ANALYSIS OF THE STABILITY OF PRECONDITIONED STOCHASTIC GRADIENT DESCENT

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the algorithmic stability of preconditioned gradient descent (PGD) for empirical risk minimization. Although preconditioning affects finite-time stability, existing analyses fail to explain PGD's long-term behavior. We propose a refined stability analysis that separates the effect of preconditioning on transient optimization dynamics from PGD's intrinsic stability, recovering the long-term behavior obscured by standard contraction-based analyses. For deterministic updates, our decomposition establishes non-asymptotic bounds, while showing that long-term stability is independent of the preconditioner. For stochastic updates, we establish an analogous decomposition with an additional term that retains a dependence on the preconditioner. In both cases, each contribution is of order $O(1/n)$, up to variance terms and transient terms that decay with the number of iterations. Our conclusions hold in two regimes. First, the regularized setting, where individual losses are Lipschitz and the empirical objective is regularized by a sample-independent strongly convex function. Second, projected PGD for constrained problems, showing that projection introduces only additional residual terms. Finally, we use these stability estimates to derive expected generalization bounds and excess-risk guarantees for regularized and constrained empirical risk minimization.

1 INTRODUCTION

Algorithmic stability studies the impact of replacing one training example on an algorithm's output. This idea has been very successful in analyzing the generalization properties of (stochastic) gradient-based learning algorithms (Hardt et al., 2016), relying on the finding that stable algorithms have a small expected generalization gap (Bousquet & Elisseeff, 2002). Beyond generalization, algorithmic stability has led to novel results in online learning (Saha et al., 2012), adaptive data analysis (Chaudhuri et al., 2011; Bassily et al., 2016), and unlearning (Ullah et al., 2021).

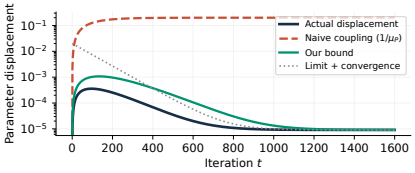


Figure 1: Prior results on PGD's stability either (i) fail to capture small initial stability (dotted, gray), or (ii) give a very conservative asymptotic bound (dashed, red). In contrast, our analysis sharply characterizes both cases (green).

Stability guarantees for gradient descent methods are typically obtained either using a direct contraction-based argument (Hardt et al., 2016) or exploiting rates of convergence (Lei & Ying, 2020). The first type of analysis relies on accumulating gradient differences through the path of the algorithm, while the latter exploits results on the algorithm's convergence rate. Another line of work studies the stability of algorithms based on their convergence to a global optimum (Charles & Papailiopoulos, 2018), although the obtained results are only informative asymptotically.

Preconditioning is a key tool for accelerating gradient-based optimization, especially on ill-conditioned problems (Boyd & Vandenberghe, 2004; Nocedal & Wright, 2006), with applications ranging from classical numerical methods (Bramble et al., 1996; Knyazev & Neymeyr, 2003; 2009) to modern machine learning, where it has achieved impressive results (Zhang et al., 2021; Li, 2017; Li et al., 2016; Gupta et al., 2018). Despite this success, little attention has been drawn to the study of these methods' algorithmic stability, and therefore to their generalization properties. To our knowledge, the only work on this matter is the one of Vary et al. (2026), whose results give a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

bound on preconditioned gradient descent's stability. However, as illustrated in Figure 1, this bound misses a key property of deterministic preconditioned methods: after a transient phase affected by the preconditioner, algorithms can recover convergence to a minimizer independent of it.

In this paper, we study the stability of preconditioned gradient descent methods using a novel technique based on an exact decomposition of the error in two parts. The goal of this decomposition is to finely separate the neighboring-minimizer displacement from transient optimization drift and stochastic fluctuations. The first part captures most of the algorithm's instability, both at small iteration count and asymptotically, while the remaining error terms constitute a second part of the error, which is zero at time zero, and eventually vanishes. This allows us to capture the complex behavior of preconditioned optimization algorithms, where the trajectory of the iterates changes but not its asymptotic limit. Indeed, a direct coupling bound can exhibit a dependence on the conditioning matrix P even after deterministic optimization has converged, although the limiting points do not depend on the algorithmic preconditioner. Convergence-based stability arguments already suggest comparing the algorithm to its limiting empirical minimizer (Charles & Papailiopoulos, 2018).

We conduct our study in two regimes. First, the non-projected case, with a data-dependent part with uniformly bounded gradients together with a data-independent strongly convex regularizer. Second, constrained strongly convex and Lipschitz-continuous objectives with projection on a feasible set and fixed preconditioner. In this constrained setting, full gradients are bounded on the feasible set. Our main contributions are the following:

- We give a sharp characterization of PGD's stability, based on a decomposition of the stability displacement in three terms: a limiting empirical-minimizer displacement, a replacement drift, and a nonlinear Hessian drift. In the projected case, an additional projection drift term appears. We bound each of these terms, which results in an upper bound that follows closely the empirical behavior of PGD (see Figure 1). These bounds recover, up to higher-order terms, the early growth in $O(\gamma t/n)$, where γ is the step size, t the iteration count, and n the number of data records, as well as the expected limit stability, without relying on bounds on the minimizer's stability itself. This shows that the preconditioner controls the transient but not the asymptotic sensitivity of the deterministic algorithm.
- For the stochastic counterpart of PGD, we propose a similar decomposition, where the replacement drift, Hessian drift, and projection drift terms are random variables with non-vanishing variance. Interestingly, constant-step stochastic PGD retains a dependence on the preconditioner, which only affects higher-order terms. Nonetheless, taking time to infinity and then the step size to zero in our bounds recovers the same preconditioner-independent ERM stability scale, confirming the soundness of our analysis.
- Finally, we leverage our refined stability bounds to derive expected generalization and excess-risk bounds for deterministic and stochastic PGD. We give bounds on the generalization of PGD's first iterates and recover classical generalization guarantees of the global optimum asymptotically.

We present the problem and algorithms we study in Section 2. Section 3 develops the regularized case first; Section 4 treats projected methods and their statistical guarantees. The proofs related to these two sections are presented in Appendix C.4 and Appendix D.5. We then apply these bounds to give generalization guarantees for PGD and PSGD in Section 5. We position these results within the broader literature in Section 6, and illustrate our theoretical results numerically in Section 7.

2 Preconditioned (Stochastic) Gradient Descent

We consider the empirical risk minimization problem with a dataset $S = (z_1, \dots, z_n) \in \mathcal{Z}^n$ and a neighboring dataset $\widetilde{S} = (\widetilde{z}_1, \dots, \widetilde{z}_n)$, where $\widetilde{z}_i = z_i$ for all $i \neq j$. Throughout, j is a fixed index in $[n] = \{1, \dots, n\}$. Let $f\colon \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ be the loss function. For every $i \in [n]$, define $f_i(x) = f(x, z_i)$ and $\widetilde{f}_i(x) = f(x, \widetilde{z}_i)$ for $x \in \mathcal{X}$. Thus $\widetilde{f}_i = f_i$ for $i \neq j$. The empirical risks associated with S and $\widetilde{S}$ are:

$$F(x) = \tfrac{1}{n}\textstyle\sum_{i=1}^n f_i(x), \qquad \widetilde{F}(x) = \tfrac{1}{n}\textstyle\sum_{i=1}^n \widetilde{f}_i(x), \tag{2.1}$$

We study preconditioned gradient descent (PGD) and its stochastic counterpart (PSGD). The preconditioner $P \succ 0$ is fixed independently of the data. Both runs use the same initialization x_0.

Algorithm 1 (Projected) Preconditioned (stochastic) gradient descent

Require: Dataset $S = (z_1, \dots, z_n)$, preconditioner $P \succ 0$, feasible set K, initial point $x_0 \in K$, step size $\gamma > 0$, number of iterations T

1: **for** $t = 0, \dots, T-1$ **do**

2: Set $g_t = \begin{cases} \nabla F(x_t) = \frac{1}{n}\sum_{i=1}^n \nabla f_i(x_t) & \text{(deterministic case)} \\ \nabla f_{I_t}(x_t) \ , \quad \text{where } I_t \sim \mathrm{Unif}(\{1,\dots,n\}) & \text{(stochastic case)} \end{cases}$

3: Update $x_{t+1} \leftarrow x_t - \gamma P g_t$

4: If projection: project $x_{t+1} \leftarrow \Pi_K^{P^{-1}}(x_{t+1})$, with $\Pi_K^{P^{-1}}$ defined in (2.2)

5: **end for**

6: **return** x_T

Throughout the stability analysis, we condition on the neighboring datasets; expectations concern algorithmic randomness. In the statistical analysis in Section 5 and its proofs, expectations also include the sampled data and any independent replacement observation.

Preconditioned Gradient Descent. A limitation of gradient descent is its poor dependence on the problem's condition number (Bubeck, 2015). Preconditioned gradient descent (PGD) (Neymeyr, 2012; Zhang et al., 2021) aims to mitigate this issue by rebalancing the weight put on each direction of the space. PGD iteratively refines an iterate by performing gradient steps, rescaled by a matrix P. This gives the following two sequences of updates, depending on whether projection is used or not.

$$x_{t+1}^{\text{PGD}} = x_t^{\text{PGD}} - \gamma P \nabla F(x_t^{\text{PGD}}) \ , \qquad \tilde{x}_{t+1}^{\text{PGD}} = \tilde{x}_t^{\text{PGD}} - \gamma P \nabla \tilde{F}(\tilde{x}_t^{\text{PGD}}) \tag{PGD}$$

$$x_{t+1}^{\text{P-PGD}} = \Pi_K^{P^{-1}}(x_t^{\text{P-PGD}} - \gamma P \nabla F(x_t^{\text{P-PGD}})) \ , \quad \tilde{x}_{t+1}^{\text{P-PGD}} = \Pi_K^{P^{-1}}(\tilde{x}_t^{\text{P-PGD}} - \gamma P \nabla \tilde{F}(\tilde{x}_t^{\text{P-PGD}})), \tag{P$-$PGD}$$

where $\Pi_K^{P^{-1}}$ is the projection on K with respect to the geometry defined by P^{-1}:

$$\Pi_K^{P^{-1}}(v) = \arg\min_{y \in K} \tfrac{1}{2} \|y - v\|_{P^{-1}}^2 \ . \tag{2.2}$$

We analyze the sequence (PGD) in Section 3.1 and the sequence (P$-$PGD) in Section 4.1.

Preconditioned Stochastic Gradient Descent. More generally, in the stochastic case, if we consider indices $I_t \in \{1, \dots, n\}$ that are i.i.d. uniform over time and shared by both runs:

$$x_{t+1}^{\text{PSGD}} = x_t^{\text{PSGD}} - \gamma P \nabla f_{I_t}(x_t^{\text{PSGD}}) \ , \qquad \tilde{x}_{t+1}^{\text{PSGD}} = \tilde{x}_t^{\text{PSGD}} - \gamma P \nabla \tilde{f}_{I_t}(\tilde{x}_t^{\text{PSGD}}) \tag{PSGD}$$

$$x_{t+1}^{\text{P-PSGD}} = \Pi_K^{P^{-1}}(x_t^{\text{P-PSGD}} - \gamma P \nabla f_{I_t}(x_t^{\text{P-PSGD}})), \ \tilde{x}_{t+1}^{\text{P-PSGD}} = \Pi_K^{P^{-1}}(\tilde{x}_t^{\text{P-PSGD}} - \gamma P \nabla \tilde{f}_{I_t}(\tilde{x}_t^{\text{P-PSGD}})). \tag{P$-$PSGD}$$

We analyze the sequences (PSGD) in Section 3.2 and the sequence (P$-$PSGD) in Section 4.2. When the algorithm is clear from context, we suppress its superscript and write $x_t, \tilde{x}_t$ for the two coupled iterates. We give the pseudo-code for PGD and PSGD together with their projected variants in Algorithm 1.

Role of the conditioning matrices. The matrix P changes the trajectories of the iterates, but not their asymptotic limit. Thus, when the preconditioner P is properly chosen, this can significantly improve the speed of convergence, even in stochastic settings (Li et al., 2016), while still finding a solution of the problem at hand. We define the common metric and constants related to the matrix P.

$$\|u\|_{P^{-1}}^2 = u^\top P^{-1} u, \quad \chi(P) = \sqrt{\lambda_{\max}(P)/\lambda_{\min}(P)}. \tag{2.3}$$

Moreover, given a μ-strongly convex and L-smooth function, we define

$$\mu_P = \mu \lambda_{\min}(P), \quad L_P = L \lambda_{\max}(P), \quad 0 < \gamma \le L_P^{-1}, \quad \rho = 1 - \gamma \mu_P. \tag{2.4}$$

Notations. For two sequences $\{x_t\}_{t \ge 0}$ and $\{\tilde{x}_t\}_{t \ge 0}$, with reference points $x_\star$ and $\tilde{x}_\star$, we define $\Delta_t = x_t - \tilde{x}_t$ and $\Delta_\star = x_\star - \tilde{x}_\star$. In the applications below, the reference points $x_\star, \tilde{x}_\star$ are the empirical minimizers of $F, \tilde{F}$, respectively, over $\mathbb{R}^d$ in the regularized setting and over K in the projected setting. We also introduce the following integrated Hessian matrices,

$$H_i = \int_0^1 \nabla^2 f_i(\tilde{x}_\star + s\Delta_\star)\, ds, \qquad H_{i,t} = \int_0^1 \nabla^2 f_i(\tilde{x}_t + s\Delta_t)\, ds, \tag{2.5}$$

and their averaged counterparts, as well as the matrix A which is central to our decompositions:

$$H = \frac{1}{n}\sum_{i=1}^{n} H_i \quad , \quad \bar{H}_t = \frac{1}{n}\sum_{i=1}^{n} H_{i,t}, \quad , \quad A = I - \gamma P H. \tag{2.6}$$

We also introduce the replacement-gradient differences at $\widetilde{x}_t$ and $\widetilde{x}_\star$, together with the nonlinear part of the gradient:

$$\delta_t = \nabla f_j(\widetilde{x}_t) - \nabla \widetilde{f}_j(\widetilde{x}_t), \quad \delta_\star = \nabla f_j(\widetilde{x}_\star) - \nabla \widetilde{f}_j(\widetilde{x}_\star), \quad \eta_t = \nabla F(x_t) - \nabla F(\widetilde{x}_t) - H\Delta_t.$$

The regularized regime uses $f_i = \ell(\cdot, z_i) + r$ on $\mathbb{R}^d$; the projected regime uses full losses on a fixed compact convex set K. Their assumptions are specified separately below. In each regime, μ and L bound the full Hessians, while G bounds data-loss gradients in the regularized case and full gradients on K in the projected case. The same symbols are reused with these regime-specific meanings.

3 Regularized Preconditioned Gradient Descent

Consider the functions $F, \widetilde{F}$ defined in (2.1), where the loss f combines a data-dependent loss term ℓ and a data-independent regularizer $r\colon \mathbb{R}^d \to \mathbb{R}$:

$$f(x, z) = \ell(x, z) + r(x), \qquad x \in \mathbb{R}^d, \quad z \in \mathcal{Z}. \tag{3.1}$$

In this section, we use the following assumptions on the loss ℓ and the regularizer r.

A 3.1 (Regularizer)**.** The function r is C^2, μ-strongly convex and L_r-smooth.

A 3.2 (Convex Lipschitz)**.** For every z, $\ell(\cdot, z)$ is C^2, convex, L_ℓ-smooth and $\sup_x \|\nabla \ell(x, z)\| \le G$.

A 3.3 (Hessian-Lipschitz)**.** Each full Hessian $\nabla^2(\ell(\cdot, z) + r)$ is L_H-Lipschitz, uniformly in z.

