$, Σ, and P. While, in the idealised scenario, Σ and $\nabla^2 f$ coincide (Amari, 1998), in the general *misspecified* learning setting where $\Sigma \neq \nabla^2 f$, the disparity creates a fundamental trade-off. This trade-off is addressed in practice in different ways by different optimisation algorithms. Methods like Adam (Kingma and Ba, 2015) and K-FAC (Martens and Grosse, 2015) target an approximate conditioning based on the uncentered second moments of the gradients, while others, such as AdaHessian (Yao et al., 2021), PROMISE (Frangella et al., 2024a), SAPPHIRE (Sun et al., 2025), SketchySGD (Frangella et al., 2024b), target the inverse of the *expected Hessian* $\nabla^2 f$. Thus, without a characterisation of the statistical properties associated with the mismatch between these geometries, the choice of preconditioner in the misspecified regime remains largely heuristic, which can lead to undesired behaviour (see Figure 1 for a graphical example). From a non-asymptotic statistical perspective, here we ask *what is the optimal choice of P with respect to $\nabla^2 f$ and Σ?*

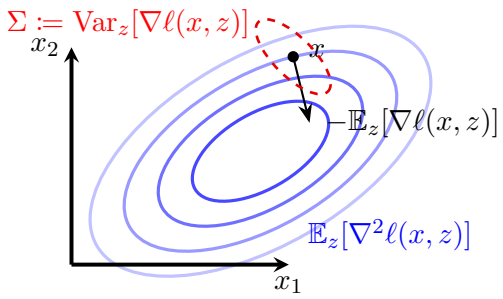

Figure 1: Illustration of model misspecification. The geometry of the expected loss curvature $\nabla^2 f$ differs from the geometry of the gradient noise (Σ). While setting $P \approx \Sigma^{-1}$ whitens the noise, it may result in unstable updates along high-curvature directions.

In this paper we are primarily interested in how excess risk $\mathbb{E}[\delta f(x_t)]$ depends on *effective dimension*

$$\text{tr}\left((\nabla^2 f)^{-1} \Sigma\right) \tag{2}$$

1. Derivatives are always taken with respect to the first argument, unless stated otherwise.

which commonly appears in statistics as a replacement for the ambient dimension. This is also known as the Takeuchi Information Criterion (TIC) in the context of information theory (Shibata, 1989). For example, the effective dimension controls excess risk bounds of linear (ridge) regression, for exact minimisers (Bach, 2024), Stochastic Gradient Descent (SGD) with iterative averaging (Neu and Rosasco, 2018), as well as asymptotic analysis in stochastic approximation (Polyak and Juditsky, 1992). While it is known that dependence of the excess risk on (2) is not improvable asymptotically, we ask here how P interacts with effective dimension in the non-asymptotic regime.

In particular, we will study this question through the lens of *generalisation error* $x \mapsto f(x) - f_S(x)$ and *algorithmic stability*, which is a classical framework dating back to the study of nearest-neighbor rules (Devroye and Wagner, 1979) and Empirical Risk Minimization (ERM) problems (Bousquet and Elisseeff, 2002). The stability approach asks whether the solution produced by the learning algorithm is insensitive to small perturbations in the training set, such as the removal of a data point or its replacement by an independent copy. Namely, if $\hat{x}^{(i)}$ is a parameter produced with such a perturbation (say when z_i is replaced by its independent copy z_i'), then the expected generalisation error is directly linked to stability gauged by the difference of losses:

$$\mathbb{E}[f(\hat{x}) - f_S(\hat{x})] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{S,z_i'}\left[\ell(\hat{x}^{(i)}, z_i) - \ell(\hat{x}, z_i)\right] . \tag{3}$$

Numerous works (Feldman and Vondrak, 2019; Bousquet et al., 2020) establish high-probability bounds on the generalisation error by controlling the *uniform* stability $\sup_{S,z,i} |\ell(\hat{x}, z) - \ell(\hat{x}^{(i)}, z)|$ for various ERM formulations. However, such notions of stability tend towards covering the *worst-case* and are not suitable to achieving our goal, since $(\nabla^2 f, \Sigma)$ are distribution-dependent quantities. Here we turn our attention to the weaker notion of *on-average* stability $\max_i \mathbb{E}_{S,z_i'}[\ell(\hat{x}^{(i)}, z_i) - \ell(\hat{x}, z_i)]$ which has, so far, largely been used to study ERM algorithms instead of the SGD-type algorithms of interest in this work (Kearns and Ron, 1997; Bousquet and Elisseeff, 2002; Elisseeff et al., 2005).

Algorithmic stability of SGD-type algorithms has been studied extensively over recent years. A seminal paper from Hardt et al. (2016) derived uniform stability bounds for simultaneously smooth Lipschitz and convex loss functions. These proof techniques were later extended by Kuzborskij and Lampert (2018); Lei (2023) to the on-average stability, observing that generalisation error can be controlled by the data dependent quantities (such as the empirical risk), leading to optimistic bounds. However, none of these works showed dependence on the effective dimension or preconditioner P. By targeting a data-dependent analysis of PSGD, one runs into several difficulties. Most commonly known, is the difficulty of managing dependence between parameter iterates and the dataset, which is usually circumvented by restricting to the single pass setting. In the present work, we consider the multi-pass setting and we develop methods to manage parameter-dataset dependence.

1.1. Our contributions

1. We develop an *on-average stability* analysis of *multipass* SGD that overcomes the technical challenge of dependence between iterates arising through reused data points — see Section 2 for the sketch of the analysis.

2. We derive excess risk bounds for multipass PSGD that depend on the effective dimension governed jointly by the loss curvature, preconditioning matrix, and gradient noise.

3. We identify a regime where an improperly chosen preconditioner leads to suboptimal effective dimension dependence in both optimisation and generalisation.

4. We complement our results by obtaining matching instance-dependent lower bounds.

Rather than working directly with $\nabla^2 f$ we employ a proxy PD matrix H such that $\nabla^2 \ell \preceq \beta H$ and perform an analysis in the geometry of the $\|\cdot\|_H$-norm. We focus on β-smooth (but not necessarily Lipschitz) losses and we consider two structural cases: strong convexity, and non-convex losses satisfying a Polyak-Łojasiewicz (PL) condition (see eq. (5)), both in $\|\cdot\|_H$-norm.

Smooth strongly convex losses. In the first setting, we consider an arbitrary choice of the preconditioner P. Proposition 10 implies that with step size $\sim 1/(t+1)$ the excess risk satisfies

$$\mathbb{E}_{S,\mathcal{A}}[\delta f(x_t)] \;\leq\; \frac{32\beta}{\lambda_{\min}(PH)^2\alpha^2}\frac{\mathbb{E}_S[\operatorname{tr}(PHP\Sigma_S)]}{t+1} \;+\; \frac{64\operatorname{tr}(P\Sigma)}{\lambda_{\min}(PH)\alpha}\left(\frac{1}{\sqrt{n(t+1)}}+\frac{1}{n}\right),$$

where $\operatorname{Var}_{i_t}[\nabla\ell(x,z_{i_t})] \preceq \Sigma_S$ for all x. Observe that the excess risk depends on the term $\operatorname{tr}(P\Sigma)$ which resembles effective dimension and multiplies $1/n$, which is a statistical rate. The term $\mathbb{E}_S[\operatorname{tr}(PHP\Sigma_S)]$ bears a similar role as it multiplies $1/t$, which is an optimiser convergence rate.[2] At the same time it is known that the optimal statistical rate is $\operatorname{tr}(H^{-1}\Sigma)/n$ and so the above suggests that the optimal choice $P = H^{-1}$ recovers the optimal rate $\operatorname{tr}(H^{-1}\Sigma)(1/t+1/n)$, while other choices will lead to the suboptimal statistical rate. This also demonstrates that the geometry required to minimise the variance in the optimisation error is identical to the geometry required to minimise finite-sample algorithmic instability. Thus, second-order information is not only a tool for speed, but a mechanism for robustness against sampling noise.

The key to the presence of $\operatorname{tr}(P\Sigma)$ stems from combination of on-average stability analysis and working with weighted Euclidean norms. This is elucidated by Lemma 8, which analyses stability of stochastic iterative algorithms that satisfy geometric update contractivity between iterates x_t and $x_t^{(i)}$ (where the latter is obtained on the perturbed training set).[3] In particular, for any such algorithm with r-contractive updates, any PD matrix M, and a constant step size η we have

$$\mathbb{E}\left[\|x_t - x_t^{(i)}\|_M^2\right] = \mathcal{O}\left(\operatorname{tr}(PMP\Sigma)\left(\frac{\eta}{nr}+\frac{1}{n^2r^2}\right)\right) \qquad \text{as} \qquad n\to\infty.$$

Note that this result only requires smoothness, but not convexity of the loss. First, the lemma captures stability of PSGD in a subspace of choice rather than globally. Choosing curvature $M = P^{-1}$ naturally leads to analysis of preconditioned SGD as iterates live in a subspace spanned by the preconditioner. Second, working with on-average stability allows us to gain dependence on Σ, whereas a stronger, uniform stability would be oblivious to geometry of the noise.

On-average stability for smooth PL losses. Next in Section 4.2 we extend our analysis to a family of non-convex smooth losses that satisfy PL condition. In particular, we show that excess risk is controlled by the effective dimension,

$$\mathbb{E}[\delta f(x_t(S))] \leq \frac{2\beta}{\mu}\mathbb{E}[\delta f_S(x_t(S))] + \frac{16\operatorname{tr}(H^{-1}\Sigma)}{\mu\, n}.$$

2. Note that $\mathbb{E}[\Sigma_S]$ can be controlled in terms of Σ with bias of order $\sqrt{\operatorname{tr}(PHP\Sigma)/n}$, see Lemma 22.
3. Update is r-contractive when $\|x-\eta P\nabla\ell(x,z)-y+\eta P\nabla\ell(y,z)\|_M^2 \leq (1-\eta r)\|x-y\|_M^2$ for $r>0$.

for large enough n. Note that, excess risk no longer depends on a particular P and behaves as if an optimal P was chosen. The expected optimisation error $\mathbb{E}[\delta f_S(x_t(S))]$ scales with the effective dimension as well and the bound is given by the standard convergence analysis for PL objectives (Karimi et al., 2016).

Lower bounds. Finally, a natural question is whether the results we presented are optimal. On one hand it is known that from both statistical and optimisation perspectives dependence on $\mathrm{tr}(H^{-1}\Sigma)$ is optimal as there exist asymptotic lower bounds (Cramér-Rao type lower bound (Polyak and Juditsky, 1992)). To this end, focusing on the strongly-convex model, in Section 5 we complement this fact in non-asymptotic sense, by showing that minimax lower bounds on the excess risk are of order $\mathrm{tr}(H^{-1}\Sigma)/(n\,\alpha)$. Clearly we cannot expect any improvement in the minimax sense, however, the message our analysis conveys is that a bad choice of the preconditioner might lead to a poor statistical performance of PSGD, and so minimax analysis is no longer appropriate. To this end we present an *instance-dependent* lower bound in Lemma 16. In particular, for a decaying step size $\eta_t \sim 1/t$, for a sufficiently large t, the expected excess risk behaves as

$$\frac{\mathrm{tr}(PHP\Sigma)}{\alpha\,\lambda_{\max}(PH)\,\lambda_{\min}(PH)} \cdot \frac{1}{t} + \frac{\mathrm{tr}(H^{-1}\Sigma)}{\alpha} \cdot \frac{1}{n}\,.$$

While for the optimal choice of the preconditioner $P = H^{-1}$ this bound matches the upper bound up to κ_ℓ factor[4], for a badly chosen preconditioner P (for instance, we can construct P that approaches rank-deficiency, i.e., $\kappa(P) = 1/\varepsilon$) the above result implies that the risk is lower bounded $\mathrm{tr}(H\Sigma)/(\varepsilon t)$ for $t > 4/\varepsilon$. In other words, for a general curvature (Σ, H), preconditioner P, and large t, the associated constant in front of the asymptotic rate of the excess risk can be arbitrarily large, even with decaying step. This, once more, emphasises the impact of the preconditioning on statistical performance of the last iterate.

Notation and terminology. For symmetric matrices A, B, we write $A \preceq B$ to denote the semidefinite order, meaning that $B - A$ is Positive Semi-Definite (PSD), and similarly $\prec$ to denote PD. We denote $\|x\|_H = \sqrt{x^\top H x}$ for a positive definite matrix $H \succ 0$. We let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the smallest and the largest eigenvalue and $\kappa(A) = \lambda_{\max}(A)/\lambda_{\min}(A)$ is the condition number of a matrix $A \in \mathbb{R}^{d\times d}$. For α-strongly convex β-smooth function w.r.t. $\|\cdot\|_H$-norm we denote $\kappa_\ell = \beta/\alpha$. Roman font Q, P_x denote probability distributions, the latter parameterised by the vector x.

2. Proof Sketch and Technical Challenges

The expected excess risk of an estimator $\hat{x}$ is typically bounded by balancing the trade-off between error terms originating from the generalisation component and those arising from offline optimisation of the empirical risk:

$$\mathbb{E}_S\left[\delta f(\hat{x})\right] = \underbrace{\mathbb{E}_S\left[f(\hat{x}) - f_S(\hat{x})\right]}_{\text{generalisation}} + \underbrace{\mathbb{E}_S\left[f_S(\hat{x}) - f_S(\tilde{x})\right]}_{\text{optimisation}}\,,$$

where $\tilde{x} = \arg\min_{x\in\mathcal{X}} f(x)$. Here the optimisation error can be further upper bounded using the ERM $x_S^* \in \arg\min_{\mathcal{X}} f_S(x)$ and noting that $\mathbb{E}_S\left[f_S(\hat{x}) - f_S(\tilde{x})\right] \le \mathbb{E}_S\left[f_S(\hat{x}) - f_S(x_S^*)\right]$.

4. The gap of κ_ℓ between the algorithmic lower bound (Lemma 16) and the upper bound (Proposition 10) arises from the fact that the former is derived for a quadratic loss while the latter applies to α-strongly convex β-smooth loss.

The generalisation term can be controlled using the standard algorithmic stability argument. Let $\hat{x}^{(i)}$ be computed from a perturbed dataset $S^{(i)} = S \setminus \{z_i\} \cup \{z'\}$, where $z' \sim \mathrm{Q}$ with the same algorithmic procedure as $\hat{x}$. Then using the standard symmetricity argument leads to observation that the generalisation term is equal to the *on-average algorithmic stability*, eq. (3).

2.1. Generalisation Geometry via On-Average Multipass Stability with Correlated Iterates

In the multi-pass setting, when x_t is computed by sampling examples from S with replacement, the iterate is not independent of previously seen samples z_{i_t} and the standard stability analysis fails. This usually forces the analysis to rely on *uniform stability bounds* (Hardt et al., 2016) that assume uniform $\ell(\cdot, z)$ is L-Lipschitz for all samples

$$\mathbb{E}_{z\sim\mathrm{Q}}[|\ell(x_t(S), z) - \ell(x_t(S^{(i)}), z)|] \leq \sup_{z\in\mathcal{Z}} |\ell(x_t(S), z) - \ell(x_t(S^{(i)}), z)| \leq L\|x_t(S) - x_t(S^{(i)})\|.$$

This step effectively removes any dependence on the data distribution and its interaction with finer geometric properties of the loss, which is commonly pointed out as a limitation (Zhang et al., 2017).

In order to reveal the generalisation geometry, we exploit that $\ell(\cdot, z)$ is β-smooth w.r.t $\|\cdot\|_H$-norm, and show that, when η_t is small, the generalisation is governed by

$$\begin{aligned}&\mathbb{E}_S\left[f(x_t(S)) - f_S(x_t(S))\right] \\ &\quad = \mathcal{O}\left(\mathrm{Var}_{z\sim\mathrm{Q}}\left[\|\nabla\ell(x_t(S), z)\|_*^2\right]^{1/2} \cdot \mathbb{E}_{\mathcal{A},S,z'}\left[\|x_t(S) - x_t(S^{(i)})\|^2\right]^{1/2}\right).\end{aligned}$$

The choice for $\|\cdot\|$-norm controlling the squared parameter stability $\varepsilon^2_{\mathrm{pstab}}(x_t(S), \|\cdot\|) := \mathbb{E}_{\mathcal{A},S,z'}[\|x_t(S) - x_t(S^{(i)})\|^2]$ plays a crucial role in two ways: it bounds the parameter stability and its dual norm will interact with the noise of gradients. We restrict ourselves to Hilbert spaces and consider $\varepsilon^2_{\mathrm{pstab}}(x_t(S), \|\cdot\|_M)$ for some $M \succ 0$. If the deterministic PGD update is r-contractive we can upper bound the parameter stability as

$$\varepsilon^2_{\mathrm{pstab}}(x_{t+1}(S)) \leq (1 - c_1\eta_t r)\,\varepsilon^2_{\mathrm{pstab}}(x_t(S)) + c_2\,\frac{\eta_t^2}{n}\,\mathbb{E}\left[\|P(\xi_t - \tilde{\xi}_t)\|_M^2\right],$$

where $\xi_t := \nabla\ell(x_t(S), z_i) - \nabla\ell(x_t(S^{(i)}), z')$ involves the challenging term with the correlated samples and parameters, and $\tilde{\xi}_t := \nabla f(x_t(S)) - \nabla f(x_t(S^{(i)}))$. We overcome the *problem of correlated iterates in the multi-pass setting* by being able to upper bound it as

$$\mathbb{E}\left[\|\xi_t - \tilde{\xi}_t\|^2_{PMP}\right] \leq \mathrm{tr}(PMP\Sigma) + c_3\beta^2\varepsilon^2_{\mathrm{pstab}}(x_t(S)).$$

We identify a condition $n \geq 4\kappa_\ell\kappa(PH)$ depending on the geometry of ℓ and P, that ensures the contribution of the correlated terms is benign, resulting in

$$\varepsilon^2_{\mathrm{pstab}}(x_t(S), \|\cdot\|_M) \;\leq\; \underbrace{c_4\,\frac{\mathrm{tr}(PMP\Sigma)}{n^2}}_{\text{Irreducible Fast Rate}} + \underbrace{c_5\,\eta\,\frac{\mathrm{tr}(PMP\Sigma)}{n}}_{\text{Optimisation Variance}} \qquad \text{for a fixed } \eta_t = \eta.$$

This decomposition is sharper than $\mathcal{O}(1/n)$ providing finer control as $\mathcal{O}(1/n^2)$ when $\eta \leq 1/n$, it isolates the intrinsic statistical complexity (the fast rate) from the noise induced by the algorithm's

step size, and establishes *on-average* stability for arbitrary t provided n is large enough. Thus, when the deterministic PGD update is r-contractive in $\|\cdot\|_M$-norm, η is small enough[5]

$$\mathbb{E}_S\left[f(\hat{x}) - f_S(\hat{x})\right] = \mathcal{O}\left(\frac{\sqrt{\operatorname{tr}(M^{-1}\Sigma)\operatorname{tr}(PMP\Sigma)}}{n}\right)$$

and selecting $\|\cdot\|_M$ determines the analysis's sensitivity to parameter stability and gradient noise.