We also introduce $L = L_\ell + L_r$. Under Assumptions 3.1 and 3.2, F and $\widetilde{F}$ admit unique minimizers on $\mathbb{R}^d$, denoted $x_\star$ and $\widetilde{x}_\star$, respectively. The gradient bound G applies to the data loss, not to the strongly convex full objective, which allows for globally unbounded gradients while still studying a strongly convex function on $\mathbb{R}^d$. The shared regularizer cancels-out from the replacement-gradient differences, so $\delta_t = \nabla \ell(\widetilde{x}_t, z_j) - \nabla \ell(\widetilde{x}_t, \widetilde{z}_j)$, and analogously for $\delta_\star$. Finally, we define:

$$D_0 = \sqrt{\lambda_{\max}(P)}\, \|x_0 - x_\star\|_{P^{-1}}, \quad \widetilde{D}_0 = \sqrt{\lambda_{\max}(P)}\, \|x_0 - \widetilde{x}_\star\|_{P^{-1}}, \quad D_\Sigma = D_0 + \widetilde{D}_0. \tag{3.2}$$

Throughout the section, proofs and precise bounds are given in Appendices A for the deterministic case and in B for the stochastic case.

3.1 Deterministic PGD: limiting displacement and transient drift

In this section, let $\{x_t\}_{t\ge 0}$ and $\{\widetilde{x}_t\}_{t\ge 0}$ denote the sequences of iterates of deterministic PGD without projection, as given by (PGD). We establish a decomposition of $\Delta_t = x_t - \widetilde{x}_t$, splitting PGD's stability into three parts.

Proposition 3.4 (Regularized deterministic decomposition)**.** *Under Assumptions 3.1 and 3.2,*

$$\Delta_t = \underbrace{(I - A^t)\Delta_\star}_{\text{intrinsic stability}} - \underbrace{\tfrac{\gamma}{n}\textstyle\sum_{k<t} A^{t-1-k} P(\delta_k - \delta_\star)}_{\text{replacement drift}} - \underbrace{\gamma \textstyle\sum_{k<t} A^{t-1-k} P \eta_k}_{\text{Hessian drift}}, \tag{3.3}$$

with Δ_t, $\Delta_\star$, δ_t, $\delta_\star$, η_t, and A as in Section 2 using the iterate sequences $\{x_t\}_{t\ge 0}$, $\{\widetilde{x}_t\}_{t\ge 0}$ of PGD.

The proof is given in Appendix A.2. This decomposition separates the main instability term $(I - A^t)\Delta_\star$ from the two other sums, which account for the accumulated errors and eventually vanish as the optimization procedure gets closer to the optimum. Note that, although this decomposition is inspired by (Sheshukova et al., 2025, Equation 28), it differs significantly as $A = I - \gamma P H$ where H is integrated between $x_\star$ and $\tilde{x}_\star$. We also use this decomposition for very different reasons: here, we aim to establish stability bounds, while Sheshukova et al. (2025) studies SGD's limit distribution.

We now introduce the term $\mathcal{B}_t = (I - A^t)\Delta_\star$, which represents the intrinsic instability of PGD. This is what we show in this next theorem.

Theorem 3.5 (Regularized deterministic stability)**.** *Under Assumptions 3.1, 3.2 and 3.3, with the Hessian Lipschitz assumption, we have*

$$\|\Delta_t - \mathcal{B}_t\| \le \left(\tfrac{2\gamma \lambda_{\max}(P) L_\ell \widetilde{D}_0}{n} + \tfrac{\gamma^2 \lambda_{\max}(P)^2 L_H G D_\Sigma}{2n} t \right) t \rho^{t-1}, \tag{3.4}$$

where $\|\mathcal{B}_t\| \le B_t := \min(2\gamma t G \lambda_{\max}(P)/n, 2G/(\mu n))$ for all $t \ge 0$. This upper bound is $O(1/n)$ uniformly in time for fixed problem constants and vanishes as $t \to \infty$.

The proof relies on bounding each term appearing in Theorem 3.4, and is given in Appendix A.3. Note that this result gives that $\|\Delta_t\| = O(\gamma t/n + \rho^t(\gamma t/n + \gamma^2 t^2/n))$, which recovers standard convex bounds as long as $\gamma t = O(1)$, while still obtaining $\|\Delta_t\| = O(G/(\mu n))$ as $t \to \infty$. This highlights the strength of our decomposition, which recovers both behaviors for small t and large t without dependence on μ_P.

3.2 Stochastic PSGD: centered fluctuations and nonlinear drift

In this section, let $\{x_t\}_{t\geq 0}$ and $\{\widetilde{x}_t\}_{t\geq 0}$ denote the sequences of iterates of stochastic PGD without projection, as given by (PSGD). We establish a decomposition of $\Delta_t = x_t - \widetilde{x}_t$, splitting PSGD's stability into five parts. The decomposition is similar to the one in the deterministic setting, with two additional terms due to stochasticity.

Proposition 3.6 (Regularized stochastic decomposition)**.** *The following identity holds:*

$$\Delta_t = \underbrace{(I - A^t)\Delta_\star}_{\textit{intrinsic stability}} - \underbrace{\gamma \textstyle\sum_{k<t} A^{t-1-k} P \zeta_k \delta_\star}_{\textit{replacement fluctuation}} - \underbrace{\gamma \textstyle\sum_{k<t} A^{t-1-k} P J_k (\delta_k - \delta_\star)}_{\textit{replacement drift}} \tag{3.5}$$
$$- \underbrace{\gamma \textstyle\sum_{k<t} A^{t-1-k} P U_k}_{\textit{Hessian drift}} - \underbrace{\gamma \textstyle\sum_{k<t} A^{t-1-k} P V_k}_{\textit{Hessian fluctuation}},$$

where we introduced $U_t = (H_{I_t,t} - H_{I_t})\Delta_t$, $V_t = (H_{I_t} - H)\Delta_t$, $J_t = \mathbf{1}_{\{I_t = j\}}$, and $\zeta_t = J_t - 1/n$, together with Δ_t, $\Delta_\star$, δ_t, $\delta_\star$, η_t, and A as in Section 2 using PSGD's iterate sequences x_t, and $\tilde{x}_t$.

The proof follows the same lines as the proof of Equation (3.4), further decomposing stochastic terms, and is given in Appendix B.2.

Theorem 3.7 (Regularized stochastic stability)**.** *Under Assumptions 3.1,3.2 and 3.3, including Lipschitz Hessians, if we define $\mathcal{B}_t = (I - A^t)\Delta_\star$, then*

$$\mathbb{E}\left[\|\Delta_t - \mathcal{B}_t\| \mid S, \widetilde{S}\right] \leq \left[\tfrac{2\gamma\lambda_{\max}(P)L_\ell \widetilde{D}_0}{n} + \gamma^2 \lambda_{\max}(P)^2 L_H G D_\Sigma \left(\tfrac{t}{2n} + \sqrt{\tfrac{t}{n}}\right)\right] t\rho^{t-1} \tag{3.6}$$
$$+ 2\lambda_{\max}(P) G \sqrt{\tfrac{\gamma(1-\rho^{2t})}{\mu_P n}} + \tfrac{2\lambda_{\max}(P)^2 G}{\mu_P}\left(2L_\ell + \tfrac{\lambda_{\max}(P) L_H G}{\mu_P}\right)\left(\tfrac{1}{n}\sqrt{\tfrac{\gamma}{\mu_P}} + \tfrac{\gamma}{\sqrt{n}}\right)\sqrt{1-\rho^{2t}}.$$

See Appendix B.3 for the proof of (3.6) and a more precise bound which uses both first-moment and martingale estimations, yielding a better dependence on n. Theorem 3.7 exhibits two types of terms: some that depend on μ_P and vanish when $t \to \infty$, and some that scale with γ and $1/\mu_P$, that are due to gradients' stochasticity. This is expected, as the preconditioner may affect the noise differently depending on its direction, but still disappears if the step size is decreased to zero. In particular,

$$\limsup_{\gamma \downarrow 0} \limsup_{t\to\infty} \mathbb{E}\,\|\Delta_t\| \leq \tfrac{2G}{\mu n}, \tag{3.7}$$

which confirms the soundness of our analysis, showing once again that PGD's limit does not depend on the preconditioner, up to noise terms that scale in γ.

4 Constrained setting

We consider the non-regularized case, where the problem is constrained to a nonempty compact convex set $K \subset \mathbb{R}^d$. We now assume that the functions $f(\cdot, z)$ are strongly convex and Lipschitz-continuous for all z, together with a Lipschitz Hessian.

A 4.1 (Loss)**.** For all z, $f(\cdot, z)$ is C^2, μ-strongly convex, G-Lipschitz, and L-smooth on K.

A 4.2 (Hessian-Lipschitz)**.** Each full Hessian $\nabla^2 f(\cdot, z)$ is L_H-Lipschitz on K, uniformly in z.

We take $x_0 = \tilde{x}_0 \in K$ and denote by $x_\star$ and $\widetilde{x}_\star$ the minimizers of F and $\widetilde{F}$ on K. The constants in (2.3) now use the constrained values of μ, L, G. The preconditioning P and the set K interact through $\Pi_K^{P^{-1}}$. Following common practice in constrained optimization (Beck & Teboulle, 2003; Bauschke & Combettes, 2011; Nemirovski et al., 2009; Bubeck, 2015), we use a projection in the P^{-1}-norm, which guarantees that the constrained minimizers remain fixed points of the update.

4.1 DETERMINISTIC PROJECTED PGD

In this section, let $\{x_t\}_{t\geq 0}$ and $\{\widetilde{x}_t\}_{t\geq 0}$ denote the sequences of iterates of P-PGD with projection on K, as given by (P−PGD). We establish a decomposition of $\Delta_t = x_t - \widetilde{x}_t$, splitting P-PGD's stability into four parts. Moreover, for any objective h, let the projection residual be

$$r_h(x) = \Pi_K^{P^{-1}}(x - \gamma P \nabla h(x)) - (x - \gamma P \nabla h(x)). \tag{4.1}$$

Center their difference as $c_t = r_F(x_t) - r_{\widetilde{F}}(\widetilde{x}_t) - \big(r_F(x_\star) - r_{\widetilde{F}}(\widetilde{x}_\star)\big)$. We first establish a decomposition of Δ_t similar to the one in the regularized setting, with an additional projection term.

Proposition 4.3 (Exact deterministic decomposition)**.** *For every $t \geq 0$,*

$$\Delta_t = \underbrace{(I - A^t)\Delta_\star}_{\textit{intrinsic stability}} - \underbrace{\tfrac{\gamma}{n}\textstyle\sum_{k<t} A^{t-1-k} P(\delta_k - \delta_\star)}_{\textit{replacement drift}} - \underbrace{\gamma \textstyle\sum_{k<t} A^{t-1-k} P \eta_k}_{\textit{Hessian drift}} + \underbrace{\textstyle\sum_{k<t} A^{t-1-k} c_k}_{\textit{projection drift}}, \tag{4.2}$$

with $\Delta_t, \Delta_\star, \delta_t, \delta_\star, \eta_t, H, A$ from Section 2.

See Appendix C.3 for the proof. This allows us to give an excess stability bound for projected PGD.

Theorem 4.4 (Deterministic projected stability)**.** *Under Assumptions 4.1 and 4.2, using the notations $D_0, \widetilde{D}_0$ and D_Σ from (3.2), it holds that*

$$\|\Delta_t - \mathcal{B}_t\| \leq \left(\tfrac{2\gamma\lambda_{\max}(P) L \widetilde{D}_0}{n} + \tfrac{\gamma^2 \lambda_{\max}(P)^2 L_H G D_\Sigma}{2n} t\right) t\rho^{t-1} + D_t^\Pi, \tag{4.3}$$

where D_t^Π is defined in Appendix C.4. In particular,

$$\|\Delta_t - \mathcal{B}_t\| \leq \left(\tfrac{2\gamma\lambda_{\max}(P) L \widetilde{D}_0}{n} + \tfrac{\gamma^2 \lambda_{\max}(P)^2 L_H G D_\Sigma}{2n} t + \rho D_\Sigma\right) t\rho^{t-1}. \tag{4.4}$$

Moreover, the bound $\|\mathcal{B}_t\| \leq B_t := 2G/(\mu n)$ still holds for any $t \geq 0$.

See Appendix C.5 for the proof. This result is very similar to the regularized case, up to an additional error term due to projection, which also vanishes as $t \to \infty$. Consequently, as in the regularized case, all the stability of projected PGD is captured by $\mathcal{B}_t$.

4.2 STOCHASTIC PROJECTED PGD

In this section, let $\{x_t\}_{t\geq 0}$ and $\{\widetilde{x}_t\}_{t\geq 0}$ denote the sequences of iterates of stochastic P-PSGD with projection on K, as given by (P−PSGD). We establish a decomposition of $\Delta_t = x_t - \widetilde{x}_t$, splitting P-PSGD's stability into six parts. Similarly to the deterministic projected case, we introduce the centered projection residual. However, this term also encompasses the fluctuations of the stochastic gradient,

$$c_t^{\mathrm{s}} = r_{f_{I_t}}(x_t) - r_{\widetilde{f}_{I_t}}(\widetilde{x}_t) - \big(r_F(x_\star) - r_{\widetilde{F}}(\widetilde{x}_\star)\big). \tag{4.5}$$

This gives the complete decomposition of the error for projected PSGD, similar to projected PGD with additional terms due to gradients' stochasticity.

Proposition 4.5 (Exact stochastic decomposition)**.** *The coupled difference satisfies*

$$\begin{aligned}\Delta_t = &\underbrace{(I - A^t)\Delta_\star}_{\textit{intrinsic stability}} - \underbrace{\gamma \textstyle\sum_{k<t} A^{t-1-k} P \zeta_k \delta_\star}_{\textit{replacement fluctuation}} - \underbrace{\gamma \textstyle\sum_{k<t} A^{t-1-k} P J_k (\delta_k - \delta_\star)}_{\textit{replacement drift}} \\ &- \underbrace{\gamma \textstyle\sum_{k<t} A^{t-1-k} P U_k}_{\textit{Hessian drift}} - \underbrace{\gamma \textstyle\sum_{k<t} A^{t-1-k} P V_k}_{\textit{Hessian fluctuation}} + \underbrace{\textstyle\sum_{k<t} A^{t-1-k} c_k^{\mathrm{s}}}_{\textit{projection drift}},\end{aligned} \tag{4.6}$$

where we introduced $U_t = (H_{I_t,t} - H_{I_t})\Delta_t$, $V_t = (H_{I_t} - H)\Delta_t$, $J_t = \mathbf{1}_{\{I_t = j\}}$, and $\zeta_t = J_t - 1/n$, together with Δ_t, $\Delta_\star$, δ_t, $\delta_\star$, η_t, and A as in Section 2 using projected PSGD's iterate sequences x_t, and $\tilde{x}_t$. The replacement and Hessian fluctuations are martingale convolutions. The projection residual is generally neither conditionally centered nor asymptotically zero at fixed step size.

See Appendix D.4 for a complete proof of this result. We now bound each of these terms, which gives our main result on the stability of projected PSGD.

Theorem 4.6 (Stochastic projected stability). *Under Assumptions 4.1 and 4.2, for $t \geq 1$,*

$$\mathbb{E}[\|\Delta_t - \mathcal{B}_t\|] \leq \mathcal{E}_t^{\mathrm{p}} + S_t^{\Pi}. \tag{4.7}$$

Here $\mathcal{E}_t^{\mathrm{p}}$ is the explicit right-hand side of (3.6) with L_ℓ replaced by L and all other constants taken from the constrained setting. Thus the gradient contribution has the same decaying transient and bounded-noise form. The additional projection coefficient S_t^{Π} is bounded as

$$\begin{aligned} S_t^{\Pi} \leq\ & (D_\Sigma + \chi(P)\,\|\Delta_\star\|)\rho^t + 2\gamma\lambda_{\max}(P)LD_\Sigma t\rho^{t-1} \\ & + \lambda_{\max}(P)\sqrt{\tfrac{\gamma}{\mu_P}}\left[2G + \tfrac{4\lambda_{\max}(P)LG}{\mu_P} + (L-\mu)\,\|\Delta_\star\| + \tfrac{\|\delta_\star\|}{\sqrt{n}}\right]\sqrt{1-\rho^{2t}}. \end{aligned} \tag{4.8}$$

We give a more precise expression for S_t^{Π} in Appendix D.5, which also gives $S_t^{\Pi} = O(1/n)$ uniformly in time and admissible step size, but losing the dependence on γ.

We prove this result in Appendix D.5 and Appendix D.6. This result shows that, even in the stochastic projected case, the stability of projected PSGD is eventually determined by $\mathcal{B}_t$. Specifically, we have

$$\limsup_{\gamma\downarrow 0}\limsup_{t\to\infty}\mathbb{E}[\|\Delta_t\|] \leq \tfrac{2G}{\mu n}, \tag{4.9}$$

recovering the correct asymptotic behavior even with stochastic gradients and projection.

5 Generalization and Excess Risk

We now leverage our results to derive generalization bounds for preconditioned gradient descent. Assume that $S = (z_1, \ldots, z_n)$ is i.i.d. from $\mathcal{D}$, that the loss is measurable in its data argument, and that x_0 and P are fixed independently of the data. For stochastic algorithms, the index sequence is independent of the data and is shared when comparing neighboring runs. Assume $\mathbb{E}_z|f(x_0, z)| < \infty$. Write $q \in \{\mathrm{r}, \mathrm{p}\}$ for the regularized and projected settings, respectively, and define

$$\mathcal{X}_{\mathrm{r}} = \mathbb{R}^d, \qquad \mathcal{X}_{\mathrm{p}} = K, \qquad \mathcal{R}(x) = \mathbb{E}_z f(x, z), \qquad x_{\mathrm{opt}} \in \arg\min_{x\in\mathcal{X}_q}\mathcal{R}(x). \tag{5.1}$$

In the regularized setting $f(x, z) = \ell(x, z) + r(x)$, and the population objective $\mathcal{R}$ includes the common regularizer. The constant G bounds data-loss gradients $\nabla\ell$ there, and full loss gradients ∇f_i on K in the projected setting.

Let $a \in \{\mathrm{d}, \mathrm{s}\}$ distinguish deterministic and stochastic updates, and write $x_t^{q,a}(S, \omega)$ for the corresponding iterate. The pairs (r, d), (r, s), (p, d), and (p, s) thus correspond to PGD, PSGD, P-PGD, and P-PSGD, respectively. Here $\omega = (I_k)_{k\geq 0}$ is the sampled-index sequence, shared by neighboring runs and independent of all data; for deterministic updates, ω is a trivial seed. For each $i \in [n]$, let $S^{(i)} = (z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n)$, where $z_i' \sim \mathcal{D}$ is independent of S and ω. We abbreviate $x_t^{q,a} = x_t^{q,a}(S, \omega)$.

Uniform parameter-stability bounds for all four settings. Let $x_r = \arg\min_{x\in\mathbb{R}^d} r(x)$ and set

$$D_{\mathrm{r}} = D_{\max} := \chi(P)(\|x_0 - x_r\| + G/\mu), \qquad D_{\mathrm{p}} := \chi(P)\,\mathrm{diam}(K). \tag{5.2}$$

Thus $D_0, \widetilde{D}_0 \leq D_q$ uniformly over neighboring datasets in the relevant setting. Use the optimization-noise radius already established in the stochastic proofs,

$$R_t = \gamma\lambda_{\max}(P)G\sqrt{(1-\rho^{2t})/(1-\rho^2)}. \tag{5.3}$$

Let $\beta_t^{\mathrm{r,d}}, \beta_t^{\mathrm{r,s}}, \beta_t^{\mathrm{p,d}}$ and $\beta_t^{\mathrm{p,s}}$ be the upper bounds on $\|\Delta_t\|$ obtained from Theorems 3.5, 3.7, 4.4, and 4.6 respectively. For every $q \in \{\mathrm{r}, \mathrm{p}\}$ and $a \in \{\mathrm{d}, \mathrm{s}\}$, $\mathbb{E}_\omega\left\|x_t^{q,a}(S, \omega) - x_t^{q,a}(S^{(i)}, \omega)\right\| \leq \beta_t^{q,a}$ uniformly over neighboring datasets and replacement indices.

Theorem 5.1 (Generalization and excess risk in the four settings). *Under Assumptions 3.1, 3.2 and 3.3 for regularized algorithms, or Assumptions 4.1 and 4.2 for projected algorithms, and the sampling conditions above, for every $t \geq 0$, the last-iterate excess-risk bounds in the regularized setting are, for* PGD *and* PSGD *respectively,*

$$\mathbb{E}\mathcal{R}(x_t^{\mathrm{r,d}}) - \mathcal{R}(x_{\mathrm{opt}}) \leq G\beta_t^{\mathrm{r,d}} + \tfrac{L}{2}D_{\mathrm{r}}^2\rho^{2t}, \quad \mathbb{E}\mathcal{R}(x_t^{\mathrm{r,s}}) - \mathcal{R}(x_{\mathrm{opt}}) \leq G\beta_t^{\mathrm{r,s}} + \tfrac{L}{2}\left(D_{\mathrm{r}}^2\rho^{2t} + R_t^2\right), \tag{5.4}$$

and similarly, for the projected case, we obtain the generalization bounds

$$\begin{aligned} \mathbb{E}\mathcal{R}(x_t^{\mathrm{p,d}}) - \mathcal{R}(x_{\mathrm{opt}}) &\leq G\beta_t^{\mathrm{p,d}} + \min\left\{GD_{\mathrm{p}}\rho^t, \tfrac{\mathrm{diam}(K)^2}{2\gamma\lambda_{\min}(P)}\rho^{2(t-1)}\right\}, \\ \mathbb{E}\mathcal{R}(x_t^{\mathrm{p,s}}) - \mathcal{R}(x_{\mathrm{opt}}) &\leq G\beta_t^{\mathrm{p,s}} + G\sqrt{D_{\mathrm{p}}^2\rho^{2t} + R_t^2}. \end{aligned} \tag{5.5}$$

The four proofs are in Appendix A.4 (regularized deterministic), Appendix B.4 (regularized stochastic), Appendix C.6 (projected deterministic), and Appendix D.7 (projected stochastic). Therefore, for either setting q, the bounds yield

$$\limsup_{t\to\infty}\big[\mathbb{E}\mathcal{R}(x_t^{q,\mathrm{d}})-\mathcal{R}(x_{\mathrm{opt}})\big]\leq(2G^2)/(\mu n),\ \limsup_{\gamma\to 0}\limsup_{t\to\infty}\big[\mathbb{E}\mathcal{R}(x_t^{q,\mathrm{s}})-\mathcal{R}(x_{\mathrm{opt}})\big]\leq(2G^2)/(\mu n).$$

It is also possible to derive bounds on the excess risk of the time-averaged parameter.

Theorem 5.2 (Averaged projected stochastic excess risk)**.** *For projected stochastic PGD, $T\geq 1$, and $\bar{x}_T=T^{-1}\sum_{t=0}^{T-1}x_t$, the additional averaged-output estimate is*

$$\mathbb{E}\mathcal{R}(\bar{x}_T)-\mathcal{R}(x_{\mathrm{opt}})\leq\tfrac{\mathrm{diam}(K)^2}{2\gamma\lambda_{\min}(P)T}+\tfrac{\gamma\lambda_{\max}(P)G^2}{2}+\tfrac{G}{T}\textstyle\sum_{t=0}^{T-1}\beta_t^{\mathrm{p,s}}. \tag{5.6}$$

See the proof in Appendix D.8 and (D.55). The final term is at most $2\lambda_{\max}(P)G^2/(\mu_P n)$, giving $O((\gamma T)^{-1}+\gamma+n^{-1})$ with the remaining parameters fixed, recovering the same asymptotic generalization guarantees as algorithms which do not use preconditioning (Hardt et al., 2016).

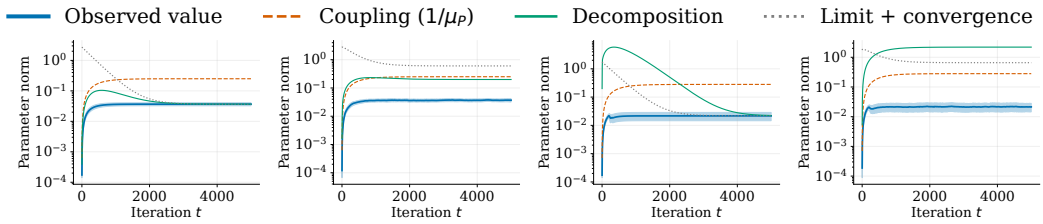
6 Related work

Linearization of constant-step-size dynamics. The technique behind our novel decomposition is inspired by the recent decomposition of SGD's error proposed in Sheshukova et al. (2025, Eq. (28)). This method allows us to isolate the linear part of the gradient from its higher-order components, and is heavily inspired by the stochastic optimization literature (Moulines & Bach, 2011; Dieuleveut et al., 2020; Gadat & Panloup, 2023). While such ideas have been used to study the stochastic component of SGD, we propose a surprising novel application of this method beyond stochastic optimization, providing new algorithmic stability even in the deterministic setting.

Stability of gradient and stochastic-gradient methods. The modern stability framework originates with the sensitivity-based analysis of Bousquet & Elisseeff (2002). Stability for randomized algorithms was developed by Elisseeff et al. (2005), while Shalev-Shwartz et al. (2010) clarified the relation between stability, learnability, and uniform convergence. Hardt et al. (2016) established the influential coupling analysis of SGD under smoothness and Lipschitz assumptions, which were later refined by: Kuzborskij & Lampert (2018) with data-dependent stability bounds; Lei & Ying (2020) with on-average model stability and relaxed bounded-gradient and smoothness assumptions; and Zhang et al. (2022) with tightness and data-dependent improvements of classical SGD stability estimates. Stability was also studied in non-smooth convex functions (Bassily et al., 2020; Harvey et al., 2019), non-convex, non-smooth ones (Lei, 2023), and decentralized settings (Le Bars et al., 2024; Sun et al., 2024). Zeng & Lei (2026) treated stability for biased gradient estimators. Another line of work related stability to optimization, rather than propagating a perturbation step by step. Charles & Papailiopoulos (2018) derived stability from stability of the global optimum together with convergence rates, while Lei & Ying (2020) directly related stability to optimization, exploiting algorithms' convergence guarantees. Although seemingly similar, our method profoundly differs from these approaches: first, our bounds are informative even in the small iteration regimes; second, they do not rely on any result on the optimum's stability; third, optimization bounds are only used to control higher-order terms, while we give an explicit expression of the non-vanishing instability term.

Preconditioning methods. Preconditioned methods have a long history in optimization (Bramble et al., 1996; Knyazev & Neymeyr, 2003; 2009), and regained interest in the past years with recent success of preconditioned gradient descent (Neymeyr, 2012; Zhang et al., 2021) and its stochastic variants (Li, 2017; Li et al., 2016; Gupta et al., 2018). They are also related to adaptive methods such as AdaGrad, Adam and their variants (Duchi et al., 2011; Defazio et al., 2022; Kingma & Ba, 2014), which use adaptive and diagonal preconditioners. In generalized linear models, Gonen & Shalev-Shwartz (2018) proved invariance of average stability under data preconditioning: this strongly differs from our setting, where the preconditioner acts on the dynamics of the optimization algorithm, not on the statistical objective, which remains unchanged.

Stability of preconditioning methods. Despite the strong interest of the community in preconditioning methods, little attention has been put on studying their stability. A notable exception is the work of Vary et al. (2026). Their results establish the stability of PSGD with decreasing step size. However, they retain an asymptotic dependence on the preconditioner, even in the deterministic case.



(a) Regularized PGD. (b) Regularized PSGD. (c) Projected PGD. (d) Projected PSGD.

Figure 2: Observed parameter stability and three upper bounds for logistic regression in the four settings. Observations are averaged over 128 datasets. We represent mean ± 1.96 standard deviations.

This is due to the use of the contraction method to derive stability, following classical analyses of SGD's stability. Our method is radically different, and relies on a careful decomposition of the error between intrinsic stability discrepancies and higher-order error terms which eventually vanish.

7 Numerical Experiments

We consider binary logistic regression with sample objectives

$$f(x;(a,y)) = \log(1+\exp(a^\top x)) - y\, a^\top x + \tfrac{\lambda}{2}\|x\|^2, \qquad y \in \{0,1\}.$$

We compare deterministic and stochastic updates, regularized and projected. Each experiment uses datasets differing in one observation, and a fixed step size γ. The stability $\|\Delta_t\|$ is averaged over 128 datasets. Experimental details and additional results are given in Appendix E.

Figure 2 compares the observed stability with three upper bounds: direct coupling, stability of the limiting minimizers combined with optimization error, and our decomposition bound. These estimates capture different aspects of the dynamics. Direct coupling accounts for the accumulation of sample perturbations but does not explicitly identify the limiting displacement. The convergence-based estimate captures the eventual behavior, although its initial-distance terms can be large early in training. Our decomposition instead isolates the response $\mathcal{B}_t = (I - A^t)\Delta_\star$ and bounds the remaining transient and stochastic contributions separately.

The comparisons illustrate both the usefulness and the limitations of this separation. Its relative sharpness depends on the regime and the iteration; it does not necessarily improve both alternatives uniformly. In particular, bounding the contributions separately can lose cancellations and produce conservative estimates when projection is active. Finally, the stochastic bounds retain fluctuation terms at fixed step size, unlike the decaying deterministic remainders.

8 Conclusion

We established novel stability bounds for preconditioned gradient descent, with deterministic and stochastic gradients, for functions defined either on $\mathbb{R}^d$ or on a compact subset of $\mathbb{R}^d$. Our analysis is based on a stability argument which decomposes the stability in two parts: one part capturing intrinsic stability, and another transient term which eventually vanishes with iterations, or if the step size is sufficiently small when gradients are stochastic. This refined analysis allows us to account for the fact that preconditioning changes the trajectory of the algorithm, affecting its stability. Interestingly, our analysis points out the fact that a poor choice of preconditioner, which may arise in practice, can exacerbate initial algorithm's error. Nonetheless, this only affects initial iterations, and the algorithm eventually recovers from this initial error, enjoying stability bounds that are independent of the preconditioner. To our knowledge, the methods exploited here to derive our theoretical bounds had never been used in algorithmic stability analysis. Yet, sharpening the residual terms and obtaining simpler bounds across all settings remains a promising direction for future work. Finally, our analysis focuses on fixed preconditioners, which does not allow for application to all types of preconditioning methods. Extending our approach to adaptive methods such as Adam would require accounting for the sensitivity of the preconditioner to the training data. More broadly, separating transient dynamics from intrinsic stability may provide a useful approach to studying momentum-based methods.

AI USE STATEMENT

In this work, we used generative AI tools for code generation, for improving the fluency and clarity of the writing, and for spotting typos.

We have reviewed all AI-assisted work. All AI-generated code was read, verified, and tested for correctness by the authors; all AI-suggested edits to the text were inspected individually and accepted or rejected by the authors, and no text was incorporated without author review.