2.2. Spectral Alignment under Geometric Mismatch

In the standard stability analysis, for α-strongly convex and β-smooth functions w.r.t. $\|\cdot\|_2$, the contractivity of the gradient descent (without preconditioning) comes from *gradient co-coercivity*:

$$\langle \nabla f(x) - \nabla f(y),\, x - y\rangle \geq \frac{\alpha\beta}{\alpha+\beta}\|x-y\|_2^2 + \frac{1}{\alpha+\beta}\|\nabla f(x) - \nabla f(y)\|_2^2.$$

When the geometry of ℓ is defined by $\|\cdot\|_H$-norm and we use preconditioning P the term to be bounded is $\langle HP(\nabla f(x) - \nabla f(y)),\, x - y\rangle$, which does not need to be positive unless $P = H^{-1}$. However, in practical settings we almost never have $P = H^{-1}$, the matrices P, H are misaligned and do not commute.

We introduce a rigorous condition for *spectral alignment* based on the matrix pencil (P, H^{-1}) and establish a *generalised co-coercivity inequality* for gradients under non-commuting preconditioning

$$\langle HP(\nabla f(x) - \nabla f(y)),\, x - y\rangle \geq \frac{\lambda_{\min}(PH)\, C_{\ell,P}}{\alpha+\beta}\Big(\alpha\beta\|x-y\|_H^2 + \|\nabla f(x) - \nabla f(y)\|_{H^{-1}}^2\Big),$$

where the constant $C_{\ell,P} \in (0, 1]$ tracks the quality of the alignment: $C_{\ell,P} = 1$ for quadratic functions ($\beta = \alpha$) and $C_{\ell,P} \to 0$ for badly aligned problems. This property allows us to show the contractivity of the preconditioned gradient update in the parameterised family of metrics $\|\cdot\|_{M_\theta}$ defined by $M_\theta = H^{1/2}(H^{1/2}PH^{1/2})^{-\theta}H^{1/2}$ interpolating between: the natural metric of the problem when $\theta = 0$ ($\|\cdot\|_H$), and the metric defined by the algorithm when $\theta = 1$ ($\|\cdot\|_{P^{-1}}$), which holds for any $P \succ 0$.

3. Preliminaries

Relative smoothness & strong convexity. We define the geometry w.r.t. the weighted norm $\|\cdot\|_H$.

Definition 1 (Smoothness w.r.t. $\|\cdot\|_H$) *Let $H \succ 0$ such that $\lambda_{\max}(H) = 1$, and $\beta > 0$. The function $f(x)$ is β-smooth w.r.t. $\|\cdot\|_H$ when $f(y) - f(x) \leq \langle \nabla f(x),\, y - x\rangle + \frac{\beta}{2}\|y-x\|_H^2$ or equivalently, $\|\nabla f(x) - \nabla f(y)\|_{H^{-1}} \leq \beta\|x-y\|_H$ for convex f.*

Definition 2 (Strong convexity w.r.t. $\|\cdot\|_H$) *Let $H \succ 0$ such that $\lambda_{\max}(H) = 1$, and $\alpha > 0$. The function $f(x)$ is α-strongly convex w.r.t. $\|\cdot\|_H$ when $f(y) - f(x) \geq \langle \nabla f(x),\, y - x\rangle + \frac{\alpha}{2}\|y-x\|_H^2$.*

The definitions are special cases of *relative smoothness and strong convexity*, see (Lu et al., 2018, Definition 1.1 and 1.2), where we choose the *reference function* to be $h(x) = \langle x, Hx\rangle$. They have

5. Or, for example, in the standard setting of $\eta_t \approx 1/t$ and $t \geq n$.

been referred to also as "matrix smoothness" employed by (Thomas et al., 2020; Li et al., 2024). We denote the condition number of the loss w.r.t. the $\|\cdot\|_H$-norm geometry by $\kappa_\ell := \beta/\alpha$.

Since κ_ℓ expresses the discrepancy between f and a quadratic function, on a bounded domain it can be bounded using higher order smoothness. If $\ell(\cdot, z)$ has γ-smooth Hessian, i.e., $\lambda_{\max}(\nabla^2\ell(x_1, z) - \nabla^2\ell(x_2, z)) \le \gamma\|x - y\|_2$, we have the following bound

$$\kappa_\ell \le \frac{1 + \gamma R/\lambda_{\min}(H)}{1 - \gamma R/\lambda_{\min}(H)}, \quad \text{for } \|x - x_0\|_2 \le R.$$

For a fixed sample $z \in \mathcal{Z}$ and $\ell(\cdot, z) \in C^2$, the combination of Definitions 1 and 2 acts as a quadratic upper and lower bound respectively: $\alpha H \preceq \nabla^2\ell(x, z) \preceq \beta H$ for all x.

Generalised co-coercivity. The following defines the spectrally aligned preconditioner when the relative condition number $\kappa(PH)$ is sufficiently bounded compared to κ_ℓ.

Definition 3 (Spectrally aligned preconditioner) *For $\ell(\cdot, z)$ that is α-strongly convex and β-smooth w.r.t. $\|\cdot\|_H$, we say that P is $C_{\ell,P}$-*spectrally aligned *with the geometry of $\ell(\cdot, z)$ for $C_{\ell,P} \in (0, 1]$ iff*

$$\kappa(PH) \le \rho_\ell^2 \quad \textit{with} \quad C_{\ell,P} = \frac{\rho_\ell^2 - \kappa(PH)}{\rho_\ell^2 - 1} \quad \textit{and} \quad \rho_\ell := \frac{\sqrt{\kappa_\ell} + 1}{\sqrt{\kappa_\ell} - 1} > 1.$$

This decomposes the conditioning misalignment into two parts: ρ_ℓ reflects how well the model of relative smoothness/strong convexity captures the actual geometry of ℓ, and $\kappa(PH)$ reflects how well the algorithm, i.e., the choice of P, captures the model curvature defined by H. Definition 3 is satisfied for many widely used choices of P and allows for a fine description of the geometry needed for generalisation.

Example 1 (Inexact-Newton Methods (q-approximate inverse curvature)) *Assume that for some $q \ge 1$ the preconditioner $P \succ 0$ satisfies $\frac{1}{q} H^{-1} \preceq P \preceq q\, H^{-1}$. Then $(1/q)I \preceq PH \preceq qI$, hence $\kappa(PH) \le q^2$. Therefore, whenever $q^2 < \rho_\ell^2$, the preconditioner is spectrally aligned, and the alignment constant is lower bounded as*

$$C_{\ell,P} \;=\; \frac{\rho_\ell^2 - \kappa(PH)}{\rho_\ell^2 - 1} \;\ge\; \frac{\rho_\ell^2 - q^2}{\rho_\ell^2 - 1}.$$

Such uniform spectral boundedness assumptions have been used in the Quasi-Newton literature to prove global convergence; see, e.g., (Nocedal and Wright, 2006, Sec. 3.3) and (Dennis and Moré, 1977). More recently Cheng and Li (2010) showed that ensuring the q-approximate inverse curvature leads to improved numerical performance.

Example 2 (Diagonal preconditioning) *Let P be a diagonal preconditioner $P := \mathrm{diag}(H)^{-1}$. If $A := D^{-1/2} H D^{-1/2}$ is strictly diagonally dominant in the sense that $\alpha := \max_i \sum_{j \ne i} |A_{ij}| < 1$, then by Gershgorin disc theorem we have that the spectrum $\lambda(A) \subset [1-\alpha, 1+\alpha]$, hence $\kappa(PH) = \kappa(A) \le \frac{1+\alpha}{1-\alpha}$. Consequently, whenever $\frac{1+\alpha}{1-\alpha} < \rho_\ell^2$,*

$$C_{\ell,P} \;\ge\; \frac{\rho_\ell^2 - \frac{1+\alpha}{1-\alpha}}{\rho_\ell^2 - 1},$$

yielding a simple explicit $C_{\ell,P}$ bound for diagonal preconditioning whenever H is close to diagonal.

Definition 3 allows to derive a generalisation of the standard gradient co-coercivity result, e.g., see (Nesterov, 2018, Theorem 2.1.12), that applies to preconditioned gradients and the specific case of relative smoothness and strong convexity (Lu et al., 2018).

Lemma 4 (Co-coercivity of spectrally aligned PSGD updates) *Let f be α-strongly convex and β-smooth w.r.t. $\|\cdot\|_H$ and P is $C_{\ell,P}$-spectrally aligned with $\ell(\cdot, z)$, i.e., $\kappa(PH) < \rho_\ell^2$ in Definition 3. Then for all $x, y \in \mathbb{R}^d$:*

$$\langle HP(\nabla f(x) - \nabla f(y)), x - y\rangle \geq \frac{\lambda_{\min}(PH) C_{\ell,P}}{\alpha + \beta} \left(\alpha\beta \|x - y\|_H^2 + \|\nabla f(x) - \nabla f(y)\|_{H^{-1}}^2\right).$$

Proof is given in Section B.1. For $P = H^{-1}$ this recovers the standard co-coercivity of gradients. Note, that Lemma 4 does not require that P and H commute.

4. Excess risk bounds of PSGD via on-average stability

In this section, we derive excess risk bounds for the PSGD algorithm via on-average stability. Throughout this section, assume the following:

Assumption 5 *Suppose that for each $z \in \mathcal{Z}$, $\ell(\cdot, z)$ is β-smooth with respect to the norm, $\|\cdot\|_H$.*

Assumption 6 *Suppose there exists $\Sigma \succ 0$ such that $\mathrm{Cov}_{z\sim Q}(\nabla \ell(x, z)) \preccurlyeq \Sigma$ for all $x \in \mathcal{X}$.*

In traditional stability analyses of SGD-type algorithms, a uniform Lipschitz assumption is typically employed to relate algorithmic stability to parameter stability. However, this Lipschitz assumption both rules out several settings of interest (e.g. strongly convex losses on non-compact domains) and often conceals the curvature information present in smoothness and convexity. To fully exploit the geometry of the problem, we proceed without the global Lipschitz assumption. We first provide a general stability result for an algorithm $\mathcal{A}$ that maps from $\mathcal{Z}^n$ to a random variable on $\mathcal{X}$.

Lemma 7 *Suppose that assumptions 5 and 6 hold and let $M \succ 0$. If the algorithm $\mathcal{A}$ is L^2-on-average parameter stable in $\|\cdot\|_M$-norm with constant $\varepsilon_{\mathrm{pstab}}^2 \geq 0$, then the expected excess risk on the parameters $x = \mathcal{A}(S)$ satisfies,*

$$\mathbb{E}_{S,\mathcal{A}}[\delta f(x_t)] \;\leq\; 2\mathbb{E}_{S,\mathcal{A}}[\delta f_S(x_t)] + 2\,\mathrm{tr}(M^{-1}\Sigma)^{1/2}\,\varepsilon_{\mathrm{pstab}} + 4\beta\lambda_{\max}(HM^{-1})\,\varepsilon_{\mathrm{pstab}}^2.$$

Note that the primary limitation borne from the Lipschitz assumption, that we are able to overcome, is that parameter stability is measured in the weaker $\|\cdot\|_M$, whereas our analysis uses $\|\cdot\|_M^2$-norm instead. While this requires a tighter control on the iterates, it allows us to use smoothness to identify the explicit role of the curvature (via H and M) in the generalisation bound. The matrix M is chosen according to its amenability to the parameter stability analysis, but to optimise the bound, M must also be chosen to align with either the curvature H or the covariance matrix Σ. Thus, under misspecification, the natural geometry to analyse parameter stability in is not immediate.

We now focus on the PSGD algorithm defined by a positive definite preconditioner, $P \succ 0$. To analyse the stability of PSGD, we will obtain that if the update is contractive in a suitable geometry

$$\|x - \eta P\nabla\ell(x, z) - y + \eta P\nabla\ell(y, z)\|_M^2 \leq (1 - \eta r)\|x - y\|_M^2. \tag{4}$$

Lemma 8 (On-average parameter stability of PSGD) *Suppose that Assumption 5 holds, choose any matrix $M \succ 0$ and constants $\bar{\eta}, r > 0$ such that for any $x, y \in \mathcal{X}, z \in \mathcal{Z}$ and $\eta \leq \bar{\eta}$, the r-contractivity property in eq. (4) holds. Then, if $\sup_s \eta_s \leq \bar{\eta} \wedge r^{-1} \wedge (\beta\gamma)^{-1}$ and $n \geq 34\beta\sqrt{\lambda_{\max}(HPMP)} \cdot \sqrt{\lambda_{\max}(M^{-1}H)}/r$, where $\gamma^2 := \lambda_{\max}(HPMP)\lambda_{\max}(M^{-1}H)$, we have that $\mathcal{A}_{P,t}$ is on-average parameter stable with constant,*

$$\varepsilon^2_{\text{pstab}} \leq 64\left(\frac{\bar{\eta}_t}{8n} + \frac{1 - e^{-T_t r/4}}{n^2 r^2}\right) \operatorname{tr}(PMP\Sigma) \,,$$

where $T_s = \sum_{s'=0}^{s-1} \eta_{s'}$ and $\bar{\eta}_t = \sum_{s<t} e^{-r\frac{T_t - T_s}{4}} \eta_s^2$.

The proof is provided in section C.1. A significant advantage of this bound over those in (Hardt et al., 2016) is the explicit dependence on the data distribution via the trace term $\operatorname{tr}(PMP\Sigma)$. Furthermore, unlike (Kuzborskij and Lampert, 2018), our result captures the exact interaction between the curvature H, the preconditioner P, and the noise Σ, while valid in the multi-pass setting.

The quantity $\bar{\eta}_t$ characterises how memory of past step-sizes decays. For standard step-size schedules, it behaves intuitively: for example, with a linearly decaying step size $\eta_t = c/t$, we have $\bar{\eta}_t \leq c^2/t$.

4.1. On-average stability and risk bounds for strongly convex smooth losses

Under additional assumption that the loss is also α-strongly convex, we can show that the PGD update is r-contractive in a specific family of $\|\cdot\|_M$-norms.

Assumption 9 *Suppose that for each $z \in \mathcal{Z}$, $\ell(\cdot, z)$ is α-strongly convex with respect to $\|\cdot\|_H$.*

Lemma 19 in Section B shows that under Assumption 9, the PGD update is contractive in $\|\cdot\|_{M_\theta}$ where $M_\theta := H^{1/2}(H^{1/2}PH^{1/2})^{-\theta}H^{1/2}$ for $\theta \in [0, 1]$ interpolating between H and P^{-1}.

By combining the stability result in Lemma 8, the contractivity result in Lemma 19, and the optimisation rates for PSGD (see Section D), we can derive explicit generalisation bounds, see Lemma 25. We consider two natural geometries for measuring convergence: the geometry induced by P^{-1} ($\theta = 1$) and the geometry induced by the Hessian H ($\theta = 0$).

Proposition 10 (Risk bounds in geometry defined by P^{-1}) *Let $P \succ 0$, suppose that Assumptions assumption 5, 6 and 9 hold, and that $n \geq 4\kappa_\ell \kappa(PH)$. Let $r := 2\,\lambda_{\min}(PH)\,\frac{\alpha\,\beta}{\alpha+\beta}$. Let $\operatorname{Var}_{i_t}[\nabla\ell(x, z_{i_t})] \preceq \Sigma_S$ for all x.*

If the stepsizes are chosen as $\eta_t := \min\{1/(\beta\,\lambda_{\max}(PH)), 8/(r(t+1))\}$, then, for all t sufficiently large, the population excess risk satisfies

$$\mathbb{E}_{S,\mathcal{A}}[\delta f(x_t)] \;\leq\; \frac{64}{r}\left(\frac{\beta}{r}\,\frac{\mathbb{E}_S[\operatorname{tr}(PHP\Sigma_S)]}{t+1} \;+\; \operatorname{tr}(P\Sigma)\left(\frac{1}{\sqrt{n(t+1)}} + \frac{1}{n}\right)\right) .$$

We get sublinear $\mathcal{O}(1/t + 1/\sqrt{t\,n} + 1/n)$ convergence rate which matches the single pass (when $n = t$) result (Rakhlin et al., 2012), however with the precise rates depending on the interplay of the curvature, variance of noise, and how well the preconditioning adapts to these. Note that $\mathbb{E}[\Sigma_S]$ can be bounded by Σ with additive bias of order $L\beta\varepsilon_{\text{pstab}}$ assuming that ℓ is L-Lipschitz, see Lemma 22.

This means that for any $P \succ 0$, the excess risk of the last iterate of PSGD converges to zero asymptotically, although the rate in the upper bound can become arbitrarily loose with large $\kappa(PH)$,

even if the variance is bounded. Corollary 20 shows that minimizing the upper bound in terms of $P \succ 0$ yields that $P = H^{-1}$ minimizes the expected risk and gives the optimal Takeuchi Information Criterion for the noisy strongly convex smooth model (theorem 15).

Remark 11 (Approximate NGD under misspecification) *Due to the connection to the natural gradient descent discussed in Section A, this result has implications for NGD under misspecification. Let $\ell(x, z) := -\log p(z|x)$ be the negative log-likelihood of the distribution of $z \sim P_x$ and $\ell(\cdot, z)$ is α strongly convex, β smooth w.r.t. $\|\cdot\|_H$. If the data distribution differs from the model family (misspecification), we have $\Sigma \neq H$. Our bounds show that choosing $P = H^{-1} \approx (F_{\mathrm{P}_x}(x))^{-1}$ achieves a generalisation bound that is optimal even under this misspecification.*

In the case where P and H^{-1} are spectrally aligned, we can get more precise bounds through an analysis in $\|\cdot\|_H$-norm.

Proposition 12 (Risk bounds in geometry defined by H) *Suppose that Assumptions Assumption 5, 6 and 9 hold, and that $n \geq \frac{8\beta}{r}\sqrt{\lambda_{\max}(HPHP)}$. Assume further that $\kappa(PH) \leq \rho_\ell^2$ and let $r := 2\,\lambda_{\min}(PH)\, C_{\ell,P}(\beta\,\alpha)/(\alpha+\beta)$.*

If the stepsizes are chosen as $\eta_t := \min\{C_{\ell,P}/(\beta\,\lambda_{\max}(PH)\,\kappa(PH)),\ 8/(r(t+1))\}$, then, for all t sufficiently large, the population excess risk satisfies

$$\mathbb{E}_{S,\mathcal{A}}[\delta f(x_t)] \leq \frac{64}{r}\left(\frac{\beta}{r}\frac{\mathbb{E}_S[\mathrm{tr}(PHP\Sigma_S)]}{t+1} + \sqrt{\mathrm{tr}(H^{-1}\Sigma)\ \mathrm{tr}(PHP\Sigma)}\left(\frac{1}{\sqrt{n(t+1)}} + \frac{1}{n}\right)\right).$$

Note, that since $\lambda_{\max}(P) = \lambda_{\max}(H) = 1$, we have that $\lambda_{\max}(PH) \leq 1$, and thus $\mathrm{tr}(PHP\Sigma) \leq \mathrm{tr}(P\Sigma)$ making the rate in Proposition 12 less than or equal to the one in Proposition 10.

4.2. Risk bounds for non-convex losses under PL-property

While the above analysis captures the generalisation properties of PSGD along the trajectory, it fails to capture what occurs at convergence. This can be seen due to the fact that, irrespective of the choice of preconditioner, the PSGD iterates should converge to the same empirical risk minimiser, and thus, exhibit the same generalisation properties. The inability of this type of stability analysis to capture generalisation at convergence is known (Hardt et al., 2016). For that reason, we turn instead to a black-box analysis of any algorithm $\mathcal{A}$ that produces parameters that approximately minimise f_S.

Here we also consider a more general setting than the previous analysis under strong convexity. In addition to β-smoothness w.r.t. $\|\cdot\|_H$ we will assume that empirical risk satisfies the following PL condition (Karimi et al., 2016): There exists $\mu > 0$ and a minimizer x^* of f_S such that for all S, x,

$$\frac{1}{2}\|\nabla f_S(x)\|^2_{H^{-1}} \geq \mu(f_S(x) - f_S(x^*)). \tag{5}$$

Our analysis is inspired by that of Charles and Papailiopoulos (2018) and we make the following assumption which is identical to their Assumption 1 and the stability analysis that follows is similar to their proof technique of Theorem 3(iii).

Assumption 13 *The empirical risk minimizers for f_S and $f_{S^{(i)}}$, i.e., $\hat{x}^*, \hat{y}^*$, satisfy $\mathrm{Proj}_S(\hat{y}^*) = \hat{x}^*$, where Proj_S is the projection on the set of empirical risk minimizers of f_S.*

Proposition 14 (Excess risk bounds for PL-losses) *Suppose that f_S is β-smooth and satisfies the μ-PL property w.r.t. $\|\cdot\|_H$ and suppose that Assumption 13 holds. Then whenever $n \geq 4\beta/\mu$, we have the excess risk bound,*

$$\mathbb{E}_{\mathcal{A},S}[\delta f(x_t(S))] \leq \frac{2\beta}{\mu}\mathbb{E}[\delta f_S(x_t(S))] + \frac{16\,\mathrm{tr}(H^{-1}\Sigma)}{\mu n} + 128\beta\frac{\mathrm{tr}(H^{-1}\Sigma)}{\mu^2 n^2}.$$

Together, these results suggest that the generalisation dynamics are governed by a delicate trade-off mediated by the preconditioner. The choice of P dictates the learning trajectory in two distinct ways:

1. **Optimisation Rate:** P determines the convergence speed of the empirical error $\mathbb{E}[\delta f_S(x_t)]$, primarily through the condition number $\kappa(PH)$.
2. **Effective Dimension:** P shapes the effective noise geometry, scaling the stability error by $\mathrm{tr}(PHP\Sigma)$ or $\mathrm{tr}(P\Sigma)$ in the worst case.

Furthermore, once the algorithm converges, the excess risk it produces becomes independent of the choice of preconditioner. The optimal choice $P \approx H^{-1}$ simultaneously maximises the convergence rate and minimises the effective dimension, acting as a benefit to both optimisation and generalisation.

5. Lower bounds on the expected risk

Theorem 15 (Lower bound) *Let $\ell(x,z) : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ be strongly convex w.r.t. $\|\cdot\|_H$-norm in the parameters x and $\mathcal{P}$ is a family of distributions such that $\forall P \in \mathcal{P}, x \in \mathcal{X}$ we have $\mathbb{E}_{z\sim P}[\nabla\ell(x,z)] = 0$ and $\mathrm{Var}_{z\sim P}(\nabla\ell(x,z)) = \Sigma$. Then we have that the expected excess risk of an estimator computed from $S \sim P$ is lower bounded as*

$$\inf_{\hat{x}\in\mathcal{X}} \sup_{P\in\mathcal{P}} \mathbb{E}_{S\sim P^n}[\delta f(\hat{x}(S))] \geq \frac{2}{27\,n\alpha}\,\mathrm{tr}(H^{-1}\Sigma)\,.$$

Proof is in Section G.1. Theorem 15 establishes that the fundamental statistical limit of the problem is governed by the interaction between the geometry of the loss (H) and the noise structure (via Σ).

Algorithmic lower bounds of multipass PSGD While previously we demonstrated that the choice of $P = H^{-1}$ results in non-asymptotically optimal rate, here we show that, even in our simple setting, choosing a bad preconditioner P can increase the risk of the last iterate by a multiplicative factor $\kappa(PH)$. These bounds can be compared with the lower bound in (Nesterov, 2018, Theorem 2.1.13), but here we have preconditioning, decaying step-sizes, and lower bound multi-pass risk. The first result shows that the rate in Proposition 12 is tight up to the constant $\kappa(PH)$ for t large enough.

Lemma 16 (Algorithmic multipass lower bound) *The expected excess risk of the $(t+1)^{\text{th}}$-iterate of multipass PSGD with $\eta_t = \min\left\{\frac{1}{2\alpha\lambda_{\max}(PH)}, \frac{2}{\alpha\lambda_{\min}(PH)(t+1)}\right\}$ on the quadratic noisy model with n samples is lower bounded as*

$$\mathbb{E}_{\mathcal{A},S}[\delta f(x_{t+1})] \geq \left(1-\frac{1}{n}\right)\frac{\mathrm{tr}(PHP\Sigma)}{2\alpha\,\lambda_{\max}(PH)\lambda_{\min}(PH)}\cdot\frac{1}{t+1} + \frac{\gamma_t(P,H)}{2\alpha n}\,\mathrm{tr}(H^{-1}\Sigma),$$

when $t \geq \lceil 4\kappa(PH)\rceil$ where $\gamma_t(P,H) = \left(1-\prod_{s=0}^{t}\left(1-\alpha\eta_s\lambda_{\min}(PH)\right)\right)^2$ and for which we have that $\lim_{t\to\infty}\gamma_t(P,H) = 1$.

Proof is in Section G.2. A similar type of lower bound, but for the single-pass case, is derived in (Martens, 2020, Theorem 5).

Consequently, for any given H, Σ, choosing a badly conditioned P can make the lower bound on the risk rate arbitrarily larger than the optimal rate.

Corollary 17 (Algorithmic lower bound for ill-conditioned P) *Choose $\varepsilon > 0$ and assume that $t > 4/\varepsilon$. Let $H, \Sigma \succ 0$ and $\lambda_{\max}(H) = 1$ and Q be the eigenbasis of H. Let (h_k, q_k) be the eigenpair of H for the index k explicitly defined by the spectrum of H and Σ. Then PSGD with a decaying stepsize $\eta = \min\{1/(2\alpha\lambda_{\max}(PH)), 2/(t\lambda_{\min}(PH)\alpha)\}$ and a preconditioner $P_\varepsilon = I - (1 - \frac{\varepsilon}{h_k})q_k q_k^\top$ has the risk lower bounded as*

$$\mathbb{E}_{\mathcal{A},S}[\delta f(x_{t+1})] \geq \left(1 - \frac{1}{n}\right)\left(1 - \frac{1}{d-1}\right) \cdot \frac{\operatorname{tr}(H\Sigma)}{2\alpha\varepsilon(t+1)} \,.$$

The proof is given in Section G.3. This shows that for a general H, Σ and t large enough, the constant in front of the excess risk rate can get arbitrarily large in general, even with a decaying stepsize, as P_ε approaches a rank-deficiency.

One would expect that $P = I$ is a relatively safe choice. However, when the problem is ill-conditioned in the form of H, even a well-conditioned P can lead to significantly worse rates. We show that for any given P and H, in the presence of low-dimensional noise Σ, the lower bound on the risk of the last iterate of PSGD is at least $\kappa(PH)$ worse than the optimal rate of $\operatorname{tr}(H^{-1}\Sigma)/t$.

Corollary 18 (Algorithmic lower bound for ill-conditioned H) *Let $P, H \succ 0$ and $\lambda_{\max}(P) = \lambda_{\max}(H) = 1$. Assume that $t > 4\kappa(PH)$. Let q_1 be the leading eigenvector of $H^{1/2}PH^{1/2}$ and set the variance of noise to $\Sigma = H^{1/2}q_1 q_1^\top H^{1/2}$ (the bound is invariant to the scale of Σ, so this choice is without loss of generality). Then PSGD with a preconditioner P, a decaying stepsize $\eta = \min\{1/(2\alpha\lambda_{\max}(PH)), 2/(t\lambda_{\min}(PH)\alpha)\}$, has the risk lower bounded as,*

$$\mathbb{E}_{\mathcal{A},S}[\delta f(x_{t+1})] \geq \left(1 - \frac{1}{n}\right)\kappa(PH) \cdot \frac{\operatorname{tr}(H^{-1}\Sigma)}{2\alpha(t+1)} \,.$$

The proof is given in Section G.4 and consists of showing that the lower bound in Lemma 16 for any $P \succ 0$ is $\kappa(PH)$ larger than the optimal rate. Even for well-conditioned P the constant in the risk bound can be arbitrarily bad by having H badly conditioned. For example, SGD, i.e., when $P = I$, has its risk at least $\kappa(H)$ (which can be arbitrarily large) worse than the optimal rate.

6. Numerical example: PSGD on noisy quadratics

We observe the excess risk of a single-pass PSGD applied to a simple noisy quadratic loss problem: $\ell(x, z) = \frac{1}{2}\|x - z\|_H^2$, with the data distribution $z \sim \mathrm{Q} := \mathcal{N}\big(0,\, H^{-1}\Sigma H^{-1}\big)$, so that $\nabla_x \ell(x, z) = H(x - z)$ with $\operatorname{Var}_{z\sim\mathrm{Q}}(\nabla_x \ell(x, z)) = \Sigma$. We use online PSGD with the schedule $\eta_t = \min\big(1/\lambda_{\max}(PH),\, 2/(\lambda_{\min}(PH)\, t)\big)$ and generate 24 problem instances in dimension $d = 10$, each with a sampled dense SPD triple (H, Σ, P) whose eigenvalues are linearly spaced in $[0.1, 1]$ over independent Haar-random orthogonal eigenbases (the matrices generically do not commute). For each instance we run 512 independent trajectories for $T = 5000$ steps and report the Monte Carlo mean of the final scaled excess risk $T\,\mathbb{E}[\delta f(x_T)]$. In Figure 2 we see that the measured excess risk correlates strongly with the predicted upper bound $\operatorname{tr}(P\Sigma)/\lambda_{\min}(PH)$ with $R^2 = 0.98$, more

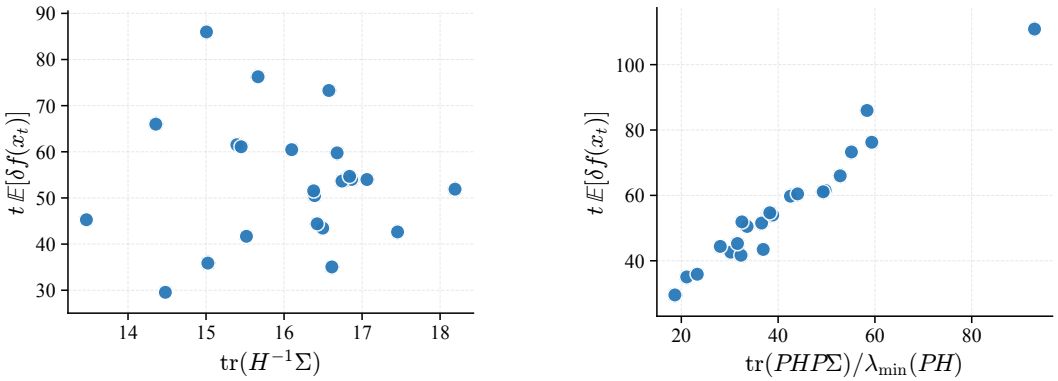

Figure 2: PSGD on randomly sampled noisy quadratic problems (H, Σ) and random matrices P.

so than with the rate predicted by local asymptotic statistics $\mathrm{tr}(H^{-1}\Sigma)$ with $R^2 = 0.03$. This confirms that the non-asymptotic behaviour of preconditioned SGD on anisotropic noisy quadratics is governed by the interaction of P, H, and Σ rather than by H and Σ alone.

Acknowledgments

Simon Vary and Patrick Rebeschini were funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number EP/Y028333/1]. Tyler Farghly was supported by Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/T517811/1] and by the DeepMind scholarship.

References

Naman Agarwal and Alon Gonen. Optimal sketching bounds for exp-concave stochastic minimization. arXiv:1805.08268, 2018.

Shun-ichi Amari. Natural Gradient Works Efficiently in Learning. *Neural Computation*, 10(2): 251–276, 1998. doi: 10.1162/089976698300017746.

Francis Bach. *Learning theory from first principles*. MIT press, 2024.

Olivier Bousquet and Andre Elisseeff. Stability and Generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper Bounds for Uniformly Stable Algorithms. In *Conference on Computational Learning Theory (COLT)*, 2020.

Zachary Charles and Dimitris Papailiopoulos. Stability and Generalization of Learning Algorithms that Converge to Global Optima. In *International Conference on Machine Learing (ICML)*, 2018.

Wanyou Cheng and Dong-Hui Li. Spectral Scaling BFGS Method. *Journal of Optimization Theory and Applications*, 146(2):305–319, August 2010. ISSN 0022-3239, 1573-2878. doi: 10.1007/s10957-010-9652-y.

John E Dennis, Jr and Jorge J Moré. Quasi-Newton Methods, Motivation and Theory. *SIAM Review*, 19(1):46–89, January 1977. ISSN 0036-1445, 1095-7200. doi: 10.1137/1019005.

Luc Devroye and Terry Wagner. Distribution-free inequalities for the deleted and holdout error estimates. *IEEE Transactions on Information Theory*, 25(2):202–207, 1979. doi: 10.1109/TIT.1979.1056032.

Simon S Du, Jason D Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient Descent Finds Global Minima of Deep Neural Networks. In *International Conference on Machine Learing (ICML)*, 2019.

Andre Elisseeff, Theodoros Evgeniou, Massimiliano Pontil, and Leslie Pack Kaelbing. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6(1), 2005.

Vitaly Feldman and Jan Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Computational Learning Theory (COLT)*, pages 1270–1279. PMLR, 2019.

Zachary Frangella, Pratik Rathore, Shipu Zhao, and Madeleine Udell. PROMISE: Preconditioned Stochastic Optimization Methods by Incorporating Scalable Curvature Estimates. *Journal of Machine Learning Research*, 25(346):1–57, 2024a.

Zachary Frangella, Pratik Rathore, Shipu Zhao, and Madeleine Udell. SketchySGD: Reliable Stochastic Optimization via Randomized Curvature Estimates. *SIAM Journal on Mathematics of Data Science*, 6(4):1173–1204, 2024b. doi: 10.1137/23M1575330.

Alon Gonen and Shai Shalev-Shwartz. Average Stability is Invariant to Data Preconditioning. Implications to Exp-concave Empirical Risk Minimization. *Journal of Machine Learning Research*, 18(222):1–13, 2018.

Moritz Hardt, Benjamin Recht, and Yoram Singer. Train Faster, Generalize Better: Stability of Stochastic Gradient Descent. In *International Conference on Machine Learing (ICML)*, 2016.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems*, 2018.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Łojasiewicz Condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2016.

Michael Kearns and Dana Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. In *Conference on Computational Learning Theory (COLT)*, 1997.

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. In *International Conference on Learning Representations (ICLR)*, 2017.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Ilja Kuzborskij and Christoph H Lampert. Data-Dependent Stability of Stochastic Gradient Descent. In *International Conference on Machine Learing (ICML)*, 2018.

Ilja Kuzborskij, Nicolò Cesa-Bianchi, and Csaba Szepesvári. Distribution-dependent analysis of Gibbs-ERM principle. In *Conference on Computational Learning Theory (COLT)*, 2019.