We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. *48th Annual Symposium on the Theory of Computing*, 2016. URL https://arxiv.org/abs/1511.02513v1.

Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. *Advances in Neural Information Processing Systems*, 33: 4381–4391, 2020.

H.H. Bauschke and Patrick Louis Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2011. doi: 10.1007/978-1-4419-9467-7. URL https://inria.hal.science/hal-00643354.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003. ISSN 0167-6377. doi: https://doi.org/10.1016/S0167-6377(02)00231-6. URL https://www.sciencedirect.com/science/article/pii/S0167637702002316.

Olivier Bousquet and André Elisseeff. Stability and generalization. *J. Mach. Learn. Res.*, 2:499–526, 2002. URL https://api.semanticscholar.org/CorpusID:1157797.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. doi: 10.1017/CBO9780511804441.

James H Bramble, Joseph E Pasciak, and Andrew V Knyazev. A subspace preconditioning algorithm for eigenvector/eigenvalue computation. *Advances in computational mathematics*, 6(1):159–189, 1996.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and trends in Machine Learning*, 8(3-4):231–357, 2015.

Zachary Charles and Dimitris Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In *International conference on machine learning*, pp. 745–754. PMLR, 2018.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(29):1069–1109, 2011. URL http://jmlr.org/papers/v12/chaudhuri11a.html.

Aaron Defazio, Baoyu Zhou, and Lin Xiao. Grad-gradagrad? a non-monotone adaptive stochastic gradient method. *arXiv preprint arXiv:2206.06900*, 2022.

Aymeric Dieuleveut, Alain Durmus, and Francis Bach. Bridging the gap between constant step size stochastic gradient descent and markov chains. *The Annals of Statistics*, 48(3):1389–1401, 2020. doi: 10.1214/19-AOS1850. URL https://doi.org.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011. URL http://jmlr.org/papers/v12/duchi11a.html.

Andre Elisseeff, Theodoros Evgeniou, and Massimiliano Pontil. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6(3):55–79, 2005. URL http://jmlr.org/papers/v6/elisseeff05a.html.

Sébastien Gadat and Fabien Panloup. Optimal non-asymptotic analysis of the ruppert–polyak averaging stochastic algorithm. *Stochastic Processes and their Applications*, 156:312–348, 2023.

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.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pp. 1842–1850. PMLR, 2018.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pp. 1225–1234. PMLR, 2016.

Nicholas JA Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory*, pp. 1579–1613. PMLR, 2019.

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

Andrew V Knyazev and Klaus Neymeyr. A geometric theory for preconditioned inverse iteration iii: A short and sharp convergence estimate for generalized eigenvalue problems. *Linear Algebra and its Applications*, 358(1-3):95–114, 2003.

Andrew V Knyazev and Klaus Neymeyr. Gradient flow approach to geometric convergence analysis of preconditioned eigensolvers. *SIAM journal on matrix analysis and applications*, 31(2):621–628, 2009.

Ilja Kuzborskij and Christoph Lampert. Data-dependent stability of stochastic gradient descent. In *international conference on machine learning*, pp. 2815–2824. PMLR, 2018.

Batiste Le Bars, Aurélien Bellet, Marc Tommasi, Kevin Scaman, and Giovanni Neglia. Improved stability and generalization guarantees of the decentralized sgd algorithm. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Yunwen Lei. Stability and generalization of stochastic optimization with nonconvex and nonsmooth problems. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 191–227. PMLR, 2023.

Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *international conference on machine learning*, pp. 5809–5819. PMLR, 2020.

Chunyuan Li, Changyou Chen, David Carlson, and Lawrence Carin. Preconditioned stochastic gradient langevin dynamics for deep neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Xi-Lin Li. Preconditioned stochastic gradient descent. *IEEE transactions on neural networks and learning systems*, 29(5):1454–1466, 2017.

Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24, 2011.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009. doi: 10.1137/070704277. URL https://doi.org/10.1137/070704277.

Klaus Neymeyr. A geometric convergence theory for the preconditioned steepest descent iteration. *SIAM Journal on Numerical Analysis*, 50(6):3188–3207, 2012.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, second edition, 2006. doi: 10.1007/978-0-387-40065-5.

Ankan Saha, Prateek Jain, and Ambuj Tewari. The interplay between stability and regret in online learning. *arXiv preprint arXiv:1211.6158*, 2012.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11(90):2635–2670, 2010. URL http://jmlr.org/papers/v11/shalev-shwartz10a.html.

Marina Sheshukova, Denis Belomestny, Alain Oliviero Durmus, Eric Moulines, Aleksei Naumov, and Sergey Samsonov. Nonasymptotic analysis of stochastic gradient descent with the richardson–romberg extrapolation. In *International Conference on Learning Representations*, volume 2025, pp. 33019–33049, 2025.

Zhenyu Sun, Xiaochun Niu, and Ermin Wei. Understanding generalization of federated learning via stability: Heterogeneity matters. In *International conference on artificial intelligence and statistics*, pp. 676–684. PMLR, 2024.

Enayat Ullah, Tung Mai, Anup Rao, Ryan A. Rossi, and Raman Arora. Machine unlearning via algorithmic stability. In Mikhail Belkin and Samory Kpotufe (eds.), *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 4126–4142. PMLR, 15–19 Aug 2021. URL https://proceedings.mlr.press/v134/ullah21a.html.

Simon Vary, Tyler Farghly, Ilja Kuzborskij, and Patrick Rebeschini. On-average stability of multipass preconditioned sgd and effective dimension. *arXiv preprint arXiv:2603.11989*, 2026.

Shuang Zeng and Yunwen Lei. Stochastic gradient methods: Bias, stability and generalization. *Journal of Machine Learning Research*, 27(6):1–55, 2026. URL http://jmlr.org/papers/v27/24-0637.html.

Jialun Zhang, Salar Fattahi, and Richard Y Zhang. Preconditioned gradient descent for over-parameterized nonconvex matrix factorization. *Advances in Neural Information Processing Systems*, 34:5985–5996, 2021.

Yikai Zhang, Wenjia Zhang, Sammy Bald, Vamsi Pingali, Chao Chen, and Mayank Goswami. Stability of sgd: Tightness analysis and improved bounds. In *Uncertainty in artificial intelligence*, pp. 2364–2373. PMLR, 2022.

A PROOFS FOR DETERMINISTIC REGULARIZED PGD

A.1 GENERALITIES

For any full local or empirical objective h, the integral Hessian along a segment lies between μI and LI. Under $x' = P^{-1/2}x$, the gradient map has derivative $I - \gamma P^{1/2}\nabla^2 h(x)P^{1/2}$ with spectrum in $[0, \rho]$. Consequently

$$\|x - y - \gamma P(\nabla h(x) - \nabla h(y))\|_{P^{-1}} \le \rho \|x - y\|_{P^{-1}}. \tag{A.1}$$

The same similarity transform gives $\|A^q P\| \le \lambda_{\max}(P)\rho^q$ and $\|A^q\| \le \chi(P)\rho^q$ for any $q \ge 0$.

Stationarity gives $\nabla F(\widetilde{x}_\star) = \delta_\star/n$ and $\nabla F(x_\star) = 0$, hence $H\Delta_\star = -\delta_\star/n$. Since $H \succeq \mu I$ and $\|\delta_\star\| \le 2G$, $\|\Delta_\star\| \le 2G/(\mu n)$.

For $g_i = \nabla \ell(x, z_i)$, cancellation of the regularizer and the variance identity give

$$\mathbb{E}_I \|\nabla f_I(x) - \nabla F(x)\|^2 = \frac{1}{n}\sum_i \|g_i\|^2 - \left\|\frac{1}{n}\sum_i g_i\right\|^2 \le G^2. \tag{A.2}$$

A.2 PROOF OF PROPOSITION 3.4

Proof. Subtracting the updates gives

$$\begin{aligned}
\Delta_{k+1} &= x_k - \widetilde{x}_k - \gamma P(\nabla F(x_k) - \nabla \widetilde{F}(\widetilde{x}_k)) && \text{(A.3)}\\
&= \Delta_k - \gamma P(\nabla F(x_k) - \nabla F(\widetilde{x}_k)) - \gamma P(\nabla F(\widetilde{x}_k) - \nabla \widetilde{F}(\widetilde{x}_k)) && \text{(A.4)}\\
&= \Delta_k - \gamma P(\eta_k + H\Delta_k) - \frac{\gamma}{n} P\delta_k && \text{(A.5)}\\
&= A\Delta_k - \gamma P\eta_k - \frac{\gamma}{n} P\delta_k. && \text{(A.6)}
\end{aligned}$$

Use $-(\gamma/n)P\delta_\star = (I - A)\Delta_\star$ and unroll from $\Delta_0 = 0$ to obtain (3.3). □

A.3 PROOF OF THEOREM 3.5

Define

$$a_{1,t} = \frac{2\gamma\lambda_{\max}(P)G}{n}\sum_{k<t}\rho^k = \frac{2\lambda_{\max}(P)G}{\mu_P n}(1 - \rho^t) \le \frac{2\gamma\lambda_{\max}(P)G}{n}t. \tag{A.7}$$

First, using contraction and the perturbation bound $2G/n$, we obtain

$$\|\Delta_k\| \le \sqrt{\lambda_{\max}(P)}\,\|\Delta_k\|_{P^{-1}} \le a_{1,k}, \qquad \|x_k - x_\star\| \le D_0\rho^k, \quad \|\widetilde{x}_k - \widetilde{x}_\star\| \le \widetilde{D}_0\rho^k. \tag{A.8}$$

For $t \ge 1$, the full finite-time stability bound given by Theorem 3.5, which we prove below, is

$$\|\Delta_t\| \le \|\mathcal{B}_t\| + \underbrace{\frac{2\gamma\lambda_{\max}(P)L_\ell\widetilde{D}_0}{n}t\rho^{t-1}}_{=D_t^{\delta,\mathrm{r}}} + \underbrace{\frac{\gamma\lambda_{\max}(P)^2 L_H G D_\Sigma}{\mu_P n}t\rho^{t-1}}_{=D_t^{H,\mathrm{r}}}. \tag{A.9}$$

Proof of Theorem 3.5. We bound each term separately.

Bounding $\|\mathcal{B}_t\|$. Let $M = H^{1/2}PH^{1/2}$. Its eigenvalues lie in $[\mu_P, L_P]$, so $0 \preceq I - (I - \gamma M)^t \preceq I$. Thus

$$\|\mathcal{B}_t\| = \frac{1}{n}\left\|H^{-1/2}[I - (I - \gamma M)^t]H^{-1/2}\delta_\star\right\| \le \frac{1}{n}\left\|H^{-1}\right\|\|\delta_\star\| \le \frac{2G}{\mu n}. \tag{A.10}$$

This argument uses unconstrained stationarity and does not work at a boundary optimum.

We can also derive the alternative upper bound:

$$\begin{aligned}\|\mathcal{B}_t\| &= \left\|(I-(I-\gamma PH)^t)\Delta_\star\right\| \\ &= \left\|\gamma\sum_{k=0}^{t-1}(I-\gamma PH)^k PH\Delta_\star\right\| \\ &\le \gamma\sum_{k=0}^{t-1}\left\|(I-\gamma PH)^k P\right\|\frac{2G}{n} \\ &\le \gamma\lambda_{\max}(P)\sum_{k=0}^{t-1}\rho^k\frac{2G}{n} \\ &\le \frac{2\gamma G\lambda_{\max}(P)}{n}t.\end{aligned}$$

Proving the $D_t^{\delta,\mathrm{r}}$ bound (replacement drift). Cancellation of r gives $\|\delta_k-\delta_\star\|\le 2L_\ell\widetilde{D}_0\rho^k$. Using the fact that $\left\|A^{t-1-k}P\right\|\le\lambda_{\max}(P)\rho^{t-1-k}$ we get

$$\left\|\frac{\gamma}{n}\sum_{k<t}A^{t-1-k}P(\delta_k-\delta_\star)\right\| \le 2L_\ell\widetilde{D}_0\frac{\gamma}{n}\lambda_{\max}(P)\sum_{k<t}\rho^{t-1-k}\rho^k \tag{A.11}$$

$$= \frac{2\gamma\lambda_{\max}(P)L_\ell\widetilde{D}_0}{n}t\rho^{t-1}. \tag{A.12}$$

Proving the $D_t^{H,\mathrm{r}}$ bound (Hessian drift). For $\bar{H}_k=\int_0^1\nabla^2F(\widetilde{x}_k+s\Delta_k)\,ds$, compare corresponding points on the two segments to obtain

$$\left\|\bar{H}_k-H\right\|\le\frac{L_H}{2}(\|x_k-x_\star\|+\|\widetilde{x}_k-\widetilde{x}_\star\|)\le\frac{L_H}{2}D_\Sigma\rho^k. \tag{A.13}$$

Since $\eta_k=(\bar{H}_k-H)\Delta_k$, its norm is at most $L_HD_\Sigma\rho^ka_{1,k}/2$. Using the fact that $\left\|A^{t-1-k}P\right\|\le\lambda_{\max}(P)\rho^{t-1-k}$, we have

$$\left\|\gamma\sum_{k<t}A^{t-1-k}P\eta_k\right\|\le\gamma\lambda_{\max}(P)\frac{L_HD_\Sigma}{2}\sum_{k<t}\rho^{t-1-k}\rho^ka_{1,k} \tag{A.14}$$

$$=\gamma\lambda_{\max}(P)\frac{L_HD_\Sigma}{2}\rho^{t-1}\sum_{k<t}a_{1,k}. \tag{A.15}$$

Using $a_{1,k}\le\dfrac{2\lambda_{\max}(P)G}{\mu_Pn}$ yields

$$\left\|\gamma\sum_{k<t}A^{t-1-k}P\eta_k\right\|\le\frac{\gamma\lambda_{\max}(P)^2L_HGD_\Sigma}{\mu_Pn}t\rho^{t-1}. \tag{A.16}$$

Using instead $a_{1,k}\le\dfrac{2\gamma\lambda_{\max}(P)G}{n}k$ yields

$$\left\|\gamma\sum_{k<t}A^{t-1-k}P\eta_k\right\|\le\frac{\gamma^2\lambda_{\max}(P)^2L_HGD_\Sigma}{2n}t^2\rho^{t-1}. \tag{A.17}$$

Without Hessian Lipschitzness, recenter the two gradient differences at their minimizing pair. Smoothness gives $\|\eta_k\|\le 2LD_\Sigma\rho^k$, hence its convolution is at most $2\gamma\lambda_{\max}(P)LD_\Sigma t\rho^{t-1}$ and still vanishes. Direct coupling and convergence give

$$\|\Delta_t\|\le\min\{a_{1,t},\,2G/(\mu n)+D_\Sigma\rho^t\}.$$

Moreover $D_\Sigma\le 2\chi(P)(\|x_0-x_r\|+G/\mu)$, by the minimizer-radius bound above.

These results prove (A.9). Both sums vanish as $t\to\infty$. Also $A^t\Delta_\star\to 0$, proving $\Delta_t\to\Delta_\star$. For fixed γ, $\sup_{t\ge1}t\rho^{t-1}<\infty$; the uniform radius bound above proves the uniform $O(1/n)$ claims. □

A.4 Deterministic regularized statistical guarantees

We prove the deterministic part of Theorem 5.1.

Proof. The bounded data gradient makes $\ell(\cdot, z)$ G-Lipschitz. The anchor integrability assumption therefore gives a finite population data loss at each fixed point. Adding r makes the population objective strongly convex and coercive, so it has a unique minimizer.