Jaehoon Lee, Lechao Xiao, Samuel S Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12): 124002, 2020. ISSN 1742-5468. doi: 10.1088/1742-5468/abc62b.

Yunwen Lei. Stability and Generalization of Stochastic Optimization with Nonconvex and Nonsmooth Problems. In *Conference on Computational Learning Theory (COLT)*, 2023.

Yunwen Lei and Yiming Ying. Fine-Grained Analysis of Stability and Generalization for Stochastic Gradient Descent. In *International Conference on Machine Learing (ICML)*, 2020.

Hanmin Li, Avetik Karagulyan, and Peter Richtárik. Det-CGD: Compressed Gradient Descent with Matrix Stepsizes for Non-Convex Optimization. In *OPT 2023: Optimization for Machine Learning at NeurIPS*, 2024.

Haihao Lu, Robert M. Freund, and Yurii Nesterov. Relatively Smooth Convex Optimization by First-Order Methods, and Applications. *SIAM Journal on Optimization*, 28(1):333–354, January 2018. ISSN 1052-6234, 1095-7189. doi: 10.1137/16M1099546.

Tianyi Ma, Kabir A. Verchand, and Richard J. Samworth. High-probability minimax lower bounds. arXiv:2406.13447, 2024.

James Martens. New Insights and Perspectives on the Natural Gradient Method. *Journal of Machine Learning Research*, 21(146):1–76, 2020.

James Martens and Roger Grosse. Optimizing Neural Networks with Kronecker-factored Approximate Curvature. In *International Conference on Machine Learing (ICML)*, 2015.

Yurii Nesterov. *Lectures on Convex Optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer International Publishing, Cham, 2018. ISBN 978-3-319-91577-7. doi: 10.1007/978-3-319-91578-4.

Gergely Neu and Lorenzo Rosasco. Iterate averaging as regularization for stochastic gradient descent. In *Conference on Computational Learning Theory (COLT)*, 2018.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, 2017.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, New York, NY, second edition edition, 2006.

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*, 2018.

Boris T Polyak and Anatoli B Juditsky. Acceleration of Stochastic Approximation by Averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, July 1992. ISSN 0363-0129, 1095-7138. doi: 10.1137/0330046.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization. In *International Conference on Machine Learing (ICML)*, 2012.

Ritei Shibata. Statistical Aspects of Model Selection. In Jan C. Willems, editor, *From Data to Model*, pages 215–240. Springer Berlin Heidelberg, Berlin, Heidelberg, 1989. doi: 10.1007/978-3-642-75007-6_5.

Jingruo Sun, Zachary Frangella, and Madeleine Udell. SAPPHIRE: Preconditioned Stochastic Variance Reduction for Faster Large-Scale Statistical Learning, January 2025.

Valentin Thomas, Fabian Pedregosa, Bart van Merriënboer, Pierre-Antoine Mangazol, 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 (AISTATS)*, 2020.

Zhewei Yao, Amir Gholami, Sheng Shen, Mustafa Mustafa, Kurt Keutzer, and Michael Mahoney. ADAHESSIAN: An Adaptive Second Order Optimizer for Machine Learning. In *Conference on Artificial Intelligence (AAAI)*, 2021.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017.

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*, 2024.

Appendix A. Additional related work

Algorithmic Stability. Algorithmic stability of SGD was first explored by Hardt et al. (2016) where they focused exclusively on the uniform stability. Their involved multipass SGD only for strongly convex, smooth, and Lipschitz losses. To this end their bounds did not involve any distribution-dependent quantities. Later on their analysis was extended to on-average stability setting by Kuzborskij and Lampert (2018), who showed that bounds on the expected generalisation gap controlled by the expected empirical risk, however their analysis is limited to a single pass over the data. Lei and Ying (2020) improved rate obtained by Kuzborskij and Lampert (2018), however their analysis still did not extend to multiple passes. Single pass limitation in on-average stability analysis is a common problem, since iterates become correlated after a single pass. Notably, multipass analysis was explored by Pillaud-Vivien et al. (2018) (Theorem 2 and 3), however their bounds become vacuous as $\lambda \to 0$. In this paper we address this limitation and recover optimal rates by exploiting smoothness and recursively controlling stability along the update trajectory (see Section 2).

The connection between on-average stability and a slightly different notion of effective dimension ($\operatorname{tr}(\nabla^2 f(\nabla^2 f + \lambda I)^{-1})$) in the context of regularized algorithms was studied by Agarwal and Gonen (2018). They showed generalisation error bounds for minimizers of smooth Lipschitz exp-concave losses that depend on such an effective dimension. Here we study stochastic iterative algorithm rather than a minimizer, we are particularly interested in the role of preconditioning and geometry of the noise.

The connection between on-average stability and preconditioning was explored by Gonen and Shalev-Shwartz (2018), who established that on-average stability is invariant to data preconditioning: In other words analyzing on-average stability of ERM one may assume the optimal preconditioning of the data – this is different from our setting as they look at asymptotic regime, in a sense $t \to \infty$. Moreover their analysis requires Lipschitzness of the loss, whereas do not require such as assumption.

Generalisation and Flatness. Relationship between generalisation and flatness (or, conversely, sharpness) is a topic of significant interest, in particular in deep learning where it was empirically and theoretically observed that neural networks trained by SGD tend to have a smaller generalisation error when they converge to 'wider' local minima (Keskar et al., 2017; Neyshabur et al., 2017), usually with some heuristic definition of width. In this paper we associate with the effective dimension, which is a natural geometric characterisation.

In the context of non-convex analysis linking effective dimension with generalisation, Kuzborskij et al. (2019) prove distribution-dependent excess risk bounds for Gibbs-ERM principle (as an idealized model of stochastic optimisation), showing that in a neighborhood of a local minimizer the excess risk is essentially controlled by an *effective dimension* $\operatorname{tr}\big(\nabla^2 f(\nabla^2 f + \lambda I)^{-1}\big)$, so flatter minima (more small-curvature directions) yield tighter generalisation control than ambient-dimension bounds. They further characterize how the Gibbs density allocates probability mass across minima, and in the low-temperature limit the selection biases toward broader basins (over global minima, probabilities scale like $1/\det(\nabla^2 f)$), making a direct connection between flatness/volume and which solutions are ultimately favored.

Thomas et al. (2020) look at the impact of the effective dimension on optimisation (they provide optimisation error bound), obtaining bound on the error that scale with $\operatorname{tr}(\Sigma PHP)$, similarly as in our paper. In addition they empirically study the correlation between empirical estimate of generalisation error δf_S. They find that (P, H, Σ) have effect on both optimisation and generalisation

gap. In this paper, we theoretically show that this is indeed the case, by proving matching upper and lower bounds on the excess risk (which involves effect of both), in terms of $\mathrm{tr}(\Sigma PHP)$.

Relevance of quadratic model approximation. Although globally non-convex, deep networks are effectively modeled by local quadratic approximations. This surrogate is justified theoretically by the *neural tangent kernel* regime, where wide networks remain close to initialisation (Jacot et al., 2018; Du et al., 2019), and validated empirically to track realistic training dynamics (Lee et al., 2020). As such, the quadratic noisy model serves as the standard framework for analyzing preconditioning and generalisation in deep learning (Martens, 2020; Thomas et al., 2020; Zhang et al., 2024).

Connection to information geometry. The setting we analyse can be understood as a natural gradient descent under misspecified model. It is well known, that when $\mathrm{P}_x = \mathrm{Q}$, the two quantities, the variance of gradients Σ and the expected Hessian $\mathbb{E}_z[\nabla^2 \ell(x, z)]$ coincide and equal to the Fisher Information Matrix (FIM):

$$F_{\mathrm{P}_x}(x) \coloneqq \mathbb{E}_{z\sim \mathrm{P}_x}\left[\nabla^2 \ell(x; z)\right] = \mathrm{Var}_{z\sim \mathrm{P}_x}\left[\nabla \ell(x, z)\right],$$

which is a consequence of simple integration of parts. The classical result of Amari (1998) states that when $\mathrm{Q} = \mathrm{P}_x$, in asymptotic regime, locally around $\tilde{x}$, and for single pass setting, the optimal choice of the preconditioning matrix is $P = F_{\mathrm{P}_x}(x)$. However, in our setup, these do not coincide

$$\mathbb{E}_{z\sim \mathrm{Q}}\left[\nabla^2 \ell(x; z)\right] \approx H \neq \mathrm{Var}_{z\sim \mathrm{Q}}\left[\nabla \ell(x, z)\right] =: \Sigma,$$

where the $\approx$ denotes $\alpha H \preceq \nabla^2 \ell(x; z) \preceq \beta H$. This is better corresponding to the practical scenario, when in general, the model is almost always misspecified, i.e. $\mathrm{Q} \neq \mathrm{P}_x$.

Additional examples of spectrally aligned constants.

Example 3 (Regularized logistic regression) *In logistic regression, curvature of H is directly related to the data distribution. Let $\ell(w, z) \coloneqq \log\big(1 + \exp(-y\, a^\top w)\big) + \frac{\lambda}{2}\|w\|_2^2$ for $\lambda > 0$.*

Since $\sigma(t)(1 - \sigma(t)) \leq 1/4$, the Hessian satisfies $\nabla^2 \ell(w; z) = aa^\top \sigma(1 - \sigma) + \lambda I \preceq \frac{1}{4}\, aa^\top + \lambda I$ and we can choose $H \coloneqq \frac{1}{4}\,\mathbb{E}[aa^\top] + \lambda I$[6]. Then one may take $\beta = 1$, and using $\nabla^2 \ell(w; z) \succeq \lambda I$ we have

$$\nabla^2 \ell(w; z) \;\succeq\; \lambda I \;\succeq\; \frac{\lambda}{\lambda + \frac{1}{4}\lambda_{\max}(\mathbb{E}[aa^\top])}\, H,$$

$$\textit{so we can choose} \quad \alpha \;\geq\; \frac{\lambda}{\lambda + \frac{1}{4}\lambda_{\max}(\mathbb{E}[aa^\top])} \quad\Longrightarrow\quad \kappa_\ell = \frac{\beta}{\alpha} \;\leq\; 1 + \frac{\lambda_{\max}(\mathbb{E}[aa^\top])}{4\lambda}.$$

Hence $\rho_\ell = \frac{\sqrt{\kappa_\ell}+1}{\sqrt{\kappa_\ell}-1}$ is explicit. Combining the above bound on κ_ℓ with any explicit bound on $\kappa(PH)$ (e.g., Examples 1–2) immediately yields an explicit $C_{\ell,P}$.

6. We rescale H if needed so that $\lambda_{\max}(H) = 1$

Appendix B. Lemmata and proofs for relative co-coercivity and contractivity

B.1. Proof of Lemma 4

Proof Fix $x, y \in \mathbb{R}^d$ and denote the parameter difference by $u := x - y$ and the gradient difference by $v := \nabla f(x) - \nabla f(y)$. Define the $H^{1/2}$-transformed coordinates

$$\tilde{u} := H^{1/2}u, \qquad \tilde{v} := H^{-1/2}v,$$

and the symmetric positive definite matrix $S := H^{1/2}PH^{1/2}$. Since PH is similar to S, as $S = H^{1/2}(PH)H^{-1/2}$, they share the same spectrum. Let $m := \lambda_{\min}(S) = \lambda_{\min}(PH)$ and $M := \lambda_{\max}(S) = \lambda_{\max}(PH)$, so that $\kappa(PH) = M/m$.

The preconditioned inner product can be written as

$$\langle u, HPv\rangle \;=\; u^\top HPv \;=\; (H^{1/2}u)^\top (H^{1/2}PH^{1/2})(H^{-1/2}v) \;=\; \tilde{u}^\top S\tilde{v}. \tag{6}$$

We can express $S = \bar{\sigma}I + E$, where $\lambda_{\max}(E) \le \delta$ for $\bar{\sigma} := \frac{M+m}{2}$ and $\delta := \frac{M-m}{2}$. We expand (6) using the decomposition of S to get

$$\langle u, HPv\rangle = \bar{\sigma}\,\tilde{u}^\top\tilde{v} + \tilde{u}^\top E\tilde{v} = \bar{\sigma}\langle u, v\rangle + \tilde{u}^\top E\tilde{v}, \tag{7}$$

where we used $\tilde{u}^\top\tilde{v} = u^\top v = \langle u, v\rangle$.

The first term is lower bounded using the standard co-coercivity inequality for functions that are α-strongly convex and β-smooth w.r.t. $\|\cdot\|_H$:

$$\langle u, v\rangle \;\ge\; \frac{\alpha\beta}{\alpha+\beta}\|u\|_H^2 \;+\; \frac{1}{\alpha+\beta}\|v\|_{H^{-1}}^2. \tag{8}$$

The second perturbation term is bounded by Cauchy–Schwarz and $\lambda_{\max}(E) \le \delta$,

$$\tilde{u}^\top E\tilde{v} \;\ge\; -\lambda_{\max}(E)\|\tilde{u}\|_2\|\tilde{v}\|_2 \;\ge\; -\delta\|u\|_H\|v\|_{H^{-1}}. \tag{9}$$

Combining (7)–(9) yields

$$\langle u, HPv\rangle \ge \bar{\sigma}\left(\frac{\alpha\beta}{\alpha+\beta}\|u\|_H^2 + \frac{1}{\alpha+\beta}\|v\|_{H^{-1}}^2\right) - \delta\|u\|_H\|v\|_{H^{-1}}. \tag{10}$$

In order to remove the cross term, we apply AM–GM in the form

$$\|u\|_H\|v\|_{H^{-1}} = \frac{1}{\sqrt{\alpha\beta}}\big(\sqrt{\alpha\beta}\|u\|_H\big)\|v\|_{H^{-1}} \le \frac{1}{2\sqrt{\alpha\beta}}\Big(\alpha\beta\|u\|_H^2 + \|v\|_{H^{-1}}^2\Big).$$

Substituting into (10) and factoring the standard co-coercivity expression gives

$$\begin{aligned}\langle u, HPv\rangle &\ge \left[\bar{\sigma} - \delta\cdot\frac{\alpha+\beta}{2\sqrt{\alpha\beta}}\right]\left(\frac{\alpha\beta}{\alpha+\beta}\|u\|_H^2 + \frac{1}{\alpha+\beta}\|v\|_{H^{-1}}^2\right) \\ &= \left[\frac{M+m}{2} - \frac{M-m}{4}\cdot\frac{\alpha+\beta}{\sqrt{\alpha\beta}}\right]\left(\frac{\alpha\beta}{\alpha+\beta}\|u\|_H^2 + \frac{1}{\alpha+\beta}\|v\|_{H^{-1}}^2\right).\end{aligned} \tag{11}$$

It remains to simplify the constant. Let $\kappa_f := \beta/\alpha$ and note that

$$\frac{\alpha+\beta}{\sqrt{\alpha\beta}} = \frac{\kappa_f+1}{\sqrt{\kappa_f}}, \qquad \rho_\ell := \frac{\sqrt{\kappa_f}+1}{\sqrt{\kappa_f}-1}.$$

We can express $M = m\,\kappa(PH)$, which after an algebraic manipulation of the bracket in (11) yields

$$\frac{M+m}{2} - \frac{M-m}{4} \cdot \frac{\kappa_f + 1}{\sqrt{\kappa_f}} \;=\; m \cdot \frac{\rho_\ell^2 - \kappa(PH)}{\rho_\ell^2 - 1} \;=\; \lambda_{\min}(PH) \cdot C_{\ell,P}. \tag{12}$$

Under the assumption $\kappa(PH) < \rho_\ell^2$, we have $C_{\ell,P} \in (0, 1]$. Substituting (12) into (11) and recalling $u = x - y$, $v = \nabla f(x) - \nabla f(y)$ completes the proof. ■

Lemma 19 (Contractivity of the preconditioned update M_θ-norm) *Suppose that Assumption 5 and 9 hold for $\ell(\cdot, z)$. Let $P \succ 0$ and $M_\theta := H^{1/2}(H^{1/2}PH^{1/2})^{-\theta}H^{1/2}$ for $\theta \in [0, 1]$. For $\kappa(PH)^{1-\theta} \le \rho_\ell^2$ the preconditioned gradient update $x^+ = x - \eta P\nabla\ell(x, z)$ is r-contractive in the $\|\cdot\|_{M_\theta}$, where*

$$r = 2\lambda_{\min}(PH)\, C_{\ell,P}^{(\theta)} \frac{\alpha\beta}{\alpha+\beta} \quad \textit{and} \quad C_{\ell,P}^{(\theta)} := \frac{\rho_\ell^2 - \kappa(PH)^{1-\theta}}{\rho_\ell^2 - 1},$$

provided the step size satisfies $\eta_t \le 2\, C_{\ell,P}^{(\theta)} / \big(\lambda_{\max}(PH)\kappa(PH)^{1-\theta}(\alpha+\beta)\big)$.

Proof Let $u_t := x_t - y_t$ and $v_t := \nabla f(x_t) - \nabla f(y_t)$. The update gives $u_{t+1} = u_t - \eta_t P v_t$. We analyze the squared M_θ-norm:

$$\begin{aligned}\|u_{t+1}\|_{M_\theta}^2 &= \langle u_t - \eta_t P v_t,\, M_\theta(u_t - \eta_t P v_t)\rangle \\ &= \|u_t\|_{M_\theta}^2 - 2\eta_t\langle v_t,\, PM_\theta u_t\rangle + \eta_t^2\langle v_t,\, PM_\theta P v_t\rangle.\end{aligned}$$

We bound the terms separately. Let $S = H^{1/2}PH^{1/2}$ and introduce the $H^{1/2}$-transformed variables

$$\tilde{u}_t := H^{1/2}u_t, \qquad \tilde{v}_t := H^{-1/2}v_t.$$

Since $M_\theta = H^{1/2}S^{-\theta}H^{1/2}$ and $P = H^{-1/2}SH^{-1/2}$, we have

$$\langle v_t,\, PM_\theta u_t\rangle = \tilde{v}_t^\top S^{1-\theta}\tilde{u}_t, \qquad \langle v_t,\, PM_\theta P v_t\rangle = \tilde{v}_t^\top S^{2-\theta}\tilde{v}_t.$$

Cross term. The matrix $S^{1-\theta}$ has eigenvalues in $[\lambda_{\min}(PH)^{1-\theta}, \lambda_{\max}(PH)^{1-\theta}]$. Applying the same decomposition argument as in Lemma 4 (with $S^{1-\theta}$ in place of S) yields

$$\tilde{v}_t^\top S^{1-\theta}\tilde{u}_t \;\ge\; \lambda_{\min}(PH)^{1-\theta}\, C_{\ell,P}^{(\theta)} \left(\frac{\alpha\beta}{\alpha+\beta}\|u_t\|_H^2 + \frac{1}{\alpha+\beta}\|v_t\|_{H^{-1}}^2\right).$$

Quadratic term. Since $S^{2-\theta} \preceq \lambda_{\max}(PH)^{2-\theta} I$, we have

$$\langle v_t,\, PM_\theta P v_t\rangle \le \lambda_{\max}(PH)^{2-\theta}\|v_t\|_{H^{-1}}^2.$$

Combine. Substituting the bounds into the expansion gives

$$\|u_{t+1}\|_{M_\theta}^2 \le \|u_t\|_{M_\theta}^2 - 2\eta_t \lambda_{\min}(PH)^{1-\theta} C_{\ell,P}^{(\theta)} \frac{\alpha\beta}{\alpha+\beta} \|u_t\|_H^2 + \eta_t \left(\eta_t \lambda_{\max}(PH)^{2-\theta} - \frac{2\lambda_{\min}(PH)^{1-\theta} C_{\ell,P}^{(\theta)}}{\alpha+\beta} \right) \|v_t\|_{H^{-1}}^2.$$

The gradient term is non-positive provided

$$\eta_t \le \frac{2\,\lambda_{\min}(PH)^{1-\theta}\, C_{\ell,P}^{(\theta)}}{\lambda_{\max}(PH)^{2-\theta}(\alpha+\beta)}.$$

Under this condition, dropping the negative term and using $\|u_t\|_H^2 \ge \lambda_{\min}(PH)^\theta \|u_t\|_{M_\theta}^2$ yields

$$\|u_{t+1}\|_{M_\theta}^2 \le \left(1 - 2\eta_t \lambda_{\min}(PH)\, C_{\ell,P}^{(\theta)} \frac{\alpha\beta}{\alpha+\beta}\right) \|u_t\|_{M_\theta}^2.$$

■

Appendix C. Lemmata and proofs for stability results

C.1. Proof of Lemma 8

Proof Let x_t and y_t be the iterate sequences of PSGD on datasets S and $S^{(i)} = S \setminus \{z_i\} \cup \{z'\}$ respectively. We analyze the evolution of the expected squared parameter distance $\delta_t \coloneqq \mathbb{E}_{\mathcal{A},S,z'}[\|x_t - y_t\|_M^2]$.

At iteration t, let j be the index of the sample selected by the algorithm $\mathcal{A}$. With probability $1 - 1/n$, $j \ne i$ (the samples match), and with probability $1/n$, $j = i$ (the samples differ). Using the linearity of expectation:

$$\delta_{t+1} = \left(1 - \frac{1}{n}\right) \mathbb{E}_{j\ne i}[\|x_{t+1} - y_{t+1}\|_M^2] + \frac{1}{n} \mathbb{E}_{j=i}[\|x_{t+1} - y_{t+1}\|_M^2]. \tag{13}$$

For the matching sample case ($j \ne i$), we use the contractivity of the PSGD update by assumption of the lemma

$$\mathbb{E}_{j\ne i}[\|x_{t+1} - y_{t+1}\|_M^2] \le (1 - \eta_t r)\delta_t.$$

For the differing sample case ($j = i$), denote the parameter difference as $\Delta_t \coloneqq x_t - y_t$, the gradient difference as $\xi_t \coloneqq P(\nabla\ell(x_t, z_i) - \nabla\ell(y_t, z'))$, and the population gradient difference as $\tilde{\xi}_t \coloneqq P(\nabla f(x_t) - \nabla f(y_t))$. We apply Young's inequality: $\|u+v\|_M^2 \le (1+\alpha)\|u\|_M^2 + (1+\frac{1}{\alpha})\|v\|_M^2$ for $\alpha > 0$ to be chosen later, and expand the update as

$$\delta_{t+1} = \|\Delta_{t+1}\|_M^2 = (1+\alpha)\|\Delta_t - \eta_t\tilde{\xi}_t\|_M^2 + \left(1 + \frac{1}{\alpha}\right)\|\eta_t(\xi_t - \tilde{\xi}_t)\|_M^2 \tag{14}$$

$$\le (1+\alpha)(1-\eta_t r)\delta_t + \left(1 + \frac{1}{\alpha}\right)\eta_t^2 \|\xi_t - \tilde{\xi}_t\|_M^2, \tag{15}$$

where for the inequality we used that the preconditioned update is contractive w.r.t the population gradient by assumption of the lemma.

Combining these terms yields the recursion:

$$\delta_{t+1} \le (1-\eta_t r)\left(1+\frac{\alpha}{n}\right)\delta_t + \left(1+\frac{1}{\alpha}\right)\frac{\eta_t^2}{n}\mathbb{E}[\|\xi_t - \tilde{\xi}_t\|_M^2].$$

To bound the gradient variance we add and subtract $\nabla\ell(y_t, z_i) - \nabla\ell(x_t, z')$

$$\begin{aligned}
\mathbb{E}[\|\xi_t - \tilde{\xi}_t\|_M^2] &= \mathbb{E}[\|\nabla\ell(x_t,z_i) - \nabla f(x_t) - (\nabla\ell(y_t,z') - \nabla f(y_t))\|_{PMP}^2] \\
&\le 4\mathbb{E}\|\nabla\ell(x_t,z_i) - \nabla\ell(y_t,z_i)\|_{PMP}^2 + 4\mathbb{E}\|\nabla\ell(y_t,z') - \nabla\ell(x_t,z')\|_{PMP}^2 \\
&\quad + 4\mathbb{E}\|\nabla\ell(x_t,z') - \nabla f(x_t)\|_{PMP}^2 + 4\mathbb{E}\|\nabla\ell(y_t,z') - \nabla f(y_t)\|_{PMP}^2 \\
&\le 8\,\mathrm{tr}(PMP\Sigma) + 8\beta^2\lambda_{\max}(HPMP)\lambda_{\max}(M^{-1}H)\delta_t,
\end{aligned}$$

and in the second inequality we bound the gradient difference using Jensen's inequality combined with the bounded variance assumption $\mathrm{Var}[\nabla\ell] \preceq \Sigma$, and smoothness to bound the cross terms $\|\nabla\ell(x_t,z_i) - \nabla\ell(y_t,z_i)\|_{PMP}$ and $\mathbb{E}\|\nabla\ell(y_t,z') - \nabla\ell(x_t,z')\|_{PMP}$.

Denote $\gamma^2 = \lambda_{\max}(HPMP)\lambda_{\max}(M^{-1}H)$ and $\tau^2 = \mathrm{tr}(PMP\Sigma)$ and substitute the δ_{t+1} bound, we obtain:

$$\delta_{t+1} \le \underbrace{\left[(1-\eta_t r)\left(1+\frac{\alpha}{n}\right) + \frac{8\eta_t^2\beta^2\gamma^2}{n}\left(1+\frac{1}{\alpha}\right)\right]}_{A_t}\delta_t + \underbrace{\frac{8\eta_t^2\tau^2}{n}\left(1+\frac{1}{\alpha}\right)}_{B_t} \tag{16}$$

We set $\alpha = \frac{n\eta_t r}{2}$, and express A_t and B_t. The first term is

$$\begin{aligned}
A_t &= (1-\eta_t r)\left(1+\frac{\eta_t r}{2}\right) + \frac{8\eta_t^2\beta^2\gamma^2}{n}\left(1+\frac{2}{n\eta_t r}\right) \\
&= 1 - \frac{\eta_t r}{2} - \frac{\eta_t^2 r^2}{2} + \frac{8\eta_t^2\beta^2\gamma^2}{n} + \frac{16\eta_t\beta^2\gamma^2}{n^2 r} \\
&\le 1 - \frac{\eta_t r}{4}.
\end{aligned}$$

For the last inequality we use the step-size cap $\eta_t\beta\gamma \le 1$ together with $n \ge 34\,\beta\gamma/r$. The cap bounds the quadratic step term linearly, $\frac{8\eta_t^2\beta^2\gamma^2}{n} = \frac{8\eta_t\beta\gamma}{n}(\eta_t\beta\gamma) \le \frac{8\eta_t\beta\gamma}{n} \le \frac{8\eta_t r}{34}$, while $\frac{16\eta_t\beta^2\gamma^2}{n^2 r} \le \frac{16\eta_t r}{34^2}$, both using $n \ge 34\,\beta\gamma/r$. Summing, the two positive terms are at most $(\frac{8}{34} + \frac{16}{34^2})\eta_t r \le \frac{\eta_t r}{4}$, which with $-\frac{\eta_t r}{2}$ gives the claim. The term B_t becomes

$$B_t = \frac{8\eta_t^2\tau^2}{n} + \frac{16\eta_t\tau^2}{n^2 r}.$$

In particular, we have a recursion of the form,

$$\delta_{t+1} \le (1-\eta_t r/4)\delta_t + B_t \le \exp(-\eta_t r/4)\delta_t + B_t.$$

Thus, we obtain that,

$$\delta_t \le \exp(-T_t r/4)\delta_0 + \sum_{s<t}\exp(-(T_t - T_s)r/4)B_s.$$

Since the second term produces a lower Riemann approximation of the integral of $\exp(-(T-s)r/4)$, we obtain,

$$\frac{16\tau^2}{n^2 r}\sum_{s<t}\exp(-(T_t-T_s)r/4)\eta_s \le \frac{16\tau^2}{n^2 r}\int_0^{T_t}\exp(-(T_t-s)r/4)ds$$
$$\le (1-\exp(-T_t r/4))\frac{64\tau^2}{n^2 r^2},$$

Then, using $\delta_0 = 0$ and the definition of $\bar{\eta}_t$ in Lemma 24, we obtain the bound,

$$\delta_t \le (1-\exp(-T_t r/4))\frac{64\tau^2}{n^2 r^2} + \frac{8\bar{\eta}_t\tau^2}{n}.$$

■

C.2. Proof of Lemma 7

Proof We decompose the excess population risk as

$$f(x_t) - f(\tilde{x}) = \underbrace{f(x_t) - f_S(x_t)}_{\text{generalization error}} + \underbrace{f_S(x_t) - f_S(x_S^*)}_{\text{optimization error}} + \underbrace{f_S(x_S^*) - f(\tilde{x})}_{\le 0 \text{ in expectation}},$$

where $x_S^* = \arg\min f_S(x)$. Taking the expectation over S and the randomness of $\mathcal{A}$, we get

$$\delta f(x_t) \coloneqq \mathbb{E}_{\mathcal{A},S}[f(x_t) - f(\tilde{x})] \le \underbrace{\mathbb{E}_{\mathcal{A},S}[f(x_t) - f_S(x_t)]}_{\text{expected generalization error}} + \underbrace{\mathbb{E}_{\mathcal{A},S}[f_S(x_t) - f_S(x_S^*)]}_{\varepsilon_{\text{opt}}(x_t) = \text{expected optimization error}}.$$

Let $z' \sim \mathrm{Q}$ be an independent sample and let $S^{(i)} \coloneqq S \setminus \{z_i\} \cup \{z'\}$ denote the perturbed dataset. Write $x_t(S)$ and $x_t(S^{(i)})$ for the corresponding PSGD iterates. The standard symmetrization argument yields

$$\mathbb{E}_{\mathcal{A},S}[f(x_t) - f_S(x_t)] = \mathbb{E}_{\mathcal{A},S}\left[\mathbb{E}_{z'}\ell(x_t(S), z') - \frac{1}{n}\sum_{i=1}^n \ell(x_t(S), z_i)\right]$$
$$= \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{S,z',\mathcal{A}}\left[\ell(x_t(S), z') - \ell(x_t(S^{(i)}), z')\right], \tag{17}$$

where the second equality uses that (S, z_i) has the same distribution as $(S^{(i)}, z')$. Fix i. By the β-smoothness of $\ell(\cdot, z')$ w.r.t. $\|\cdot\|_H$, we have

$$\mathbb{E}[\ell(x_t(S), z') - \ell(x_t(S^{(i)}), z')] \le \mathbb{E}\left[\langle\nabla\ell(x_t(S), z'),\, x_t(S) - x_t(S^{(i)})\rangle\right] + \frac{\beta}{2}\mathbb{E}\left[\left\|x_t(S) - x_t(S^{(i)})\right\|_H^2\right]$$
$$\le \mathbb{E}\left[\|\nabla\ell(x_t(S), z')\|_{M^{-1}}^2\right]^{1/2}\mathbb{E}\left[\|x_t(S) - x_t(S^{(i)})\|_M^2\right]^{1/2}$$
$$+ \frac{\beta}{2}\mathbb{E}\left[\left\|x_t(S) - x_t(S^{(i)})\right\|_H^2\right], \tag{18}$$

where we applied the Cauchy-Schwarz inequality w.r.t. the $\|\cdot\|_M$ norm.

By the bias-variance decomposition and the assumption $\mathrm{Var}_z[\nabla\ell(x,z)] \preceq \Sigma$, we express the gradient factor in (18) as

$$\begin{aligned}\mathbb{E}_{z'}\|\nabla\ell(x_t(S),z')\|_{M^{-1}}^2 &\le \|\nabla f(x_t(S))\|_{M^{-1}}^2 + \mathrm{tr}(M^{-1}\Sigma)\\ &\le \lambda_{\max}(HM^{-1})\|\nabla f(x_t(S))\|_{H^{-1}}^2 + \mathrm{tr}(M^{-1}\Sigma)\\ &\le 2\beta\lambda_{\max}(HM^{-1})(f(x_t(S)) - f(\tilde{x})) + \mathrm{tr}(M^{-1}\Sigma),\end{aligned}$$

where the second inequality follows from matrix operator bounds and the third from the smoothness of the population risk. Taking an expectation over S and $\mathcal{A}$ yields

$$\mathbb{E}_{S,z',\mathcal{A}}\|\nabla\ell(x_t(S),z')\|_{M^{-1}}^2 \le 2\beta\lambda_{\max}(HM^{-1})\delta f(x_t) + \mathrm{tr}(M^{-1}\Sigma).$$

Using the parameter stability assumption $\mathbb{E}[\|x_t(S) - x_t(S^{(i)})\|_M^2] \le \varepsilon_{\text{pstab}}^2$ and the norm inequality $\|v\|_H^2 \le \lambda_{\max}(M^{-1}H)\|v\|_M^2$, we substitute back into (17) and (18):

$$\delta f(x_t) - \delta f_S(x_t) \le \big(2\beta\lambda_{\max}(HM^{-1})\delta f(x_t) + \mathrm{tr}(M^{-1}\Sigma)\big)^{1/2}\varepsilon_{\text{pstab}} + \frac{\beta\lambda_{\max}(M^{-1}H)\varepsilon_{\text{pstab}}^2}{2}.$$

This inequality is of the form $Y \le \sqrt{AY+B}\,\varepsilon_{\text{pstab}} + C$, where $Y = \delta f(x_t)$, $A = 2\beta\lambda_{\max}(HM^{-1})$, $B = \mathrm{tr}(M^{-1}\Sigma)$, and $C = \frac{\beta\lambda_{\max}(M^{-1}H)\varepsilon_{\text{pstab}}^2}{2} = \frac{A\varepsilon_{\text{pstab}}^2}{4}$. Using the sub-additivity of the square root $\sqrt{AY+B} \le \sqrt{A}\sqrt{Y} + \sqrt{B}$, this becomes a quadratic inequality in $\sqrt{Y}$,

$$Y \;\le\; \sqrt{A}\,\varepsilon_{\text{pstab}}\sqrt{Y} + \big(\delta f_S(x_t) + C + \sqrt{B}\,\varepsilon_{\text{pstab}}\big).$$

Solving for $\sqrt{Y}$ via the quadratic formula and using $(p+q)^2 \le 2p^2 + 2q^2$ yields

$$\delta f(x_t) \;\le\; A\,\varepsilon_{\text{pstab}}^2 + 2\big(\delta f_S(x_t) + C + \sqrt{B}\,\varepsilon_{\text{pstab}}\big) \;=\; 2\,\delta f_S(x_t) + \tfrac{3}{2}A\,\varepsilon_{\text{pstab}}^2 + 2\sqrt{B}\,\varepsilon_{\text{pstab}}.$$

Bounding $\frac{3}{2}A \le 2A$ gives the form stated in the lemma,

$$\delta f(x_t) \;\le\; 2\,\delta f_S(x_t) + 4\beta\lambda_{\max}(HM^{-1})\,\varepsilon_{\text{pstab}}^2 + 2\,\mathrm{tr}(M^{-1}\Sigma)^{1/2}\,\varepsilon_{\text{pstab}}.$$

■

Corollary 20 (Optimal choice of P) *We have that* $P \coloneqq \lambda_{\min}(H)H^{-1} = \arg\min_{P\succ 0}\mathrm{tr}(P\Sigma)/\lambda_{\min}(PH)$.

Proof Let $A = H^{1/2}PH^{1/2}$, so we have $P = H^{-1/2}AH^{-1/2}$. By definition of A we have that $PH = H^{-1/2}AH^{1/2}$, thus PH is similar to A, which means their eigenvalues are equal, which implies that $\lambda_{\min}(PH) = \lambda_{\min}(A)$. Furthermore, define $\hat{\Sigma} \coloneqq H^{-1/2}\Sigma H^{-1/2}$ for which we get by cyclicality of the trace that $\mathrm{tr}(P\Sigma) = \mathrm{tr}(A\hat{\Sigma})$. Denoting $\sum_{i=1}^d a_i v_i v_i^\top$ to be the spectral decomposition of A, the objective becomes

$$\begin{aligned}\frac{\mathrm{tr}(A\hat{\Sigma})}{\lambda_{\min}(A)} &= \frac{1}{a_d}\,\mathrm{tr}\left(\left(\sum_{i=1}^d a_i v_i v_i^\top\right)\hat{\Sigma}\right)\\ &= \sum_{i=1}^d \frac{a_i}{a_d} v_i^\top\hat{\Sigma}v_i \ge \mathrm{tr}(\hat{\Sigma}),\end{aligned}$$

where the lower bound comes from the fact that all $a_i/a_d \geq 1$ and is attained when $a_i = a$ for all $i \in [d]$. By A being symmetric this happens when $A = aI$.