Let $S^{(i)}$ replace observation i with an independent copy. Exchangeability and cancellation of the common regularizer give

$$\mathbb{E}[\mathcal{R}(x_t) - F_S(x_t)] = \frac{1}{n}\sum_i \mathbb{E}[\ell(x_t(S^{(i)}), z_i) - \ell(x_t(S), z_i)].$$

Lipschitzness and the deterministic stability estimate bound the absolute value by $G\beta_t^{\mathrm{r,d}}$. Unconstrained stationarity and smoothness give

$$F_S(x_t) - F_S(x_\star) \le \frac{L}{2}\|x_t - x_\star\|^2 \le \frac{L}{2} D_{\max}^2 \rho^{2t}.$$

Finally, split the expected excess population risk into:

$$\mathbb{E}\mathcal{R}(x_t) - \mathcal{R}(x_{\mathrm{opt}}) \tag{A.18}$$

$$= \mathbb{E}[\mathcal{R}(x_t) - F_S(x_t)] + \mathbb{E}[F_S(x_t) - F_S(x_\star)] + \underbrace{\mathbb{E}[F_S(x_\star) - F_S(x_{\mathrm{opt}})]}_{\le 0} \tag{A.19}$$

$$\le G\beta_t^{\mathrm{r,d}} + \frac{L}{2} D_{\max}^2 \rho^{2t}. \tag{A.20}$$

As $t \to \infty$, its upper bound tends to at most $2G^2/(\mu n)$. □

B Proofs for stochastic regularized PGD

B.1 Stochastic moment envelopes and optimization error

All auxiliary sequences used in the stochastic proofs are collected here:

$$s_t = \sum_{k<t} \rho^{t-1-k} = \frac{1-\rho^t}{1-\rho}, \qquad v_t = \sum_{k<t} \rho^{2(t-1-k)} = \frac{1-\rho^{2t}}{1-\rho^2}, \qquad a_{1,t} = \frac{2\gamma\lambda_{\max}(P)G}{n} s_t. \tag{B.1}$$

$$R_t = \gamma\lambda_{\max}(P)G\sqrt{v_t}, \qquad R_\infty = \frac{\gamma\lambda_{\max}(P)G}{\sqrt{1-\rho^2}}. \tag{B.2}$$

$$a_{2,\infty} = 2\gamma\lambda_{\max}(P)G\sqrt{\frac{1}{n^2(1-\rho)^2} + \frac{1}{n(1-\rho^2)}}. \tag{B.3}$$

The sums vanish at $t = 0$; $0^0 = 1$ is used when $\rho = 0$.

The same contraction argument as in the deterministic setting yields

$$\|\Delta_t\| \le \sqrt{\lambda_{\max}(P)}\,\|\Delta_t\|_{P^{-1}} \le 2\gamma\lambda_{\max}(P)G\sum_{k<t}\rho^{t-1-k}J_k. \tag{B.4}$$

Since the J_k are independent Bernoulli variables of mean $1/n$, the first moment is at most $a_{1,t}$ and the second moment is at most $a_{2,t}^2$, where

$$a_{2,t} = 2\gamma\lambda_{\max}(P)G\left[\frac{(1-\rho^t)^2}{n^2(1-\rho)^2} + \frac{1}{n}\left(1-\frac{1}{n}\right)\frac{1-\rho^{2t}}{1-\rho^2}\right]^{1/2}. \tag{B.5}$$

To control optimization error, write $\xi_k = \nabla f_{I_k}(x_k) - \nabla F(x_k)$. Conditionally on $\mathcal{F}_k$, this is mean zero and has second moment at most G^2. Expanding the squared metric distance to the stationary point $x_\star$, the cross term vanishes and (A.1) gives

$$\mathbb{E}[\|x_{k+1} - x_\star\|_{P^{-1}}^2 \mid \mathcal{F}_k] \le \rho^2\|x_k - x_\star\|_{P^{-1}}^2 + \gamma^2\lambda_{\max}(P)G^2. \tag{B.6}$$

Iteration and conversion to Euclidean norm give

$$\mathbb{E}\,\|x_t - x_\star\|^2 \le D_0^2\rho^{2t} + R_t^2, \qquad (\mathbb{E}\,\|x_t - x_\star\|^2)^{1/2} \le D_0\rho^t + R_t. \tag{B.7}$$

The same holds for the altered trajectory. Taking the triangle inequality through both minimizers yields the convergence envelope $\mathbb{E}\,\|\Delta_t\| \le 2G/(\mu n) + D_\Sigma\rho^t + 2R_t$.

Lemma B.1. *For deterministic matrices A, P and martingale differences M_k, it holds that*

$$\mathbb{E}\left\|\gamma\sum_{k<t} A^{t-1-k}PM_k\right\| \le \gamma\lambda_{\max}(P)\left(\sum_{k<t}\rho^{2(t-1-k)}\mathbb{E}\,\|M_k\|^2\right)^{1/2}. \tag{B.8}$$

Proof. Indeed, we first apply Jensen's inequality, and notice that for $r < k$ the earlier increment and both matrices are measurable while $\mathbb{E}[M_k \mid \mathcal{F}_k] = 0$. □

B.2 Proof of Proposition 3.6

Proof. The proof is similar to the one in the deterministic setting.

Subtracting the stochastic updates gives $\Delta_{k+1} = A\Delta_k - \gamma P(U_k + V_k) - \gamma P J_k\delta_k$. Expand $J_k\delta_k = \delta_\star/n + \zeta_k\delta_\star + J_k(\delta_k - \delta_\star)$ and use $H\Delta_\star = -\delta_\star/n$ to obtain (3.5). □

B.3 Proof of Theorem 3.7

Set $\Lambda = L - \mu$, $d_\star = \|\delta_\star\| \le 2G$. We give slightly more precise bounds than the ones used in Theorem 3.7:

$$W_t^\delta = 2\gamma\lambda_{\max}(P)G\min\left\{\frac{2}{n}\left(1-\frac{1}{n}\right)\frac{1-\rho^t}{1-\rho},\quad \sqrt{\frac{1}{n}\left(1-\frac{1}{n}\right)\frac{1-\rho^{2t}}{1-\rho^2}}\right\}, \tag{B.9}$$

$$S_t^\delta = \frac{2\gamma\lambda_{\max}(P)L_\ell}{n}\sum_{k<t}\rho^{t-1-k}(\widetilde{D}_0\rho^k + R_k),$$

$$S_t^H = \gamma\lambda_{\max}(P)\sum_{k<t}\rho^{t-1-k}\min\left\{\frac{L_H}{2}(D_\Sigma\rho^k + 2R_k)a_{2,k},\ \Lambda a_{1,k}\right\},$$

$$W_t^H = \gamma\lambda_{\max}(P)L_\ell\min\left\{\sum_{k<t}\rho^{t-1-k}a_{1,k},\quad \left(\sum_{k<t}\rho^{2(t-1-k)}a_{2,k}^2\right)^{1/2}\right\}.$$

Without Hessian Lipschitzness, retain only $\Lambda a_{1,k}$ in S_t^H; the associated small-step rate for that separated drift is then not claimed.

The four coefficients in (B.9) bound the four expected convolution norms in (3.5); their sum bounds $\mathbb{E}\,\|\Delta_t - \mathcal{B}_t\|$. For $t \ge 1$ they satisfy

$$W_t^\delta \le 2\gamma\lambda_{\max}(P)G\min\left\{\frac{2(1-1/n)}{n}s_t, \sqrt{\frac{1-1/n}{n}v_t}\right\}, \tag{B.10}$$

$$S_t^\delta \le \frac{2\gamma\lambda_{\max}(P)L_\ell}{n}(\widetilde{D}_0 t\rho^{t-1} + R_\infty s_t),$$

$$S_t^H \le \frac{\gamma\lambda_{\max}(P)L_H a_{2,\infty}}{2}(D_\Sigma t\rho^{t-1} + 2R_\infty s_t),$$

$$W_t^H \le \gamma\lambda_{\max}(P)L_\ell\min\left\{\frac{2\lambda_{\max}(P)G}{\mu_P n}s_t, a_{2,\infty}\sqrt{v_t}\right\}.$$

Alternatively, removing inverse curvature from the transient gives

$$S_t^H \le \gamma^2\lambda_{\max}(P)^2 L_H G D_\Sigma\left(\frac{t^2}{2n} + \sqrt{\frac{1-1/n}{n}}t^{3/2}\right)\rho^{t-1} + \gamma\lambda_{\max}(P)L_H R_\infty a_{2,\infty}s_t. \tag{B.11}$$

The componentwise large-time orders, with constants independent of n and admissible γ, are

$$W_t^\delta : O(\min\{n^{-1}, \sqrt{\gamma/n}\}), \qquad S_t^\delta : O(\sqrt{\gamma}/n), \tag{B.12}$$
$$S_t^H : O(\min\{n^{-1}, \sqrt{\gamma}/n + \gamma/\sqrt{n}\}), \qquad W_t^H : O(\min\{n^{-1}, \sqrt{\gamma}/n + \gamma/\sqrt{n}\}).$$

Proof of Theorem 3.7.

Proving the W_t^δ bound (replacement fluctuation). The first entry of W_t^δ is given by the triangle inequality, and the second one is obtained using (B.8) and $\mathbb{E}|\zeta_k| = 2(1-1/n)/n$, $\mathbb{E}\zeta_k^2 = (1-1/n)/n$.

Using $\|\delta_\star\| \le 2G$ and keeping the two geometric sums gives the W_t^δ line of (B.10).

Proving the S_t^δ bound (replacement drift). Independence of J_k from $\mathcal{F}_k$ and (B.7) give $\mathbb{E}[J_k \|\delta_k - \delta_\star\|] \le (2L_\ell/n)(\widetilde{D}_0\rho^k + R_k)$, proving the expression of S_t^δ.

To prove its explicit form, use $R_k \le R_\infty$ but retain the factor ρ^k in the initial-distance term:

$$\sum_{k<t} \rho^{t-1-k}(\widetilde{D}_0\rho^k + R_k) \le \widetilde{D}_0 t\rho^{t-1} + R_\infty s_t.$$

Multiplication by $2\gamma\lambda_{\max}(P)L_\ell/n$ proves the S_t^δ line of (B.10).

Proving the S_t^H bound (Hessian drift). The samplewise version of (A.13), Cauchy–Schwarz, and Minkowski's inequality give

$$\mathbb{E}\,\|U_k\| \le \frac{L_H}{2}(D_\Sigma\rho^k + 2R_k)a_{2,k}. \tag{B.13}$$

Moreover both matrices appearing in the third sum have spectra in $[\mu, L]$, so also $\mathbb{E}\,\|U_k\| \le \Lambda a_{1,k}$. Combining this with the bound on the norm of $A^{t-1-k}P$ proves the expression of S_t^H.

Using $a_{2,k} \le a_{2,\infty}$ and $R_k \le R_\infty$ gives

$$S_t^H \le \frac{\gamma\lambda_{\max}(P)L_H a_{2,\infty}}{2}(D_\Sigma t\rho^{t-1} + 2R_\infty s_t),$$

which is the usual explicit Hessian bound in (B.10).

Alternative S_t^H bound. From the second-moment formula and $s_k, v_k \le k$,

$$a_{2,k} \le 2\gamma\lambda_{\max}(P)G\left(\frac{k}{n} + \sqrt{\frac{1-1/n}{n}}\sqrt{k}\right).$$

For the initial-distance part of the convolution, retaining both geometric factors gives

$$\begin{aligned}\sum_{k<t} \rho^{t-1-k}\rho^k a_{2,k} &\le 2\gamma\lambda_{\max}(P)G\rho^{t-1}\left[\frac{t(t-1)}{2n} + \sqrt{\frac{1-1/n}{n}}\sum_{k<t}\sqrt{k}\right] \\ &\le 2\gamma\lambda_{\max}(P)G\rho^{t-1}\left[\frac{t^2}{2n} + \sqrt{\frac{1-1/n}{n}}t^{3/2}\right].\end{aligned} \tag{B.14}$$

For the noise part, keep the bounded envelopes instead:

$$\sum_{k<t} \rho^{t-1-k}R_k a_{2,k} \le R_\infty a_{2,\infty} s_t.$$

Combining these with $S_t^H \le (\gamma\lambda_{\max}(P)L_H/2)\sum_{k<t}\rho^{t-1-k}(D_\Sigma\rho^k + 2R_k)a_{2,k}$ proves (B.11). Thus the polynomial loss is confined to a geometrically decaying transient and never applied to persistent noise.

Proving the W_t^H bound (Hessian fluctuation). For V_k, the regularizer r's secant cancels in $H_i - H$, and the data-loss secants lie between 0 and $L_\ell I$, giving $\|H_i - H\| \le L_\ell$.

Moreover, Δ_k is $\mathcal{F}_k$-measurable and $\mathbb{E}[H_{I_k} \mid \mathcal{F}_k] = H$. Thus V_k is a martingale difference with first moment at most $L_\ell a_{1,k}$ and second moment at most $L_\ell^2 a_{2,k}^2$. The triangle inequality and (B.8) prove the expression of W_t^H.

Combining all four estimates with the bias bound proves the theorem.

For the rates, uniformly in k,

$$a_{1,k} \le \frac{2\lambda_{\max}(P)G}{\mu_P n},$$

$$a_{2,k} \le a_{2,\infty} := 2\gamma\lambda_{\max}(P)G\left[\frac{1}{n^2(1-\rho)^2} + \frac{1}{n(1-\rho^2)}\right]^{1/2} = O(n^{-1} + \sqrt{\gamma/n}).$$

Since $1-\rho = \gamma\mu_P$, $1-\rho^2 = \gamma\mu_P(2-\gamma\mu_P)$, and $R_\infty = O(\sqrt{\gamma})$, summing proves the rates for W_t^δ, W_t^H. For the two drift terms,

$$S_t^\delta \le \frac{2\gamma\lambda_{\max}(P)L_\ell}{n}\left(\widetilde{D}_0 t\rho^{t-1} + R_\infty\frac{1-\rho^t}{1-\rho}\right), \tag{B.15}$$

$$S_t^H \le \frac{\gamma\lambda_{\max}(P)L_H a_{2,\infty}}{2}\left(D_\Sigma t\rho^{t-1} + 2R_\infty\frac{1-\rho^t}{1-\rho}\right). \tag{B.16}$$

The alternative $\Lambda a_{1,k}$ entry gives the $O(1/n)$ minimum for S_t^H. For fixed γ, all first-moment entries and the uniform radius bound give $O(1/n)$ bounds uniform in t. All four large-time bounds vanish as $\gamma \downarrow 0$ for fixed n, establishing (3.7). This also follows from the convergence envelope $\mathbb{E}\,\|\Delta_t\| \le 2G/(\mu n) + D_\Sigma\rho^t + 2R_t$. □

B.4 Stochastic regularized statistical guarantees

Proof. The proof is the same as the one given in Section A.4 for the deterministic setting.

The only difference is the fact that (B.7) replaces the bound on the deterministic optimization error. □

C Proofs for deterministic projected PGD

C.1 Generalities

Work in the Hilbert norm $\|\cdot\|_{P^{-1}}$, with inner product $\langle u, v\rangle_{P^{-1}} = u^\top P^{-1} v$. The strictly convex projection problem has a unique solution. Its first-order condition is

$$q = \Pi_K^{P^{-1}}(v) \quad \Longleftrightarrow \quad \left\langle P^{-1}(v-q), y-q\right\rangle \le 0 \quad \text{for every } y \in K. \tag{C.1}$$

For $q = \Pi_K^{P^{-1}}(u)$ and $r = \Pi_K^{P^{-1}}(v)$, apply this inequality with $y = r$ and $y = q$, respectively, and add. It follows that

$$\|q-r\|_{P^{-1}}^2 \le \langle q-r, u-v\rangle_{P^{-1}}. \tag{C.2}$$

Cauchy–Schwarz proves nonexpansiveness of projection. Define $Q(v) = \Pi_K^{P^{-1}}(v) - v$. Expanding the square and using (C.2) gives:

$$\begin{aligned}\|Q(u) - Q(v)\|_{P^{-1}}^2 &= \|u-v\|_{P^{-1}}^2 + \|q-r\|_{P^{-1}}^2 - 2\langle q-r, u-v\rangle_{P^{-1}} \\ &\le \|u-v\|_{P^{-1}}^2 - \|q-r\|_{P^{-1}}^2 \le \|u-v\|_{P^{-1}}^2.\end{aligned} \tag{C.3}$$

Thus the residual map is nonexpansive.