From the requirement that $\lambda_{\max}(P) = 1$ we get

$$1 = \lambda_{\max}(P) = \lambda_{\max}(H^{-1/2} a I H^{-1/2}) = \frac{a}{\lambda_{\min}(H)},$$

implying that $a = \lambda_{\min}(H)$. Substituting $A = \lambda_{\min}(H) I$ into the formula for P yields that the optimal $P = \lambda_{\min}(H) H^{-1}$. ■

Appendix D. Optimization error bounds results

Lemma 21 (Preconditioned PL-Growth Condition) *Let f be α-strongly convex and β-smooth w.r.t. $\|\cdot\|_H$. Let x^* be the global minimizer. If the preconditioner is spectrally aligned, i.e., $\kappa(PH) < \rho^2 := \frac{\sqrt{\kappa_f}+1}{\sqrt{\kappa_f}-1}$, then, for all $x \in \mathbb{R}^d$:*

$$\langle x - x^*, HP\nabla f(x)\rangle \geq \frac{2\alpha}{\alpha+\beta}\lambda_{\min}(PH) C_{\ell,P}\left(f(x) - f(x^*) + \frac{\beta}{2}\|x - x^*\|_H^2\right), \tag{19}$$

where the co-coercivity constant $C_{\ell,P}$ is given in the statement of Lemma 4.

Proof Let $u = x - x^*$ and $v = \nabla f(x)$. Note that $\nabla f(x^*) = 0$. By Lemma 4 and the condition $\kappa(PH) < \rho^2$, we have

$$\langle u, HPv\rangle \geq C_{\ell,P}\lambda_{\min}(PH)\left(\frac{\alpha\beta}{\alpha+\beta}\|u\|_H^2 + \frac{1}{\alpha+\beta}\|v\|_{H^{-1}}^2\right). \tag{20}$$

Since f is α-strongly convex, it satisfies the Polyak-Łojasiewicz (PL) inequality w.r.t. the H-norm:

$$\|\nabla f(x)\|_{H^{-1}}^2 \geq 2\alpha(f(x) - f(x^*)).$$

We substitute this lower bound for the gradient norm term in (20):

$$\begin{aligned}\langle u, HPv\rangle &\geq C_{\ell,P}\lambda_{\min}(PH)\left(\frac{\alpha\beta}{\alpha+\beta}\|u\|_H^2 + \frac{2\alpha}{\alpha+\beta}(f(x) - f(x^*))\right)\\ &= C_{\ell,P}\lambda_{\min}(PH)\frac{2\alpha}{\alpha+\beta}\left(\frac{\beta}{2}\|u\|_H^2 + f(x) - f(x^*)\right).\end{aligned}$$

■

The following lemma allows to relate Σ_S to Σ.

Lemma 22 *Assume that $x \mapsto \ell(x, z)$ is L-Lipschitz for any z. Then, for any $i \in [n]$, under conditions of Lemma 8,*

$$\|\mathrm{Var}(\nabla\ell(x_t, z_{i_t})) - \mathrm{Var}(\nabla\ell(x_t, z))\|_2 \leq 32L\beta\sqrt{\left(\frac{\bar{\eta}_t}{8n} + \frac{1 - e^{-T_t r/4}}{n^2 r^2}\right)\mathrm{tr}(PMP\Sigma)}\,.$$

Proof Note that $\mathbb{E}[\nabla\ell(x_t, z_{i_t}) \mid x_t] = \nabla f_S(x_t)$ while $\mathbb{E}[\nabla\ell(x_t, z) \mid x_t] = \nabla f(x_t)$. Now,

$$\begin{aligned}
&\|\mathrm{Var}(\nabla\ell(x_t, z_{i_t})) - \mathrm{Var}(\nabla\ell(x_t, z))\|_2 \\
&\quad\le \|\mathbb{E}[\nabla\ell(x_t, z_{i_t})\nabla\ell(x_t, z_{i_t})^\top - \nabla\ell(x_t, z)\nabla\ell(x_t, z)^\top]\|_2 \\
&\quad+ \|\mathbb{E}[\nabla f(x_t)\nabla f(x_t)^\top - \nabla f_S(x_t)\nabla f_S(x_t)^\top]\|_2 \\
&\quad= \|\mathbb{E}[\nabla\ell(x_t^{(i)}, z)\nabla\ell(x_t^{(i)}, z)^\top - \nabla\ell(x_t, z)\nabla\ell(x_t, z)^\top]\|_2 && \text{(Here } i \equiv i_t\text{)} \\
&\quad+ \|\mathbb{E}[\nabla f(x_t)\nabla f(x_t)^\top - \nabla f_S(x_t)\nabla f_S(x_t)^\top]\|_2
\end{aligned}$$

We first bound the first term on the r.h.s. by observe that for any unit vector u

$$\begin{aligned}
&\mathbb{E}\left\langle u, \nabla\ell(x_t^{(i)}, z)\right\rangle^2 - \mathbb{E}[\langle u, \nabla\ell(x_t, z)\rangle^2 \\
&\quad= \mathbb{E}\left[\left\langle u, \nabla\ell(x_t^{(i)}, z) - \nabla\ell(x_t, z)\right\rangle\left\langle u, \nabla\ell(x_t^{(i)}, z) + \nabla\ell(x_t, z)\right\rangle\right] \\
&\quad\le 2L\,\mathbb{E}\left[\left|\left\langle u, \nabla\ell(x_t^{(i)}, z) - \nabla\ell(x_t, z)\right\rangle\right|\right] \\
&\quad\le 2L\,\mathbb{E}\left[\|\nabla\ell(x_t^{(i)}, z) - \nabla\ell(x_t, z)\|_{H^{-1}}\right] && \text{(Since } H \succ 0\text{)} \\
&\quad\le 2L\,\beta\,\mathbb{E}[\|x_t - x_t^{(i)}\|_H] && (\ell \text{ is } L\text{-Lipschitz}) \\
&\quad\le 2L\beta\sqrt{64\left(\frac{\bar\eta_t}{8n} + \frac{1 - e^{-T_t r/4}}{n^2 r^2}\right)\mathrm{tr}(PMP\Sigma)}
\end{aligned}$$

by Lemma 8. The same chain of inequalities hold for the second term. ■

Lemma 23 (Optimization rate of PSGD under PL and smoothness) *Let $P \succ 0$. Let $f_S(x) = \frac{1}{n}\sum_{i=1}^n \ell(x, z_i)$ be μ-PL and β-smooth w.r.t. $\|\cdot\|_H$ and it attains its minimal value $f_S^* = \min_{x\in\mathcal{X}} f_S(x)$. Let $\Sigma_S \succ 0$, so that $\mathrm{Var}\left[\nabla\ell(x_t, z_{i_t}) \,|\, x_t\right] \preceq \Sigma_S$. Then, for $\eta_t \le \frac{1}{\beta\lambda_{\max}(PH)}$ the expected empirical excess optimization error is bounded as*

$$\mathbb{E}_{\mathcal{A}}[f_S(x_t) - f_S^*] \le e^{-\lambda_{\min}(PH)\mu T_t}(f_S(x_0) - f_S^*) + \frac{\beta}{2}\,\mathrm{tr}(PHP\Sigma_S)\bar\eta_t,$$

where T_s and $\bar\eta_t$ are defined as in the statement of Lemma 8.

Proof Define the suboptimality process $\phi_t \coloneqq \mathbb{E}_{\mathcal{A}}[f_S(x_t) - f_S^*]$. Let $g_t \coloneqq \nabla\ell(x_t, z_{i_t})$ denote the stochastic gradient and note that $\mathbb{E}[g_t \mid x_t] = \nabla f_S(x_t)$. By β-smoothness of f_S w.r.t. $\|\cdot\|_H$

$$f_S(x_{t+1}) \le f_S(x_t) - \eta_t\langle\nabla f_S(x_t), Pg_t\rangle + \frac{\beta\,\eta_t^2}{2}\|Pg_t\|_H^2.$$

Taking conditional expectation given x_t and using $\mathbb{E}[g_t \mid x_t] = \nabla f_S(x_t)$ gives

$$\mathbb{E}[f_S(x_{t+1}) \mid x_t] \le f_S(x_t) - \eta_t\|\nabla f_S(x_t)\|_P^2 + \frac{\beta\,\eta_t^2}{2}\,\mathbb{E}\big[\|Pg_t\|_H^2 \mid x_t\big]. \tag{21}$$

Using the covariance bound and variance-bias decomposition,

$$\mathbb{E}[\|Pg_t\|_H^2 \mid x_t] = \|P\nabla f_S(x_t)\|_H^2 + \mathrm{tr}\big(PHP\,\mathrm{Cov}(g_t \mid x_t)\big) \le \|P\nabla f_S(x_t)\|_H^2 + \mathrm{tr}(PHP\Sigma_S).$$

By $\|P\nabla f_S(x_t)\|_H^2 \le \lambda_{\max}(PH)\|\nabla f_S(x_t)\|_P^2$ and substituting into (21) yields

$$\mathbb{E}[f_S(x_{t+1}) \mid x_t] \le f_S(x_t) - \eta_t\Big(1 - \tfrac{\beta\lambda_{\max}(PH)}{2}\eta_t\Big)\|\nabla f_S(x_t)\|_P^2 + \frac{\beta}{2}\eta_t^2\ \mathrm{tr}(PHP\Sigma_S).$$

Whenever $\eta_t \le \frac{1}{\beta\lambda_{\max}(PH)}$, we have $1 - \frac{\beta\lambda_{\max}(PH)}{2}\eta_t \ge \frac{1}{2}$, and thus

$$\mathbb{E}[f_S(x_{t+1}) - f_S(\hat{x}_*) \mid x_t] \le f_S(x_t) - f_S(\hat{x}_*) - \frac{\eta_t}{2}\|\nabla f_S(x_t)\|_P^2 + \frac{\beta}{2}\eta_t^2\ \mathrm{tr}(PHP\Sigma_S). \quad (22)$$

Since f_S satisfies μ-PL property w.r.t. $\|\cdot\|_H$

$$\frac{1}{\lambda_{\min}(PH)}\|\nabla f_S(x_t)\|_P^2 \ge \|\nabla f_S(x_t)\|_{H^{-1}}^2 \ge 2\mu\big(f_S(x_t) - f_S(\hat{x}_*)\big).$$

Substituting into (22) yields the scalar recursion

$$\mathbb{E}[f_S(x_{t+1}) - f_S(\hat{x}_*) \mid x_t] \le \big(1 - \eta_t\mu\lambda_{\min}(PH)\big)\big(f_S(x_t) - f_S(\hat{x}_*)\big) + \frac{\beta}{2}\eta_t^2\ \mathrm{tr}(PHP\Sigma_S). \quad (23)$$

Taking total expectation gives

$$\phi_{t+1} \le (1 - \eta_t\mu\lambda_{\min}(PH))\phi_t + \eta_t^2 B, \qquad \text{where } B \coloneqq \frac{\beta}{2}\,\mathrm{tr}(PHP\Sigma_S).$$

Denote $a = \mu\lambda_{\min}(PH)$. Using $1 - x \le e^{-x}$ for $x \ge 0$,

$$\phi_{t+1} \le e^{-a\eta_t}\phi_t + \eta_t^2 B$$

and unrolling, using that $T_s = \sum_{s'=0}^{s-1}\eta_{s'}$ and so $T_{t+1} - T_t = \eta_t$, gives

$$\phi_t \le e^{-aT_t}\phi_0 + B\sum_{s=0}^{t-1} e^{-a(T_t - T_s)}\eta_s^2 \le e^{-aT_t}\phi_0 + B\bar{\eta}_t,$$

where the last inequality uses $a = \mu\lambda_{\min}(PH) \ge r/4$, so that $\sum_{s<t} e^{-a(T_t-T_s)}\eta_s^2 \le \sum_{s<t} e^{-r(T_t-T_s)/4}\eta_s^2 = \bar{\eta}_t$. ∎

Lemma 24 (Capped-harmonic bound for $\bar{\eta}_t$) *Let $T_s = \sum_{s'=0}^{s-1}\eta_{s'}$ and $\bar{\eta}_t = \sum_{s<t} e^{-r\frac{T_t-T_s}{4}}\eta_s^2$ as in Lemma 8. Fix $\eta_0 > 0$ and $c > 0$, and define $\eta_t \coloneqq \min\Big\{\eta_0, \frac{c}{t+1}\Big\}$, $t_0 \coloneqq \left\lceil \frac{c}{\eta_0} \right\rceil - 1$, and $\alpha \coloneqq \frac{rc}{4}$. For $\alpha > 1$. Then for every $t \ge t_0 + 1$,*

$$\bar{\eta}_t \ \le\ \frac{C_{\mathrm{burn}} + C_{\mathrm{harm}}}{t+1}, \qquad \forall\, t \ge t_0 + 1,$$

where $C_{\mathrm{harm}} \coloneqq \frac{c^2}{\alpha - 1}$ and $C_{\mathrm{burn}} \coloneqq \eta_0^2\,(t_0+2)^{(\alpha+1)}$.

Proof Fix $t \ge t_0 + 1$ and split the sum defining $\bar{\eta}_t$ into "burn-in" and "harmonic tail" parts:

$$\bar{\eta}_t = \sum_{s=0}^{t_0}\exp\Big(-\frac{r}{4}(T_t - T_s)\Big)\eta_s^2 \ +\ \sum_{s=t_0+1}^{t-1}\exp\Big(-\frac{r}{4}(T_t - T_s)\Big)\eta_s^2.$$

Tail part ($s \geq t_0 + 1$). For $s \geq t_0 + 1$ we have $\eta_k = c/(k+1)$ for all $k \geq s$, hence

$$T_t - T_s = \sum_{k=s}^{t-1} \frac{c}{k+1} \;\geq\; c \int_{s+1}^{t+1} \frac{dx}{x} = c \log\Big(\frac{t+1}{s+1}\Big).$$

Therefore,

$$\exp\Big(-\frac{r}{4}(T_t - T_s)\Big) \leq \exp\Big(-\frac{r}{4}\, c \log \frac{t+1}{s+1}\Big) = \Big(\frac{s+1}{t+1}\Big)^{\alpha}.$$

Using also $\eta_s^2 = c^2/(s+1)^2$ on the tail,

$$\sum_{s=t_0+1}^{t-1} \exp\Big(-\frac{r}{4}(T_t - T_s)\Big)\eta_s^2 \leq \frac{c^2}{(t+1)^{\alpha}} \sum_{s=t_0+1}^{t-1} (s+1)^{\alpha-2}.$$

Since $\alpha > 1$, we can bound the sum by an integral:

$$\sum_{s=t_0+1}^{t-1} (s+1)^{\alpha-2} \leq \int_{t_0+2}^{t+1} x^{\alpha-2}\, dx = \frac{(t+1)^{\alpha-1} - (t_0+2)^{\alpha-1}}{\alpha - 1} \leq \frac{(t+1)^{\alpha-1}}{\alpha - 1}.$$

Thus the tail contribution is at most

$$\sum_{s=t_0+1}^{t-1} \exp\Big(-\frac{r}{4}(T_t - T_s)\Big)\eta_s^2 \leq \frac{c^2}{\alpha - 1} \cdot \frac{1}{t+1}.$$

Initial part ($s \leq t_0$). For $s \leq t_0$, we only use $\eta_s \leq \eta_0$ and monotonicity of T_s:

$$\sum_{s=0}^{t_0} \exp\Big(-\frac{r}{4}(T_t - T_s)\Big)\eta_s^2 \leq (t_0+1)\eta_0^2 \exp\Big(-\frac{r}{4}(T_t - T_{t_0+1})\Big).$$

For $t \geq t_0 + 1$ the segment from $t_0 + 1$ to t is harmonic, so

$$T_t - T_{t_0+1} = \sum_{k=t_0+1}^{t-1} \frac{c}{k+1} \geq c \int_{t_0+2}^{t+1} \frac{dx}{x} = c \log\Big(\frac{t+1}{t_0+2}\Big),$$

and hence

$$\exp\Big(-\frac{r}{4}(T_t - T_{t_0+1})\Big) \leq \Big(\frac{t_0+2}{t+1}\Big)^{\alpha}.$$

Therefore the initial contribution is bounded by

$$\sum_{s=0}^{t_0} \exp\Big(-\frac{r}{4}(T_t - T_s)\Big)\eta_s^2 \leq (t_0+1)\eta_0^2 \Big(\frac{t_0+2}{t+1}\Big)^{\alpha}.$$

Combining initial and tail bounds yields the first displayed inequality. The simplified bound $\bar{\eta}_t \leq (C_{\mathrm{burn}} + C_{\mathrm{harm}})/(t+1)$ follows since $(t_0+1)\eta_0^2 \big(\frac{t_0+2}{t+1}\big)^{\alpha} \leq C_{\mathrm{burn}}/(t+1)$ for $t \geq t_0 + 1$. ■

Appendix E. Proofs and lemmata for risk bounds in M_θ geometry

The following is a risk bound arising from analysing the generalization in $\|\cdot\|_{M_\theta}$-norm, where $M_\theta := H^{1/2}(H^{1/2}PH^{1/2})^{-\theta}H^{1/2}$ for $\theta \in [0,1]$ interpolating between H and P^{-1}.