For a differentiable convex h, $x_\star$ minimizes h on K if and only if $\langle \nabla h(x_\star), y - x_\star\rangle \ge 0$ for every $y \in K$. Taking $v = x_\star - \gamma P\nabla h(x_\star)$ and $q = x_\star$ in (C.1) proves the fixed-point property and $r_h(x_\star) = \gamma P\nabla h(x_\star)$.

Lemma C.1 (Basic estimates). *For a differentiable function h, define $T_h(x) = \Pi_K^{P^{-1}}(x - \gamma P \nabla h(x))$. Under Assumption 4.1, for $h = F, \widetilde{F}, f_i$, or $\widetilde{f}_i$, and $x, y \in K$,*

$$\|T_h(x) - T_h(y)\|_{P^{-1}} \le \rho \|x - y\|_{P^{-1}}, \tag{C.4}$$

$$\|x_\star - \widetilde{x}_\star\| \le \frac{2G}{\mu n}. \tag{C.5}$$

If $\mu I \preceq H \preceq LI$ and $A = I - \gamma PH$, then for any $q \ge 0$

$$\|A^q P\| \le \lambda_{\max}(P)\rho^q, \qquad \|A^q u\|_{P^{-1}} \le \rho^q \|u\|_{P^{-1}}, \qquad \|A^q\| \le \chi(P)\rho^q \quad (q \ge 0). \tag{C.6}$$

For a uniform index I,

$$\mathbb{E}_I \|\nabla f_I(x) - \nabla F(x)\|^2 \le G^2. \tag{C.7}$$

Proof. Existence of the empirical minimizers follows from continuity on compact K; uniqueness follows from strong convexity on the convex subset K. For any of the objectives h, the fundamental theorem of calculus gives

$$\nabla h(x) - \nabla h(y) = H_{x,y}(x - y), \qquad H_{x,y} = \int_0^1 \nabla^2 h(y + s(x - y))\, ds, \tag{C.8}$$

where $\mu I \preceq H_{x,y} \preceq LI$. Writing $V_h(x) = x - \gamma P \nabla h(x)$, we have

$$P^{-1/2}(V_h(x) - V_h(y)) = (I - \gamma P^{1/2} H_{x,y} P^{1/2}) P^{-1/2}(x - y). \tag{C.9}$$

The symmetric matrix in parentheses has eigenvalues in $[1 - \gamma L_P, 1 - \gamma \mu_P] \subset [0, \rho]$. Consequently V_h contracts by ρ in the metric norm, and composing with the nonexpansive projection proves (C.4). Combining (C.3) with (C.9) also proves

$$\|r_h(x) - r_h(y)\|_{P^{-1}} \le \rho \|x - y\|_{P^{-1}}. \tag{C.10}$$

Let $\Delta_\star = x_\star - \widetilde{x}_\star$ and $\delta_\star = \nabla f_j(\widetilde{x}_\star) - \nabla \widetilde{f}_j(\widetilde{x}_\star)$. The two minimizing variational inequalities imply

$$\left\langle \nabla F(x_\star) - \nabla \widetilde{F}(\widetilde{x}_\star), \Delta_\star \right\rangle \le 0. \tag{C.11}$$

Subtract and add $\nabla F(\widetilde{x}_\star)$, and use strong monotonicity to obtain

$$\mu \|\Delta_\star\|^2 \le -\frac{1}{n} \langle \delta_\star, \Delta_\star \rangle \le \frac{2G}{n} \|\Delta_\star\|. \tag{C.12}$$

If $\Delta_\star = 0$ the conclusion is immediate; otherwise divide by its norm. This proves (C.5) without assuming interior minimizers.

For a fixed H, put $B = I - \gamma P^{1/2} H P^{1/2}$. Then $A = P^{1/2} B P^{-1/2}$ and $A^q P = P^{1/2} B^q P^{1/2}$, where $\|B^q\| \le \rho^q$. These identities prove all three bounds in (C.6). Finally, with $g_i = \nabla f_i(x)$ and $\bar{g} = n^{-1} \sum_i g_i$,

$$\mathbb{E}_I \|g_I - \bar{g}\|^2 = \frac{1}{n} \sum_i \|g_i\|^2 - \|\bar{g}\|^2 \le G^2, \tag{C.13}$$

which proves (C.7). □

C.2 Contraction, neighboring trajectories, and residual differences

Repeated application of (C.4) to the minimizer fixed points gives

$$\|x_k - x_\star\| \le D_0 \rho^k, \qquad \|\widetilde{x}_k - \widetilde{x}_\star\| \le \widetilde{D}_0 \rho^k. \tag{C.14}$$

The empirical gradient discrepancy is $\delta(x)/n$ with norm at most $2G/n$. Adding and subtracting $T_F(\widetilde{x}_k)$ yields

$$\|\Delta_{k+1}\|_{P^{-1}} \le \rho \|\Delta_k\|_{P^{-1}} + \frac{2\gamma \sqrt{\lambda_{\max}(P)} G}{n}. \tag{C.15}$$

Here projection is nonexpansive and $\|Pu\|_{P^{-1}}^2 = u^\top Pu \le \lambda_{\max}(P)\,\|u\|^2$. Since $\Delta_0 = 0$, unrolling and multiplying by $\sqrt{\lambda_{\max}(P)}$ proves

$$\sqrt{\lambda_{\max}(P)}\,\|\Delta_k\|_{P^{-1}} \le \frac{2\lambda_{\max}(P)G}{\mu_P n}(1-\rho^k), \qquad \|\Delta_k\| \le \frac{2\lambda_{\max}(P)G}{\mu_P n}(1-\rho^k). \tag{C.16}$$

Similarly, (C.3) implies for any $x, y \in K$

$$\left\|r_F(x) - r_{\widetilde{F}}(y)\right\|_{P^{-1}} \le \rho\,\|x-y\|_{P^{-1}} + \frac{2\gamma\sqrt{\lambda_{\max}(P)}G}{n}. \tag{C.17}$$

Indeed, subtract the preprojection arguments, add and subtract $V_F(y)$, and use (C.9).

C.3 Proof of Proposition 4.3

Proof. We have $x_{k+1} = x_k - \gamma P\nabla F(x_k) + r_F(x_k)$, and the corresponding altered update. Their difference is

$$\Delta_{k+1} = A\Delta_k - \frac{\gamma}{n}P\delta_k - \gamma P\eta_k + r_F(x_k) - r_{\widetilde{F}}(\widetilde{x}_k). \tag{C.18}$$

By the fundamental theorem of calculus, $H\Delta_\star = \nabla F(x_\star) - \nabla F(\widetilde{x}_\star)$. Adding $\delta_\star/n$ and using the identities

$$r_F(x_\star) = \gamma P\nabla F(x_\star), \qquad r_{\widetilde{F}}(\widetilde{x}_\star) = \gamma P\nabla\widetilde{F}(\widetilde{x}_\star), \tag{C.19}$$

we obtain

$$\gamma P(H\Delta_\star + \delta_\star/n) = r_F(x_\star) - r_{\widetilde{F}}(\widetilde{x}_\star). \tag{C.20}$$

Subtract this from the residual difference in (C.18). The result is

$$\Delta_{k+1} = A\Delta_k + (I-A)\Delta_\star - \frac{\gamma}{n}P(\delta_k - \delta_\star) - \gamma P\eta_k + c_k. \tag{C.21}$$

Unroll from $\Delta_0 = 0$ and use $\sum_{k<t} A^{t-1-k}(I-A) = I - A^t$ to obtain (4.2). □

C.4 Bounds on the deterministic projection term

Define

$$C_t^{\Pi,\mathrm{d}} = \sum_{k<t} A^{t-1-k}c_k, \qquad s_t = \sum_{k=0}^{t-1}\rho^k, \qquad a_{1,t} = 2\gamma\lambda_{\max}(P)Gs_t/n.$$

Lemma C.2 (Deterministic projection-convolution bound)**.** *Under Assumption 4.1, without Hessian Lipschitzness, define, for $t \ge 1$,*

$$Z_t^{\Pi,\mathrm{d}} = \sum_{k<t}\rho^{t-1-k}\min\left\{D_\Sigma\rho^{k+1},\ \rho a_{1,k} + \frac{2G}{n}\left(\frac{\rho\chi(P)}{\mu} + 2\gamma\lambda_{\max}(P)\right)\right\}, \tag{C.22}$$

$$T_t^{\Pi,\mathrm{d}} = (D_\Sigma + \chi(P)\,\|\Delta_\star\|)\rho^t + 2\gamma\lambda_{\max}(P)LD_\Sigma t\rho^{t-1}. \tag{C.23}$$

Then

$$\left\|C_t^{\Pi,\mathrm{d}}\right\| \le D_t^\Pi := \min\{Z_t^{\Pi,\mathrm{d}}, T_t^{\Pi,\mathrm{d}}\}. \tag{C.24}$$

In particular,

$$D_t^\Pi \le \min\left\{D_\Sigma t\rho^t,\ \frac{2G}{n}\left[\frac{\rho\lambda_{\max}(P)}{\mu_P} + \frac{\rho\chi(P)}{\mu} + 2\gamma\lambda_{\max}(P)\right]s_t\right\}. \tag{C.25}$$

For fixed admissible $\gamma > 0$, $D_t^\Pi \to 0$.

Proof.

Step 1: proving the $Z_t^{\Pi,\mathrm{d}}$ bound. By residual contraction (C.10) on each trajectory and the metric optimization estimates,

$$\sqrt{\lambda_{\max}(P)}\,\|c_k\|_{P^{-1}} \le D_\Sigma \rho^{k+1}. \tag{C.26}$$

Alternatively, apply the paired-residual inequality (C.17) at $(x_k, \widetilde{x}_k)$ and at $(x_\star, \widetilde{x}_\star)$. The coupling estimate (C.16) and $\sqrt{\lambda_{\max}(P)}\,\|\Delta_\star\|_{P^{-1}} \le \chi(P)\,\|\Delta_\star\| \le 2\chi(P)G/(\mu n)$ give

$$\sqrt{\lambda_{\max}(P)}\,\|c_k\|_{P^{-1}} \le \rho a_{1,k} + \frac{2G}{n}\left(\frac{\rho\chi(P)}{\mu} + 2\gamma\lambda_{\max}(P)\right). \tag{C.27}$$

The metric kernel bound (C.6) yields

$$\left\|C_t^{\Pi,\mathrm{d}}\right\| \le \sum_{k<t} \rho^{t-1-k}\sqrt{\lambda_{\max}(P)}\,\|c_k\|_{P^{-1}}\,.$$

Taking the smaller of (C.26) and (C.27) at each k proves $\left\|C_t^{\Pi,\mathrm{d}}\right\| \le Z_t^{\Pi,\mathrm{d}}$.

Step 2: proving the $T_t^{\Pi,\mathrm{d}}$ bound. Set $e_k = \Delta_k - \Delta_\star$, $g_k = \nabla F(x_k) - \nabla\widetilde{F}(\widetilde{x}_k)$, $g_\star = \nabla F(x_\star) - \nabla\widetilde{F}(\widetilde{x}_\star)$, and $Q_k = g_k - g_\star - He_k$. The difference of the updates and the fixed-point balance give

$$c_k = \Delta_{k+1} - \Delta_k + \gamma P(g_k - g_\star) = e_{k+1} - Ae_k + \gamma PQ_k.$$

Multiplying by A^{t-1-k} and summing telescopes. Since $e_0 = -\Delta_\star$,

$$C_t^{\Pi,\mathrm{d}} = e_t + A^t\Delta_\star + \gamma\sum_{k<t} A^{t-1-k}PQ_k.$$

Contraction gives $\|e_k\| \le D_\Sigma\rho^k$, while smoothness gives

$$\|Q_k\| \le 2L(\|x_k - x_\star\| + \|\widetilde{x}_k - \widetilde{x}_\star\|) \le 2LD_\Sigma\rho^k.$$

Using (C.6) on all three terms and $\sum_{k<t}\rho^{t-1-k}\rho^k = t\rho^{t-1}$ proves the $T_t^{\Pi,\mathrm{d}}$ bound.

Step 3: combining the candidates and taking the limit. We have proved (C.24). Bounding $a_{1,k}$ by $2\lambda_{\max}(P)G/(\mu_P n)$ in Step 1 proves the second entry of (C.25); summing the transient entry proves the first.

For $0 < \rho < 1$, both ρ^t and $t\rho^{t-1}$ tend to zero. When $\rho = 0$, the displayed transient is zero for $t \ge 2$. This proves the limit. □

Decaying alternative without inverse μ_P. The first candidate directly gives

$$D_t^\Pi \le \sum_{k<t}\rho^{t-1-k}D_\Sigma\rho^{k+1} = D_\Sigma t\rho^t. \tag{C.28}$$

This removes inverse μ_P coefficients while preserving decay.

C.5 Proof of Theorem 4.4

For $t \ge 1$, we use the exact norm of the minimizer response and the following drift bounds:

$$\|\mathcal{B}_t\| = \left\|(I - A^t)\Delta_\star\right\|, \tag{C.29}$$

$$D_t^\delta = \frac{2\gamma\lambda_{\max}(P)L\widetilde{D}_0}{n}t\rho^{t-1}, \tag{C.30}$$

$$D_t^H = \frac{\gamma\lambda_{\max}(P)^2L_HGD_\Sigma}{\mu_P n}\sum_{k<t}\rho^{t-1-k}\rho^k(1-\rho^k) \le \frac{\gamma\lambda_{\max}(P)^2L_HGD_\Sigma}{\mu_P n}t\rho^{t-1}. \tag{C.31}$$

Without Assumption 4.2, replace D_t^H by $\gamma\lambda_{\max}(P)\sum_{k<t}\rho^{t-1-k}\min\{2LD_\Sigma\rho^k, (L - \mu)\frac{2\lambda_{\max}(P)G}{\mu_P n}(1-\rho^k)\}$.

Then Theorem 4.4 states that

$$\|\Delta_t\| \le \|\mathcal{B}_t\| + D_t^\delta + D_t^H + D_t^\Pi. \tag{C.32}$$

The three drift terms bound the norms of the corresponding convolutions in (4.2), and the projection coefficient D_t^Π is defined in (C.24).

Proof. Let us bound each term from Proposition 4.3.

Proving the D_t^δ bound (replacement drift). Smoothness of both replaced losses gives

$$\|\delta_k - \delta_\star\| \le 2L \|\widetilde{x}_k - \widetilde{x}_\star\| \le 2L\widetilde{D}_0\rho^k. \tag{C.33}$$

Convolving with $\gamma A^{t-1-k}P/n$ and using Theorem C.1 upper-bounds the replacement drift by D_t^δ.