Lemma 25 (Risk bounds in M_θ geometry) *Suppose that Assumption 5, 9, and 6 hold, and that* $n \geq \frac{8\beta}{r}\sqrt{\lambda_{\max}(HPM_\theta P)}\sqrt{\lambda_{\max}(M_\theta^{-1}H)}$. *Assume further that* $\kappa(PH)^{(1-\theta)} \leq \rho_\ell^2$ *and define and let* $r := 2\,\lambda_{\min}(PH)\,C_{\ell,P}^{(\theta)}\frac{\beta\,\alpha}{\alpha+\beta}$. *If the stepsizes are chosen as*

$$\eta_t := \min\Big\{\frac{C_{\ell,P}^{(\theta)}}{\beta\,\lambda_{\max}(PH)\,\kappa(PH)^{1-\theta}},\ \frac{8}{r(t+1)}\Big\},$$

then, for all t sufficiently large, the population excess risk satisfies

$$\mathbb{E}_{S,\mathcal{A}}[\delta f(x_t)] \leq \frac{64}{r}\left(\frac{\beta}{r}\frac{\mathbb{E}_S[\mathrm{tr}(PHP\Sigma_S)]}{t+1} + \sqrt{\mathrm{tr}(M_\theta^{-1}\Sigma)\ \mathrm{tr}(PM_\theta P\Sigma)}\left(\frac{1}{\sqrt{n(t+1)}}+\frac{1}{n}\right)\right).$$

Proof Taking the expectation over S in the optimization inequality from Lemma 23 yields

$$\mathbb{E}_{S,\mathcal{A}}[\delta f_S(x_t)] \leq \mathbb{E}_S\big[e^{-\lambda_{\min}(PH)\alpha T_t}(f_S(x_0)-f_S^*)\big] + \frac{\beta}{2}\mathbb{E}_S[\mathrm{tr}(PHP\Sigma_S)]\,\bar\eta_t.$$

Plugging this into the expected excess risk result in Lemma 7 gives

$$\begin{aligned}\mathbb{E}_{S,\mathcal{A}}[\delta f(x_t)] \leq\ & 2\,\mathbb{E}_S\big[e^{-\lambda_{\min}(PH)\alpha T_t}(f_S(x_0)-f_S^*)\big] + \beta\,\mathbb{E}_S[\mathrm{tr}(PHP\Sigma_S)]\,\bar\eta_t \\ & + 2\sqrt{\mathrm{tr}(M_\theta^{-1}\Sigma)}\,\varepsilon_{\mathrm{pstab}} + 4\beta\lambda_{\max}(HM_\theta^{-1})\varepsilon_{\mathrm{pstab}}^2.\end{aligned}$$

By the stability result in Lemma 8 and $\sqrt{u+v}\leq\sqrt{u}+\sqrt{v}$ and $1-e^{-T_t r/4}\leq 1$, we obtain

$$\varepsilon_{\mathrm{pstab}} \leq 8\sqrt{\mathrm{tr}(PM_\theta P\Sigma)}\Big(\sqrt{\tfrac{\bar\eta_t}{n}}+\tfrac{1}{n\,r}\Big), \qquad \varepsilon_{\mathrm{pstab}}^2 \leq 64\ \mathrm{tr}(PM_\theta P\Sigma)\Big(\tfrac{\bar\eta_t}{n}+\tfrac{1}{n^2r^2}\Big).$$

Substituting and absorbing numerical constants yields

$$\begin{aligned}\mathbb{E}_{S,\mathcal{A}}[\delta f(x_t)] \leq\ & 2\,\mathbb{E}_S\big[e^{-\lambda_{\min}(PH)\alpha T_t}(f_S(x_0)-f_S^*)\big] \\ & + 64\Big(\beta\,\mathbb{E}_S[\mathrm{tr}(PHP\Sigma_S)]\,\bar\eta_t + \sqrt{\mathrm{tr}(M_\theta^{-1}\Sigma)\,\mathrm{tr}(PM_\theta P\Sigma)}\Big(\sqrt{\tfrac{\bar\eta_t}{n}}+\tfrac{1}{n\,r}\Big)\Big). \quad (24)\end{aligned}$$

Define

$$\eta_t := \min\Big\{\frac{C_{\ell,P}^{(\theta)}}{\beta\,\lambda_{\max}(PH)\,\kappa(PH)^{1-\theta}},\ \frac{8}{t+1}\frac{\alpha+\beta}{2\,\lambda_{\min}(PH)\,C_{\ell,P}^{(\theta)}\,\alpha\,\beta}\Big\},$$

and recall that

$$r = 2\lambda_{\min}(PH)C_{\ell,P}^{(\theta)}\frac{\beta\,\alpha}{\alpha+\beta}.$$

With this choice, the harmonic phase satisfies $\eta_t = \frac{8}{r(t+1)}$ for all t large enough, and the bound in Lemma 24 yields

$$\bar\eta_t \leq \frac{64}{r^2}\cdot\frac{1}{t+1}, \qquad \sqrt{\bar\eta_t} \leq \frac{8}{r}\cdot\frac{1}{\sqrt{t+1}},$$

where the burn-in contribution decays faster and is absorbed into constants.

Moreover, since $T_t \gtrsim \frac{8}{r}\log(t)$ in the harmonic regime, the exponential bias term decays at least as $O(1/t^2)$ and is negligible relative to the $1/t$ term.

Substituting the above bounds into (24) and simplifying gives, for all t sufficiently large,

$$\mathbb{E}_{S,\mathcal{A}}[\delta f(x_t)] \;\le\; \frac{64}{r}\left(\frac{\beta}{r}\,\frac{\mathbb{E}_S[\mathrm{tr}(PHP\Sigma_S)]}{t+1} + \sqrt{\mathrm{tr}(M_\theta^{-1}\Sigma)\,\mathrm{tr}(PM_\theta P\Sigma)}\Big(\frac{1}{\sqrt{n(t+1)}} + \frac{1}{n}\Big)\right),$$

which is exactly the stated bound. ■

Appendix F. Proofs for non-convex PL-losses in Section 4.2

F.1. Proof for Proposition 14

Proof We begin with the decomposition,

$$\delta f(x_t(S)) \le \mathbb{E}[f(x_t(S)) - f(\hat{x}^*)] + \delta f(\hat{x}^*), \tag{25}$$

where we use the shorthand $\hat{x}^* = \mathrm{Proj}_S(x_t(S))$ for brevity. We first analyse the parameter stability of this empirical risk minimiser. Setting $\hat{y}^* = \mathrm{Proj}_{S^{(i)}}(x^*)$ and using the quadratic growth property implied by the μ-PL inequality, we obtain,

$$\begin{aligned}
\|\hat{x}^* - \hat{y}^*\|_H &\le \frac{1}{\mu}\|\nabla f_S(\hat{y}^*)\|_{H^{-1}} \\
&= \frac{1}{\mu}\left\|\nabla f_{S^{(i)}}(\hat{y}^*) + \frac{1}{n}\nabla\ell(\hat{y}^*, z_i) - \frac{1}{n}\nabla\ell(\hat{y}^*, z')\right\|_{H^{-1}} \\
&= \frac{1}{\mu n}\|\nabla\ell(\hat{y}^*, z_i) - \nabla\ell(\hat{y}^*, z')\|_{H^{-1}},
\end{aligned}$$

where we use that $\nabla f_{S^{(i)}}(\hat{y}^*) = 0$. Taking the expectation and adding/subtracting population gradients:

$$\begin{aligned}
&\mathbb{E}[\|\nabla\ell(\hat{y}^*, z_i) - \nabla\ell(\hat{y}^*, z')\|_{H^{-1}}^2]^{1/2} \\
&\quad\le \mathbb{E}[\|\nabla\ell(\hat{y}^*, z_i) - \nabla f(\hat{y}^*)\|_{H^{-1}}^2]^{1/2} + \mathbb{E}[\|\nabla f(\hat{y}^*) - \nabla f(\hat{x}^*)\|_{H^{-1}}^2]^{1/2} \\
&\qquad + \mathbb{E}[\|\nabla f(\hat{x}^*) - \nabla\ell(\hat{x}^*, z')\|_{H^{-1}}^2]^{1/2} + \mathbb{E}[\|\nabla\ell(\hat{y}^*, z') - \nabla\ell(\hat{x}^*, z')\|_{H^{-1}}^2]^{1/2} \\
&\quad\le 2\,\mathrm{tr}(H^{-1}\Sigma)^{1/2} + 2\beta\mathbb{E}[\|\hat{x}^* - \hat{y}^*\|_H^2]^{1/2}.
\end{aligned}$$

Substituting this back into the bound for $\|\hat{x}^* - \hat{y}^*\|_H$ and rearranging yields:

$$\mathbb{E}[\|\hat{x}^* - \hat{y}^*\|_H^2]^{1/2} \le \left(1 - \frac{2\beta}{\mu n}\right)^{-1}\frac{2\,\mathrm{tr}(H^{-1}\Sigma)^{1/2}}{\mu n}.$$

Assuming $n \ge 4\beta/\mu$, the pre-factor is bounded by 2. Squaring gives the bound,

$$\mathbb{E}[\|\hat{x}^* - \hat{y}^*\|_H^2] \le \frac{16\,\mathrm{tr}(H^{-1}\Sigma)}{\mu^2 n^2}. \tag{26}$$

By Lemma 7 with $M = H$ (so $\lambda_{\max}(HM^{-1}) = 1$) and the stability bound on the ERM minimizer, we have that,

$$\mathbb{E}_{S,\mathcal{A}}[\delta f(\hat{x}^*)] \leq \frac{8\,\mathrm{tr}(H^{-1}\Sigma)}{\mu n} + \beta\frac{64\,\mathrm{tr}(H^{-1}\Sigma)}{\mu^2 n^2}.$$

Now, to bound the second term of (25), we use smoothness to obtain,

$$\begin{aligned}\mathbb{E}[f(x_t(S)) - f(\hat{x}^*)] &\leq \mathbb{E}[\langle \nabla f(\hat{x}^*), x_t(S) - \hat{x}^*\rangle] + \frac{\beta}{2}\mathbb{E}[\|x_t(S) - \hat{x}^*\|_H^2] \\ &\leq \frac{1}{2\beta}\mathbb{E}[\|\nabla f(\hat{x}^*)\|_{H^{-1}}^2] + \beta\mathbb{E}[\|x_t(S) - \hat{x}^*\|_H^2] \\ &\leq \mathbb{E}[\delta f(\hat{x}^*)] + \frac{2\beta}{\mu}\mathbb{E}[\delta f_S(x_t(S))].\end{aligned}$$

■

Appendix G. Proofs for lower bounds in Section 5

The following formulation of Assouad's lemma is from (Ma et al., 2024, Lemma 23).

Lemma 26 (Assouad's lemma) *Let $d \in \mathbb{N}$, $\Phi := \{0,1\}^d$. For $\phi \in \Phi$, let $x_\phi \in \mathcal{X}$ and $P_\phi \in \mathcal{P}_{x_\phi}$. For $\phi, \phi' \in \Phi$, we write $\phi \sim \phi'$ whenever ϕ and ϕ' differ in precisely one coordinate, and $\phi \sim_j \phi'$ when that coordinate is j^{th}. Supposed now that the loss function is of the form*

$$\ell(x_1, x_2) := \sum_{j\in[d]} g(\rho_j(x_1, x_2)),$$

for $x_1, x_2 \in \mathcal{X}$, where $\rho_1, \ldots \rho_d$ are pseudo metrics on $\mathcal{X}$ with $\rho_j(x_\phi, x_{\phi'}) \geq \delta_j$ whenever $\phi \sim_j \phi'$, and where g is an increasing function satisfying $g(t_1 + t_2) \leq A(g(t_1) + g(t_2))$ for all $t_1, t_2 \geq 0$ and some $A > 0$. Then, for $\mathcal{X}_0 := \{x_\phi : \phi \in \Phi\}$, we have

$$\begin{aligned}\inf_{\hat{x}} \sup_{x\in\mathcal{X}} \sup_{P_\theta\in\mathcal{P}_\theta} \mathbb{E}[\ell(\hat{x}, x)] &\geq \inf_{\hat{x}} \max_{x_0\in\mathcal{X}_0} \sup_{P_{x_0}\in\mathcal{P}_{x_0}} \mathbb{E}[\ell(\hat{x}, x_0)] \\ &\geq \frac{1}{2A}\left(1 - \max_{\phi,\phi'\in\Phi:\phi\sim\phi'} \mathrm{TV}\left(P_\phi, P_{\phi'}\right)\right) \sum_{j\in[d]} g(\delta_j),\end{aligned}$$

where $\hat{x}$ is computed from a sample of P_{x_0}.

G.1. Proof of Theorem 15

Proof Let $\alpha > 0$ and $v \in \{0,1\}^d$. Let $\ell(x, z) := \frac{\alpha}{2}\|x - z\|_H^2$, where $z \sim P_v$ for $P_v = \mathcal{N}(\mu_v, H^{-1}\Sigma H^{-1}/\alpha^2)$, and μ_v will be specified later. By definition, $\ell(\cdot, z)$ is α-strongly convex in $\|\cdot\|_H$-norm and $\mathrm{Var}_{z\in P_v}(\nabla\ell(x, z)) = \mathrm{Var}(\alpha H z) = \Sigma$.

We have the following equivalence

$$\mathbb{E}_{P_v}\left[\frac{\alpha}{2}\|x - z\|_H^2\right] = \mathbb{E}_{P_v}\left[\frac{\alpha}{2}\|H^{-1/2}\bar{x} - H^{-1/2}\bar{z}\|_H^2\right] = \mathbb{E}_{\bar{P}_\phi}\left[\frac{\alpha}{2}\|\bar{x} - \bar{z}\|_2^2\right],$$

after we substituted $\bar{x} \coloneqq H^{1/2}x$, $\bar{z} \coloneqq H^{1/2}z$ and $\bar{z} \sim \bar{P}_v$ where $\bar{P}_v \coloneqq \mathcal{N}(\hat{\mu}_v, H^{-1/2}\Sigma H^{-1/2}/\alpha^2)$ and $\hat{\mu}_v \coloneqq H^{1/2}\mu_v$.

Let $\bar{\Sigma} \coloneqq H^{-1/2}\Sigma H^{-1/2}$ and $\bar{\Sigma} = Q\Lambda Q^\top$ be its spectral decomposition. Consider the set of $\bar{\mu}_v = \sum_{j\in[d]} \delta_j \theta_j q_j$, where $\delta_j = \frac{4}{3\alpha}\sqrt{\lambda_j/n}$, or in a matrix form $\bar{\mu}_v = QDv$ where $D = \frac{4}{3\alpha\sqrt{n}}\Lambda^{1/2}$. Define $\mathcal{M}_0 = \{\bar{\mu}_v : v \in \{0,1\}^d\}$. For $v \sim_j v'$ we have $|q_j^\top(QDv - QDv')| \geq |\delta_j|$. By Pinsker inequality we have

$$\begin{aligned}\mathrm{TV}(\bar{P}_v, \bar{P}_{v'}) &\leq \left(\frac{n}{2}\mathrm{KL}(\mathcal{N}(\bar{\mu}_v, \bar{\Sigma}/\alpha^2), \mathcal{N}(\bar{\mu}_{v'}, \bar{\Sigma}/\alpha^2))\right)^{1/2} \\ &= \left(\frac{n}{4}\frac{16}{9n}\|v - v'\|_2^2\right)^{1/2} = 2/3.\end{aligned}$$

By Lemma 26 with $g(t) = \frac{\alpha}{2}t^2$, for which $A = 2$, we have that

$$\begin{aligned}\inf_{\hat{x}\in\mathcal{X}} \sup_{P\in\mathcal{P}} \mathbb{E}_{S\sim P^n}[f(\hat{x}(S)) - f(\tilde{x})] &\geq \frac{1}{2A}\left(1 - \max_{v,v',v\sim v'} \mathrm{TV}(\bar{P}_v, \bar{P}_{v'})\right) \sum_{j\in[d]} \frac{\alpha}{2}\delta_j^2 \\ &\geq \frac{2}{27n\alpha}\,\mathrm{tr}(H^{-1}\Sigma).\end{aligned}$$

■

Lemma 27 (Decaying step-size bounds) *Let $0 < a < b$ and $\eta_t = \min\left(\frac{1}{2b}, \frac{1}{at}\right)$. Let $t_0 = \lceil \frac{2b}{a} \rceil$.*

1. ***Upper Bound:*** *The recurrence $r_{t+1} \leq (1 - 2a\eta_t)r_t + \eta_t^2 B$ satisfies*

$$r_t \leq \left(\frac{2b}{e^2 a}r_0 + \frac{B}{a^2}\right)\frac{1}{t} \quad \textit{for all } t \geq t_0.$$

2. ***Lower Bound:*** *The recurrence $r_{t+1} \geq (1 - 2b\eta_t)r_t + \eta_t^2 B$ satisfies*

$$r_t \geq \frac{B}{2abt} \quad \textit{for all } t \geq t_0.$$

Proof Upper Bound. *Phase 1 (Constant Step):* For $t < t_0$, the recurrence $r_{t+1} \leq (1 - \frac{a}{b})r_t + \frac{B}{4b^2}$ implies linear convergence to a noise floor. Unrolling from $t = 0$ to t_0:

$$r_{t_0} \leq \left(1 - \frac{a}{b}\right)^{t_0} r_0 + \frac{B}{4b^2}\sum_{i=0}^{t_0-1}\left(1 - \frac{a}{b}\right)^i \leq e^{-2}r_0 + \frac{B}{4ab}.$$

Phase 2 (Decaying Step): For $t \geq t_0$, we prove $r_t \leq \nu/t$ by induction. Assume $r_t \leq \nu/t$. Substituting $\eta_t = 1/at$:

$$r_{t+1} \leq \left(1 - \frac{2}{t}\right)\frac{\nu}{t} + \frac{B}{a^2t^2} = \frac{\nu}{t} - \frac{1}{t^2}\left(2\nu - \frac{B}{a^2}\right).$$

We require $r_{t+1} \leq \frac{\nu}{t+1}$. Using the inequality $\frac{1}{t+1} \geq \frac{1}{t} - \frac{1}{t^2}$, it suffices that the drop in the recurrence is at least ν/t^2.

$$\frac{1}{t^2}\left(2\nu - \frac{B}{a^2}\right) \geq \frac{\nu}{t^2} \implies \nu \geq \frac{B}{a^2}.$$

The definition of ν satisfies this condition and ensures the bound holds at the transition t_0 (since $\nu/t_0 \geq r_{t_0}$).

Lower Bound. We prove $r_t \geq \kappa/t$ with $\kappa = \frac{B}{2ab}$ for $t \geq t_0$. *Base Case ($t = t_0$):* Unrolling the recurrence with $\eta_t = 1/2b$ implies r_{t_0} accumulates noise terms summing to at least $\frac{B}{4b^2}$. Checking the bound: $\frac{\kappa}{t_0} = \frac{B}{2abt_0}$. Since $t_0 \geq 2b/a$, we have $\frac{B}{2abt_0} \leq \frac{B}{2ab(2b/a)} = \frac{B}{4b^2}$, so the base case holds.

Inductive Step ($t > t_0$): Assume $r_t \geq \kappa/t$. Using $\eta_t = 1/at$, the recurrence drop is:

$$r_{t+1} \geq \frac{\kappa}{t} - \frac{1}{t^2}\left(\frac{2b\kappa}{a} - \frac{B}{a^2}\right).$$

We need this to be $\geq \frac{\kappa}{t+1} \geq \frac{\kappa}{t} - \frac{\kappa}{t^2}$. This requires the coefficient of the drop to satisfy:

$$\frac{2b\kappa}{a} - \frac{B}{a^2} \leq \kappa \implies \kappa\left(\frac{2b}{a} - 1\right) \leq \frac{B}{a^2}.$$

Substituting $\kappa = \frac{B}{2ab}$:

$$\frac{B}{2ab}\left(\frac{2b-a}{a}\right) = \frac{B(2b-a)}{2a^2b} = \frac{B}{a^2}\left(1 - \frac{a}{2b}\right) < \frac{B}{a^2}.$$

The inequality holds strictly, validating the lower bound. ■

G.2. Proof of Lemma 16

Proof Let $\ell(x,z) = \frac{\alpha}{2}\|x-z\|_H^2$ and $z \sim \mathrm{Q} := \mathcal{N}(\mu, \frac{1}{\alpha^2}H^{-1}\Sigma H^{-1})$. Then

$$\nabla\ell(x,z) = \alpha H(x-z), \qquad \mathrm{Var}_{z\sim\mathrm{Q}}(\nabla\ell(x,z)) = \Sigma,$$

and the population risk is $f(x) = \mathbb{E}_{z\sim\mathrm{Q}}[\ell(x,z)] = \frac{\alpha}{2}\|x-\mu\|_H^2 + \frac{1}{2\alpha}\,\mathrm{tr}(H^{-1}\Sigma)$.

Fix the dataset $S = \{z_i\}_{i=1}^n$ and define

$$m_t := \mathbb{E}_{\mathcal{A}}[x_t \mid S], \quad \text{and} \quad u_t := x_t - m_t,$$

where $\mathcal{A}$ reflects the randomness of uniform sampling $i_t \sim \mathrm{Unif}([n])$. Since i_t is sampled independently of $i_1, \ldots, i_{t-1}$, conditional on S the current sample z_{i_t} is independent of u_t. By linearity of the quadratic update, we can then express explicitly

$$m_{t+1} = m_t - \eta_t P \nabla f_S(m_t), \qquad u_{t+1} = (I - \alpha\eta_t PH)u_t + \alpha\eta_t PH(z_{i_t} - \bar{z}_S),$$

where $f_S(x) = \frac{1}{n}\sum_{i=1}^n \ell(x,z_i)$ and $\bar{z}_S := \frac{1}{n}\sum_{i=1}^n z_i$ is the empirical mean.

Since $x_t - \mu = (m_t - \mu) + u_t$ and $\mathbb{E}_{\mathcal{A}}[u_t \mid S] = 0$, the conditional population excess risk decomposes as

$$\mathbb{E}_{\mathcal{A}}[\delta f(x_t) \mid S] = \underbrace{\frac{\alpha}{2}\,\mathbb{E}_{\mathcal{A}}[\|u_t\|_H^2 \mid S]}_{=:\widetilde{\delta}_t(S)} + \underbrace{\frac{\alpha}{2}\|m_t - \mu\|_H^2}_{=:\zeta_t(S)},$$

where the first term denoted as $\widetilde{\delta}_t(S)$ can be improved by optimization and the second term $\zeta_t(S)$ will result in the asymptotic statistical floor of the ERM.

First term $\widetilde{\delta}_t(S)$. We first analyze $\widetilde{\delta}_t(S)$. Conditional on S, the noise term $\bar{z}_S - z_{i_t}$ has mean zero and is independent of u_t. Therefore

$$\begin{aligned}\widetilde{\delta}_{t+1}(S) &= \frac{\alpha}{2}\mathbb{E}_{\mathcal{A}}\big[\|(I-\alpha\eta_t PH)u_t + \alpha\eta_t PH(\bar{z}_S - z_{i_t})\|_H^2 \,\big|\, S\big] \\ &= \frac{\alpha}{2}\mathbb{E}_{\mathcal{A}}[\|(I-\alpha\eta_t PH)u_t\|_H^2 \mid S] + \frac{\alpha^3\eta_t^2}{2}\mathbb{E}_{\mathcal{A}}[\|PH(\bar{z}_S - z_{i_t})\|_H^2 \mid S].\end{aligned}$$

The first term is bounded by the spectrum and the stepsize $\eta_t \le 1/\lambda_{\max}(PH)$ as

$$\frac{\alpha}{2}\mathbb{E}_{\mathcal{A}}[\|(I-\alpha\eta_t PH)u_t\|_H^2 \mid S] \ge (1 - 2\alpha\eta_t\lambda_{\max}(PH))\,\widetilde{\delta}_t(S).$$

To bound the second term, denote the sample covariance as $\Sigma_S := \mathrm{Var}_{i\sim\mathrm{Unif}([n])}(\nabla\ell(x,z_i) \mid S)$. Then, since in the quadratic model $\nabla\ell(x,z) = \alpha H(x-z)$, the sample covariance Σ_S is independent of x, resulting in

$$\frac{\alpha^3\eta_t^2}{2}\mathbb{E}_{\mathcal{A}}[\|PH(\bar{z}_S - z_{i_t})\|_H^2 \mid S] = \frac{\alpha\eta_t^2}{2}\,\mathrm{tr}(PHP\Sigma_S).$$

Hence

$$\widetilde{\delta}_{t+1}(S) \ge (1 - 2\alpha\eta_t\lambda_{\max}(PH))\,\widetilde{\delta}_t(S) + \frac{\alpha\eta_t^2}{2}\,\mathrm{tr}(PHP\Sigma_S)$$

and applying the lower-bound part of Lemma 27 as in the single-pass case yields

$$\widetilde{\delta}_{t+1}(S) \ge \frac{\mathrm{tr}(PHP\Sigma_S)}{2\alpha\,\lambda_{\max}(PH)\lambda_{\min}(PH)} \cdot \frac{1}{t+1}, \qquad t \ge t_0 := \lfloor 4\kappa(PH)\rfloor.$$

It remains to compute $\mathbb{E}_S[\Sigma_S]$. Since $\nabla\ell(x,z) = \alpha H(x-z)$, the covariance over $i \sim \mathrm{Unif}([n])$ is

$$\Sigma_S = \alpha^2 H\left(\frac{1}{n}\sum_{i=1}^n (z_i - \bar{z}_S)(z_i - \bar{z}_S)^\top\right) H.$$

Using the identity

$$\mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^n (z_i - \bar{z}_S)(z_i - \bar{z}_S)^\top\right] = \left(1 - \frac{1}{n}\right)\mathrm{Var}_{z\sim\mathrm{Q}}(z),$$

we obtain $\mathbb{E}_S[\Sigma_S] = \big(1 - \frac{1}{n}\big)\,\Sigma$.

Second term $\zeta_t(S)$. We now examine the behaviour of the second term $\zeta_t(S)$. Define $A_t := \prod_{s=0}^{t-1}(I - \alpha\eta_s PH)$, which governs the dynamics of the conditional expectation as:

$$m_t = A_t x_0 + (I - A_t)\bar{z}_S, \qquad m_t - \mu = A_t(x_0 - \mu) + (I - A_t)(\bar{z}_S - \mu).$$

Taking expectation over S and using $\mathbb{E}_S[\bar{z}_S - \mu] = 0$, we get

$$\mathbb{E}_S[\|m_t - \mu\|_H^2] = \|A_t(x_0 - \mu)\|_H^2 + \mathbb{E}_S[\|(I - A_t)(\bar{z}_S - \mu)\|_H^2] \ge \mathbb{E}_S[\|(I - A_t)(\bar{z}_S - \mu)\|_H^2].$$

Since $\mathrm{Cov}_S(\bar{z}_S) = \frac{1}{\alpha^2 n}H^{-1}\Sigma H^{-1}$, we get

$$\frac{\alpha}{2}\mathbb{E}_S[\|m_t - \mu\|_H^2] \ge \frac{1}{2\alpha n}\,\mathrm{tr}\Big((I - A_t)^\top H(I - A_t)H^{-1}\Sigma H^{-1}\Big).$$

It remains to analyze the dynamics of A_t on the right-hand-side. Let

$$T := H^{1/2} P H^{1/2}, \qquad p_t(\lambda) := \prod_{s=0}^{t-1} (1 - \alpha\eta_s \lambda).$$

By $A_t = H^{-1/2} p_t(T) H^{1/2}$ and the cyclicity of the trace, we get

$$\operatorname{tr}\Big((I - A_t)^\top H (I - A_t) H^{-1} \Sigma H^{-1}\Big) = \operatorname{tr}\Big((I - p_t(T))^2 H^{-1/2} \Sigma H^{-1/2}\Big).$$

Because $\eta_s \le 1/\lambda_{\max}(PH)$, we have $0 \preceq p_t(T) \preceq I$, and therefore

$$\operatorname{tr}\Big((I - p_t(T))^2 H^{-1/2} \Sigma H^{-1/2}\Big) \ge \lambda_{\min}\big((I - p_t(T))^2\big) \operatorname{tr}(H^{-1}\Sigma).$$

Since $p_t(\lambda)$ is decreasing on $[0, \lambda_{\max}(PH)]$, we have $\lambda_{\min}\big((I - p_t(T))^2\big) = \Big(1 - p_t(\lambda_{\min}(PH))\Big)^2$. Thus

$$\frac{\alpha}{2} \mathbb{E}_S[\|m_{t+1} - \mu\|_H^2] \ge \frac{\gamma_t(P, H)}{2\alpha n} \operatorname{tr}(H^{-1}\Sigma),$$

where

$$\gamma_t(P, H) := \left(1 - \prod_{s=0}^{t} \big(1 - \alpha\eta_s \lambda_{\min}(PH)\big)\right)^2.$$

Combining the two lower bounds and averaging over S, we get

$$\mathbb{E}_{\mathcal{A},S}[\delta f(x_{t+1})] \ge \left(1 - \frac{1}{n}\right) \frac{\operatorname{tr}(PHP\Sigma)}{2\alpha\, \lambda_{\max}(PH)\lambda_{\min}(PH)} \cdot \frac{1}{t+1} + \frac{\gamma_t(P, H)}{2\alpha n} \operatorname{tr}(H^{-1}\Sigma).$$

■

G.3. Proof of Corollary 17

Proof From Lemma 16, dropping the nonnegative statistical floor term, we have

$$\mathbb{E}_{\mathcal{A},S}[\delta f(x_{t+1})] \ge \left(1 - \frac{1}{n}\right) \frac{\operatorname{tr}(PHP\Sigma)}{2\alpha\, \lambda_{\max}(PH)\lambda_{\min}(PH)} \cdot \frac{1}{t+1}$$

for $t \ge t_0 := \lfloor 4\kappa(PH) \rfloor$.

Let $H = Q \operatorname{diag}(h) Q^\top$ be the spectral decomposition where $h = (h_1, \dots, h_d)$ and assume without loss of generality that $h_1 = \lambda_{\max}(H) = 1$. Define $\gamma_i := h_i q_i^\top \Sigma q_i$. Then $\operatorname{tr}(H\Sigma) = \sum_{i=1}^d h_i q_i^\top \Sigma q_i = \sum_{i=1}^d \gamma_i$. Thus by averaging there exists an index $k \ne 1$ such that $\gamma_k \le \frac{1}{d-1} \operatorname{tr}(H\Sigma)$.

Construct $P_\varepsilon := I - (1 - \frac{\varepsilon}{h_k}) q_k q_k^\top$ with $\varepsilon \le \min_{j \ne k} h_j$. Its eigenvalues are 1 on $\{q_j\}_{j \ne k}$ and ε/h_k on q_k, so $\lambda_{\max}(P_\varepsilon) = 1$. Since $P_\varepsilon H = \sum_{j \ne k} h_j q_j q_j^\top + \varepsilon\, q_k q_k^\top$, the eigenvalues of $P_\varepsilon H$ are $\{h_j : j \ne k\} \cup \{\varepsilon\}$, and because $k \ne 1$ the largest eigenvalue $h_1 = 1$ is retained. Hence $\lambda_{\min}(P_\varepsilon H) = \varepsilon$ and $\lambda_{\max}(P_\varepsilon H) = 1$, giving $\kappa(P_\varepsilon H) = 1/\varepsilon$.

$$\begin{aligned}\operatorname{tr}(P_\varepsilon H P_\varepsilon \Sigma) &= \operatorname{tr}(H\Sigma) - h_k q_k^\top \Sigma q_k + \frac{\varepsilon^2}{h_k} q_k^\top \Sigma q_k \\ &\geq \operatorname{tr}(H\Sigma) - \gamma_k = \operatorname{tr}(H\Sigma)\left(1 - \frac{1}{d-1}\right),\end{aligned}$$

where we dropped the last term. Putting it together, and using $\lambda_{\max}(P_\varepsilon H)\lambda_{\min}(P_\varepsilon H) = \varepsilon$, we have that

$$\frac{\operatorname{tr}(P_\varepsilon H P_\varepsilon \Sigma)}{\lambda_{\max}(P_\varepsilon H)\lambda_{\min}(P_\varepsilon H)} \geq \frac{\operatorname{tr}(H\Sigma)}{\varepsilon}\left(1 - \frac{1}{d-1}\right).$$

Substituting this into the lemma bound above yields

$$\mathbb{E}_{\mathcal{A},S}[\delta f(x_{t+1})] \geq \left(1 - \frac{1}{n}\right)\left(1 - \frac{1}{d-1}\right)\frac{\operatorname{tr}(H\Sigma)}{2\alpha\,\varepsilon\,(t+1)}.$$

■

G.4. Proof of Corollary 18

Proof Set $A := PHP$ and $B := H^{-1}$. For any $\Sigma \succ 0$ write $\Sigma = B^{-1/2} X B^{-1/2}$ with $X \succ 0$. Then

$$\begin{aligned}&\frac{\operatorname{tr}(PHP\,\Sigma)}{\operatorname{tr}(H^{-1}\Sigma)} = \frac{\operatorname{tr}\big(B^{-1/2} A B^{-1/2}\, X\big)}{\operatorname{tr}(X)} = \frac{\operatorname{tr}(MX)}{\operatorname{tr}(X)}, \\ &\text{with} \quad M := B^{-1/2} A B^{-1/2} = H^{1/2}(PHP)H^{1/2} = \big(H^{1/2} P H^{1/2}\big)^2 \succ 0.\end{aligned}$$

By the variational characterization over $\{X \succeq 0 : \operatorname{tr}(X) = 1\}$,

$$\frac{\operatorname{tr}(PHP\,\Sigma)}{\operatorname{tr}(H^{-1}\Sigma)} \;\leq\; \lambda_{\max}(M) = \lambda_{\max}\big(H^{1/2} P H^{1/2}\big)^2 = \lambda_{\max}(PH)^2,$$

where we used that $H^{1/2} P H^{1/2}$ is similar to PH and thus has the same (positive) spectrum. Dividing by $\lambda_{\max}(PH)\lambda_{\min}(PH)$ gives

$$\frac{\operatorname{tr}(PHP\,\Sigma)}{\lambda_{\max}(PH)\lambda_{\min}(PH)\,\operatorname{tr}(H^{-1}\Sigma)} \;\leq\; \frac{\lambda_{\max}(PH)}{\lambda_{\min}(PH)}.$$

The equality is attained by taking $X = q_1 q_1^\top$, where q_1 is a top eigenvector of M; equivalently, q_1 is a top eigenvector of $H^{1/2} P H^{1/2}$ (i.e., of PH), and the choice in the statement

$$\Sigma \;=\; H^{1/2} q_1 q_1^\top H^{1/2}$$

gives $\frac{\operatorname{tr}(PHP\,\Sigma)}{\operatorname{tr}(H^{-1}\Sigma)} = \lambda_{\max}(PH)^2$, yielding equality in the bound above. Substituting into the first term of Lemma 16 (and dropping the nonnegative statistical floor term) gives

$$\mathbb{E}_{\mathcal{A},S}[\delta f(x_{t+1})] \;\geq\; \left(1 - \frac{1}{n}\right)\frac{\operatorname{tr}(PHP\Sigma)}{2\alpha\,\lambda_{\max}(PH)\lambda_{\min}(PH)\,(t+1)} \;=\; \left(1 - \frac{1}{n}\right)\kappa(PH)\frac{\operatorname{tr}(H^{-1}\Sigma)}{2\alpha\,(t+1)},$$

which is the stated bound.

For a fixed P, if one is allowed to vary $H \succ 0$ under only the constraint $\rho(H) = 1$, the quantity can be made arbitrarily large because $\frac{\lambda_{\max}(PH)}{\lambda_{\min}(PH)}$ is unbounded in H, e.g., take $H = \operatorname{diag}(1, \varepsilon, \ldots, \varepsilon)$ in an eigenbasis of P and let $\varepsilon \to 0$. ■