Proving the D_t^H bound (Hessian drift). For the Hessian term, let $\bar{H}_k = \int_0^1 \nabla^2 F(\widetilde{x}_k + s\Delta_k)\, ds$. Then

$$\left\|\bar{H}_k - H\right\| \le \frac{L_H}{2}(\|x_k - x_\star\| + \|\widetilde{x}_k - \widetilde{x}_\star\|). \tag{C.34}$$

Since $\eta_k = (\bar{H}_k - H)\Delta_k$, equations (C.14) and (C.16) yield

$$\|\eta_k\| \le \frac{L_H}{2}(\|x_k - x_\star\| + \|\widetilde{x}_k - \widetilde{x}_\star\|)\,\|\Delta_k\| \tag{C.35}$$

$$\le \frac{\lambda_{\max}(P)L_H G D_\Sigma}{\mu_P n}\rho^k(1-\rho^k). \tag{C.36}$$

Let $a_{1,k} = 2\gamma\lambda_{\max}(P)Gs_k/n$. Keeping the geometric factors before choosing an envelope gives

$$\left\|\gamma\sum_{k<t} A^{t-1-k}P\eta_k\right\| \le \frac{\gamma\lambda_{\max}(P)L_H D_\Sigma}{2}\sum_{k<t}\rho^{t-1-k}\rho^k a_{1,k} \tag{C.37}$$

$$= \frac{\gamma\lambda_{\max}(P)L_H D_\Sigma}{2}\rho^{t-1}\sum_{k<t} a_{1,k}.$$

Using $a_{1,k} \le 2\lambda_{\max}(P)G/(\mu_P n)$ gives the usual $t\rho^{t-1}/\mu_P$ estimate in (C.31). Using instead $a_{1,k} \le 2\gamma\lambda_{\max}(P)Gk/n$ gives

$$\left\|\gamma\sum_{k<t} A^{t-1-k}P\eta_k\right\| \le \frac{\gamma^2\lambda_{\max}(P)^2 L_H G D_\Sigma}{2n}t(t-1)\rho^{t-1} \tag{C.38}$$

$$\le \frac{\gamma^2\lambda_{\max}(P)^2 L_H G D_\Sigma}{2n}t^2\rho^{t-1}.$$

This proves the alternative Hessian contribution in (4.3), exactly as in the regularized deterministic proof.

If no Hessian Lipschitz bound is assumed, $\mu I \preceq \bar{H}_k, H \preceq LI$, so $\left\|\bar{H}_k - H\right\| \le L - \mu$. Also $\eta_k \to 0$ has the explicit estimate

$$\|\eta_k\| \le \|\nabla F(x_k) - \nabla F(x_\star)\| + \|\nabla F(\widetilde{x}_k) - \nabla F(\widetilde{x}_\star)\| + \|H(\Delta_k - \Delta_\star)\| \tag{C.39}$$

$$\le 2LD_\Sigma\rho^k. \tag{C.40}$$

This proves the stated alternative convolution bound.

Proving the D_t^Π bound (centered projection drift). The definition and proof of D_t^Π are collected in Appendix C.4. In particular, its transient candidate is $D_t^\Pi \le \sum_{k<t}\rho^{t-1-k}D_\Sigma\rho^{k+1} = D_\Sigma t\rho^t$. Combining this with the alternative Hessian bound proves (4.4). Together with the two gradient-drift estimates above, this gives the remainder bound.

The triangle inequality in (4.2) now proves (C.32).

For fixed γ, $\frac{1-\rho^t}{1-\rho} \le (1-\rho)^{-1}$ and $t\rho^{t-1} \le (1-\rho)^{-2}$. Also $D_\Sigma \le 2\chi(P)\operatorname{diam}(K)$ and $\frac{2\lambda_{\max}(P)G}{\mu_P n}$, $\frac{2G}{\mu n}$ are constant multiples of $1/n$. Equations (C.30), (C.31), and the second entry of (C.25) prove the uniform $1/n$ bounds. The same conclusion holds for the Hessian alternative using its $\Lambda\frac{2\lambda_{\max}(P)G}{\mu_P n}(1-\rho^k)$ entry. The sequences $t\rho^{t-1}$ and $t\rho^t$ tend to zero. The geometric entry in the alternative Hessian bound also gives a multiple of $t\rho^{t-1}$. Thus all drifts vanish; $A^t\Delta_\star \to 0$ by (C.6).

It also follows directly from convergence of both trajectories to their constrained minimizers. The triangle inequality applied only to the three drift convolutions proves

$$\|\Delta_t - \mathcal{B}_t\| \le D_t^\delta + D_t^H + D_t^\Pi. \tag{C.41}$$

□

C.6 Deterministic projected generalization and excess risk

This subsection proves the deterministic part of Theorem 5.1.

Proof. Similarly to before, under our assumptions all expected values are finite.

Use $S^{(i)}$ and ω as defined in Section 5. We have

$$|\mathbb{E}[\mathcal{R}(x_t) - F_S(x_t)]| = \left|\frac{1}{n}\sum_i \mathbb{E}[\ell(x_t(S^{(i)}), z_i) - \ell(x_t(S), z_i)]\right| \le G\beta_t^{\mathrm{d}}.$$

Let $x^+ = T_F(x)$ and $x_\star$ minimize F on K. Projection optimality (C.1) yields

$$\left\langle \nabla F(x), x^+ - x_\star \right\rangle \le \frac{1}{\gamma}\left\langle x - x^+, x^+ - x_\star \right\rangle_{P^{-1}}. \tag{C.42}$$

Convexity at x and smoothness between x and x^+ give

$$\begin{aligned} F(x^+) - F(x_\star) &\le \left\langle \nabla F(x), x^+ - x_\star \right\rangle + \frac{L}{2}\left\|x^+ - x\right\|^2 && \text{(C.43)}\\ &\le \frac{1}{2\gamma}(\|x - x_\star\|_{P^{-1}}^2 - \left\|x^+ - x_\star\right\|_{P^{-1}}^2 - \left\|x^+ - x\right\|_{P^{-1}}^2) \\ &\quad + \frac{L\lambda_{\max}(P)}{2}\left\|x^+ - x\right\|_{P^{-1}}^2 && \text{(C.44)}\\ &\le \frac{1}{2\gamma}(\|x - x_\star\|_{P^{-1}}^2 - \left\|x^+ - x_\star\right\|_{P^{-1}}^2), && \text{(C.45)} \end{aligned}$$

where the last step uses $\gamma L\lambda_{\max}(P) \le 1$. For $x = x_{t-1}$, contraction implies

$$F(x_t) - F(x_\star) \le \frac{1}{2\gamma}\rho^{2(t-1)}\left\|x_0 - x_\star\right\|_{P^{-1}}^2. \tag{C.46}$$

Independently, Lipschitzness and deterministic distance convergence give $F(x_t) - F(x_\star) \le GD_0\rho^t$.

For any output in K, decompose

$$\begin{aligned} \mathbb{E}\mathcal{R}(x_t) - \mathcal{R}(x_{\mathrm{opt}}) &= \mathbb{E}[\mathcal{R}(x_t) - F_S(x_t)] + \mathbb{E}[F_S(x_t) - F_S(x_\star)] \\ &\quad + \underbrace{\mathbb{E}[F_S(x_\star) - F_S(x_{\mathrm{opt}})]}_{\le 0} + \underbrace{\mathbb{E}[F_S(x_{\mathrm{opt}}) - \mathcal{R}(x_{\mathrm{opt}})]}_{=0}. \end{aligned} \tag{C.47}$$

Bound the first term by $G\beta_t^{\mathrm{d}}$ and the second by the smaller of the two deterministic optimization estimates above. Use $D_0 \le \chi(P)\operatorname{diam}(K)$ and $\|x_0 - x_\star\|_{P^{-1}}^2 \le \operatorname{diam}(K)^2/\lambda_{\min}(P)$.

This proves the upper bound. Letting $t \to \infty$ yields $2G^2/(\mu n)$. □

D Proofs for stochastic projected PGD

D.1 General bounds

Use s_t, v_t from (B.1) and set $a_{1,t} = 2\gamma\lambda_{\max}(P)Gs_t/n$. The second-moment and optimization-noise scales are

$$a_{2,t} = 2\gamma\lambda_{\max}(P)G\sqrt{\frac{s_t^2}{n^2} + \frac{1 - 1/n}{n}v_t}, \qquad R_t = \gamma\lambda_{\max}(P)G\sqrt{v_t}. \tag{D.1}$$

Both vanish at $t = 0$, and

$$R_t \le R_\infty = \frac{\lambda_{\max}(P)G\sqrt{\gamma}}{\sqrt{\mu_P(2 - \gamma\mu_P)}}, \qquad \rho = 1 - \gamma\mu_P. \tag{D.2}$$

Use s_t, v_t from (B.1) and define the stationary envelopes

$$R_\infty = \frac{\gamma\lambda_{\max}(P)G}{\sqrt{1 - \rho^2}}, \qquad a_{2,\infty} = 2\gamma\lambda_{\max}(P)G\sqrt{\frac{1}{n^2(1-\rho)^2} + \frac{1}{n(1-\rho^2)}}. \tag{D.3}$$

Theorem D.1. *Under Assumption 4.1, without Hessian Lipschitzness,*

$$\|\Delta_t\| \le 2\gamma\lambda_{\max}(P)G\sum_{k<t}\rho^{t-1-k}J_k \quad \textit{on every index path}, \tag{D.4}$$

$$\mathbb{E}\,\|\Delta_t\| \le \frac{2\lambda_{\max}(P)G}{\mu_P n}(1-\rho^t), \qquad (\mathbb{E}\,\|\Delta_t\|^2)^{1/2} \le a_{2,t}, \tag{D.5}$$

$$\mathbb{E}\,\|x_t - x_\star\|_{P^{-1}}^2 \le \rho^{2t}\,\|x_0 - x_\star\|_{P^{-1}}^2 + \gamma^2\lambda_{\max}(P)G^2\frac{1-\rho^{2t}}{1-\rho^2}. \tag{D.6}$$

The analogous optimization bound holds on the altered dataset, and $\mathbb{E}\,\|x_t - x_\star\| \le D_0\rho^t + R_t$. *Consequently,*

$$\mathbb{E}\,\|\Delta_t\| \le \min\left\{\frac{2\lambda_{\max}(P)G}{\mu_P n}(1-\rho^t),\ \frac{2G}{\mu n} + D_\Sigma\rho^t + 2R_t\right\}. \tag{D.7}$$

$$\limsup_{t\to\infty}\mathbb{E}\,\|\Delta_t\| \le \min\left\{\frac{2\lambda_{\max}(P)G}{\mu_P n},\ \frac{2G}{\mu n} + 2R_\infty\right\}, \qquad \limsup_{\gamma\downarrow 0}\limsup_{t\to\infty}\mathbb{E}\,\|\Delta_t\| \le \frac{2G}{\mu n}. \tag{D.8}$$

The coupling bound is $O(1/n)$ *uniformly in time and step size; the optimization comparison gives the preconditioner-independent ERM limit with an* $O(\sqrt{\gamma})$ *correction.*

Proof. We prove coupling, then optimization.

Coupling. Contraction and the replacement perturbation give $\|\Delta_{k+1}\|_{P^{-1}} \le \rho\,\|\Delta_k\|_{P^{-1}} + 2\gamma\sqrt{\lambda_{\max}(P)}GJ_k$. Unroll from zero:

$$\sqrt{\lambda_{\max}(P)}\,\|\Delta_t\|_{P^{-1}} \le 2\gamma\lambda_{\max}(P)G\sum_{k<t}\rho^{t-1-k}J_k. \tag{D.9}$$

Since

$$\|\Delta_t\| \le \sqrt{\lambda_{\max}(P)}\,\|\Delta_t\|_{P^{-1}}\,, \tag{D.10}$$

this proves (D.4). For $Z_t = \sum_{k<t}\rho^{t-1-k}J_k$, independence gives

$$\mathbb{E}Z_t = s_t/n, \qquad \mathbb{E}Z_t^2 = s_t^2/n^2 + (1-1/n)v_t/n.$$

These prove (D.5), including its stronger metric form $\sqrt{\lambda_{\max}(P)}\mathbb{E}\,\|\Delta_t\|_{P^{-1}} \le a_{1,t}$.

Optimization. Let $\xi_k = \nabla f_{I_k}(x_k) - \nabla F(x_k)$, conditionally centered with second moment at most G^2. Nonexpansiveness and the minimizer fixed point give

$$\|x_{k+1} - x_\star\|_{P^{-1}}^2 \le \|V_F(x_k) - V_F(x_\star) - \gamma P\xi_k\|_{P^{-1}}^2\,.$$

The conditional cross term is zero. By (C.9),

$$\mathbb{E}[\|x_{k+1} - x_\star\|_{P^{-1}}^2 \mid \mathcal{F}_k] \le \rho^2\,\|x_k - x_\star\|_{P^{-1}}^2 + \gamma^2\lambda_{\max}(P)G^2.$$

Iteration proves (D.6); conversion to Euclidean norm gives

$$(\mathbb{E}\,\|x_k - x_\star\|^2)^{1/2} \le \sqrt{D_0^2\rho^{2k} + R_k^2} \le D_0\rho^k + R_k. \tag{D.11}$$

Apply the same argument to the altered dataset. The triangle inequality through the minimizers proves (D.7); $\rho^t \to 0$ and $R_\infty = O(\sqrt{\gamma})$ prove (D.8). □

D.2 GRADIENT-CONVOLUTION COEFFICIENTS

Put $\Lambda = L - \mu$ and $d_\star = \|\delta_\star\| \le 2G$. Define

$$W_t^\delta = \gamma\lambda_{\max}(P) d_\star \min\left\{\frac{2(1-1/n)}{n} s_t, \sqrt{\frac{1-1/n}{n}} v_t\right\}, \tag{D.12}$$

$$S_t^\delta = \frac{2\gamma\lambda_{\max}(P)L}{n} \sum_{k<t} \rho^{t-1-k}(\widetilde{D}_0\rho^k + R_k), \tag{D.13}$$

$$S_t^H = \gamma\lambda_{\max}(P) \sum_{k<t} \rho^{t-1-k} \min\left\{\frac{L_H}{2}(D_\Sigma\rho^k + 2R_k)a_{2,k}, \Lambda a_{1,k}\right\}, \tag{D.14}$$

$$W_t^H = \gamma\lambda_{\max}(P)\Lambda \min\left\{\sum_{k<t} \rho^{t-1-k} a_{1,k}, \left(\sum_{k<t} \rho^{2(t-1-k)} a_{2,k}^2\right)^{1/2}\right\}. \tag{D.15}$$

With Lipschitz Hessians and $t \ge 1$, an explicit componentwise version is

$$\begin{aligned}
W_t^\delta &\le 2\gamma\lambda_{\max}(P)G \min\left\{\frac{2(1-1/n)}{n} s_t, \sqrt{\frac{1-1/n}{n}} v_t\right\}, \\
S_t^\delta &\le \frac{2\gamma\lambda_{\max}(P)L}{n}\left(\widetilde{D}_0 t\rho^{t-1} + R_\infty s_t\right), \\
S_t^H &\le \frac{\gamma\lambda_{\max}(P)L_H a_{2,\infty}}{2}\left(D_\Sigma t\rho^{t-1} + 2R_\infty s_t\right), \\
W_t^H &\le \gamma\lambda_{\max}(P)(L-\mu)\min\left\{\frac{2\lambda_{\max}(P)G}{\mu_P n} s_t, a_{2,\infty}\sqrt{v_t}\right\}.
\end{aligned} \tag{D.16}$$

Alternatively, inverse μ_P factors can be removed from the whole transient Hessian contribution:

$$\begin{aligned}
S_t^H \le{}& \gamma^2\lambda_{\max}(P)^2 L_H G D_\Sigma \left(\frac{t^2}{2n} + \sqrt{\frac{1-1/n}{n}}\, t^{3/2}\right)\rho^{t-1} \\
&+ \gamma\lambda_{\max}(P) L_H R_\infty a_{2,\infty} s_t.
\end{aligned} \tag{D.17}$$

Without Assumption 4.2, retain only $\Lambda a_{1,k}$ in (D.14). With Lipschitz Hessians, the large-time orders are

$$W_t^\delta : O\Big(\min\{n^{-1}, \sqrt{\gamma/n}\}\Big), \qquad S_t^\delta : O(\sqrt{\gamma}/n), \tag{D.18}$$

$$S_t^H : O\big(\min\{n^{-1}, \sqrt{\gamma}/n + \gamma/\sqrt{n}\}\big), \qquad W_t^H : O\big(\min\{n^{-1}, \sqrt{\gamma}/n + \gamma/\sqrt{n}\}\big). \tag{D.19}$$

These are limsup bounds as $t \to \infty$, uniform in n and admissible γ; projection rates are in Theorem D.2.

D.3 MARTINGALE ESTIMATE AND EXACT DECOMPOSITION

For square-integrable, $\mathcal{F}_{k+1}$-measurable M_k with $\mathbb{E}[M_k \mid \mathcal{F}_k] = 0$, and deterministic A, P satisfying (C.6),

$$\mathbb{E}\left\|\gamma \sum_{k<t} A^{t-1-k} P M_k\right\| \le \gamma\lambda_{\max}(P)\left(\sum_{k<t} \rho^{2(t-1-k)} \mathbb{E}\,\|M_k\|^2\right)^{1/2}. \tag{D.20}$$

Indeed, Jensen's inequality reduces the claim to a second moment. For $r < k$, conditional centering gives $\mathbb{E}\left\langle A^{t-1-r} P M_r, A^{t-1-k} P M_k\right\rangle = 0$. Only diagonal terms remain, each bounded using $\|A^q P\| \le \lambda_{\max}(P)\rho^q$. Matrices deterministic conditional on the datasets suffice.

D.4 Proof of Proposition 4.5

Proof. Subtracting updates gives

$$\Delta_{k+1} = A\Delta_k - \gamma P(U_k + V_k) - \gamma P J_k \delta_k + r_{f_{I_k}}(x_k) - r_{\widetilde{f}_{I_k}}(\widetilde{x}_k).$$

Expand $J_k\delta_k = \delta_\star/n + \zeta_k\delta_\star + J_k(\delta_k - \delta_\star)$. Fixed-point balance (C.20) replaces the constant terms by $(I - A)\Delta_\star$ and the remaining residual by c_k^{s}. Unrolling from zero proves (4.6). Since Δ_k is $\mathcal{F}_k$-measurable and $\mathbb{E}[H_{I_k} \mid \mathcal{F}_k] = H$, both V_k and $\zeta_k\delta_\star$ are martingale differences. □

D.5 Stochastic projection term

Under Assumption 4.1, no Hessian Lipschitzness is needed. Set

$$\Lambda = L - \mu, \quad s_t = \sum_{k<t} \rho^{t-1-k} = \frac{1-\rho^t}{1-\rho}, \quad v_t = \sum_{k<t} \rho^{2(t-1-k)} = \frac{1-\rho^{2t}}{1-\rho^2}. \tag{D.21}$$

Both sums are zero at $t = 0$. Write

$$C_t^{\Pi} = \sum_{k<t} A^{t-1-k} c_k^{\mathrm{s}}, \qquad d_\star^{\Pi} = r_F(x_\star) - r_{\widetilde{F}}(\widetilde{x}_\star), \qquad a_{1,t} = \frac{2\lambda_{\max}(P)G}{\mu_P n}(1-\rho^t). \tag{D.22}$$

We bound the signed convolution C_t^{Π}, retaining cancellation between corrections.

Triangle bound.

$$Z_t^{\Pi} = \sum_{k<t} \rho^{t-1-k} \min \left\{ \begin{array}{c} \frac{2G}{n}\left[\frac{\rho\lambda_{\max}(P)}{\mu_P}(1-\rho^k) + \frac{\rho\chi(P)}{\mu} + 2\gamma\lambda_{\max}(P)\right], \\ \rho(D_\Sigma\rho^k + 2R_k) + 2\gamma\lambda_{\max}(P)G \end{array} \right\}. \tag{D.23}$$

Explicit bounds. The remaining candidates are

$$\begin{aligned} E_t^{\Pi} &= 4\gamma\lambda_{\max}(P)G s_t, \\ N_t^{\Pi} &= \left(3 + \frac{L}{\mu}\right) a_{1,t} + \gamma\lambda_{\max}(P)\Lambda \sum_{k<t} \rho^{t-1-k} a_{1,k}, \\ \nu_\star &= \Lambda \left\|\Delta_\star\right\| + \left\|\delta_\star\right\| \sqrt{\frac{1}{n}\left(1 - \frac{1}{n}\right)}, \\ T_t^{\Pi} &= (D_\Sigma + \chi(P)\left\|\Delta_\star\right\|)\rho^t + 2R_t \\ &\quad + 2\gamma\lambda_{\max}(P)L \sum_{k<t} \rho^{t-1-k}(D_\Sigma\rho^k + 2R_k) + \gamma\lambda_{\max}(P)\nu_\star\sqrt{v_t}. \end{aligned} \tag{D.24}$$

Use their minimum throughout:

$$S_t^{\Pi} = \min\{Z_t^{\Pi}, E_t^{\Pi}, N_t^{\Pi}, T_t^{\Pi}\}, \qquad \mathbb{E}\left\|C_t^{\Pi}\right\| \le S_t^{\Pi}. \tag{D.25}$$

Here $S_0^{\Pi} = 0$, although T_0^{Π} may be positive. Both Z_t^{Π} and E_t^{Π} also bound the sum of expected norms.

Theorem D.2 (Projection-convolution bound)**.** *Under Assumption 4.1, without Hessian Lipschitzness, the vector $C_t^{\Pi} = \sum_{k<t} A^{t-1-k} c_k^{\mathrm{s}}$ satisfies*

$$\mathbb{E}\left\|C_t^{\Pi}\right\| \le S_t^{\Pi}, \qquad \sup_{t\ge 0,\ 0<\gamma\le L_P^{-1}} S_t^{\Pi} = O(n^{-1}), \qquad \limsup_{t\to\infty} S_t^{\Pi} = O(\min\{n^{-1}, \sqrt{\gamma}\}).$$

The constants depend on the problem parameters, but not on n or admissible γ.

Proof. We prove the four candidates, then their rates.

Proving Z_t^{Π}. Use the paired-residual argument (C.17) samplewise at the current pair, with perturbation $2\gamma\sqrt{\lambda_{\max}(P)GJ_k}$, and empirically at the minimizing pair. The metric coupling bound, $\mathbb{E}J_k = 1/n$, and $\sqrt{\lambda_{\max}(P)}\,\|\Delta_\star\|_{P^{-1}} \le 2\chi(P)G/(\mu n)$ give

$$\sqrt{\lambda_{\max}(P)}\mathbb{E}\,\|c_k^{\mathrm{s}}\|_{P^{-1}} \le \rho\left(\frac{2\lambda_{\max}(P)G}{\mu_P n}(1-\rho^k) + \chi(P)\frac{2G}{\mu n}\right) + 4\gamma\lambda_{\max}(P)G/n. \tag{D.26}$$

Alternatively, residual contraction and nonexpansiveness give

$$\|r_{f_i}(x_k) - r_F(x_\star)\|_{P^{-1}} \le \rho\,\|x_k - x_\star\|_{P^{-1}} + \gamma\sqrt{\lambda_{\max}(P)}\,\|\nabla f_i(x_\star) - \nabla F(x_\star)\|\,.$$

The last norm has expectation at most G by (C.7). Apply this on both datasets and use (D.6):

$$\sqrt{\lambda_{\max}(P)}\mathbb{E}\,\|c_k^{\mathrm{s}}\|_{P^{-1}} \le \rho(D_\Sigma\rho^k + 2R_k) + 2\gamma\lambda_{\max}(P)G. \tag{D.27}$$

Convolving the pointwise minimum of these two bounds with the metric kernel proves (D.23).

Proving E_t^{Π}. For feasible x, $Q(x) = 0$, where $Q = \Pi_K^{P^{-1}} - \mathrm{Id}$. By (C.3),

$$\|r_h(x)\|_{P^{-1}} = \|Q(x - \gamma P\nabla h(x)) - Q(x)\|_{P^{-1}} \le \gamma\,\|P\nabla h(x)\|_{P^{-1}} \le \gamma\sqrt{\lambda_{\max}(P)}G. \tag{D.28}$$

The four residuals in c_k^{s} therefore give $\sqrt{\lambda_{\max}(P)}\,\|c_k^{\mathrm{s}}\|_{P^{-1}} \le 4\gamma\lambda_{\max}(P)G$. Hence

$$\mathbb{E}\left\|C_t^{\Pi}\right\| \le \sqrt{\lambda_{\max}(P)}\sum_{k<t}\rho^{t-1-k}\mathbb{E}\,\|c_k^{\mathrm{s}}\|_{P^{-1}} \le 4\gamma\lambda_{\max}(P)Gs_t = \frac{4\lambda_{\max}(P)G}{\mu_P}(1-\rho^t). \tag{D.29}$$

Proving N_t^{Π}. Put $g_k = \nabla f_{I_k}(x_k) - \nabla\widetilde{f}_{I_k}(\widetilde{x}_k) = \bar{H}_{I_k,k}\Delta_k + J_k\delta_k$. Subtracting the updates gives

$$c_k^{\mathrm{s}} = \Delta_{k+1} - \Delta_k + \gamma P g_k - d_\star^{\Pi}. \tag{D.30}$$

Use $I - A = \gamma PH$ and telescope from $\Delta_0 = 0$:

$$\begin{aligned} C_t^{\Pi} &= \Delta_t + \gamma\sum_{k<t}A^{t-1-k}P(\bar{H}_{I_k,k} - H)\Delta_k \\ &\quad + \gamma\sum_{k<t}A^{t-1-k}PJ_k\delta_k - \sum_{k<t}A^{t-1-k}d_\star^{\Pi}. \end{aligned} \tag{D.31}$$

Here $\left\|\bar{H}_{I_k,k} - H\right\| \le \Lambda$, $\mathbb{E}\,\|\Delta_k\| \le a_{1,k}$, and $\|\delta_k\| \le 2G$. Fixed-point balance gives $d_\star^{\Pi} = \gamma P(H\Delta_\star + \delta_\star/n)$, with

$$\|H\Delta_\star + \delta_\star/n\| \le L\,\|\Delta_\star\| + \frac{2G}{n} \le \frac{2G}{n}\left(1 + \frac{L}{\mu}\right). \tag{D.32}$$

Thus the four terms in (D.31) are bounded by $a_{1,t}$, $\gamma\lambda_{\max}(P)\Lambda\sum_{k<t}\rho^{t-1-k}a_{1,k}$, $a_{1,t}$, and $(1 + L/\mu)a_{1,t}$: their sum is N_t^{Π}. The coupling bound used here is independent of the projection-convolution estimate. Since $a_{1,k} \le 2\lambda_{\max}(P)G/(\mu_P n)$,

$$N_t^{\Pi} \le \frac{2\lambda_{\max}(P)G}{\mu_P n}\left(3 + \frac{L}{\mu} + \frac{\lambda_{\max}(P)(L-\mu)}{\mu_P}\right)(1-\rho^t). \tag{D.33}$$

The constant is uniform in t, n, γ. Equivalently, for $t \ge 1$,

$$\sum_{k<t}\rho^{t-1-k}a_{1,k} = \frac{2\lambda_{\max}(P)G}{\mu_P n}(s_t - t\rho^{t-1}),$$

with the usual $0^0 = 1$ convention.

Proving T_t^{Π}. Set $e_k = \Delta_k - \Delta_\star$ and define

$$g_i^\star = \nabla f_i(x_\star) - \nabla \widetilde{f}_i(\widetilde{x}_\star), \tag{D.34}$$
$$\bar{g}^\star = \frac{1}{n}\sum_i g_i^\star = \nabla F(x_\star) - \nabla \widetilde{F}(\widetilde{x}_\star), \qquad \xi_k^\star = g_{I_k}^\star - \bar{g}^\star.$$

Then $d_\star^{\Pi} = \gamma P \bar{g}^\star$ and $\xi_k^\star$ is a martingale difference. With

$$Q_k = g_k - g_{I_k}^\star - He_k, \tag{D.35}$$

substitution into (D.30) cancels the fixed mean:

$$c_k^{\mathrm{s}} = e_{k+1} - Ae_k + \gamma P Q_k + \gamma P \xi_k^\star. \tag{D.36}$$

Telescoping, using $e_0 = -\Delta_\star$, yields

$$C_t^{\Pi} = e_t + A^t \Delta_\star + \gamma \sum_{k<t} A^{t-1-k} P Q_k + \gamma \sum_{k<t} A^{t-1-k} P \xi_k^\star. \tag{D.37}$$

The optimization and kernel estimates imply

$$\mathbb{E}\,\|e_k\| \le D_\Sigma \rho^k + 2R_k, \qquad \left\|A^t \Delta_\star\right\| \le \chi(P)\rho^t \left\|\Delta_\star\right\|. \tag{D.38}$$

Samplewise smoothness gives $\|Q_k\| \le 2L(\|x_k - x_\star\| + \|\widetilde{x}_k - \widetilde{x}_\star\|)$, so

$$\mathbb{E}\,\|Q_k\| \le 2L(D_\Sigma \rho^k + 2R_k). \tag{D.39}$$

Finally $g_i^\star = H_i \Delta_\star + \mathbf{1}_{\{i=j\}}\delta_\star$, hence

$$\xi_k^\star = (H_{I_k} - H)\Delta_\star + (J_k - 1/n)\delta_\star. \tag{D.40}$$

Minkowski's inequality (not independence of these two terms) gives

$$(\mathbb{E}\,\|\xi_k^\star\|^2)^{1/2} \le \Lambda \left\|\Delta_\star\right\| + \left\|\delta_\star\right\| \sqrt{\frac{1}{n}\left(1 - \frac{1}{n}\right)} = \nu_\star. \tag{D.41}$$

By (D.20), the last convolution in (D.37) is at most $\gamma \lambda_{\max}(P)\nu_\star \sqrt{v_t}$ in expected norm. The four estimates prove T_t^{Π}, and taking the minimum proves (D.25).

Explicit transient and rates. Use $R_k \le R_\infty$ and $\sum_{k<t} \rho^{t-1-k}\rho^k = t\rho^{t-1}$:

$$\begin{aligned} T_t^{\Pi} &\le (D_\Sigma + \chi(P)\left\|\Delta_\star\right\|)\rho^t + 2\gamma\lambda_{\max}(P) L D_\Sigma t \rho^{t-1} \\ &\quad + 2R_t + 4\gamma\lambda_{\max}(P) L R_\infty s_t + \gamma \lambda_{\max}(P)\nu_\star \sqrt{v_t}. \end{aligned} \tag{D.42}$$

To derive the simpler main-text formula (4.8), use

$$R_t \le \lambda_{\max}(P) G \sqrt{\frac{\gamma}{\mu_P}}\sqrt{1-\rho^{2t}}, \quad R_\infty \le \lambda_{\max}(P) G \sqrt{\frac{\gamma}{\mu_P}}, \quad \gamma s_t = \frac{1-\rho^t}{\mu_P}.$$

Also $\gamma\sqrt{v_t} \le \sqrt{\gamma/\mu_P}\sqrt{1-\rho^{2t}}$, $1-\rho^t \le \sqrt{1-\rho^{2t}}$, and $\nu_\star \le (L-\mu)\left\|\Delta_\star\right\| + \left\|\delta_\star\right\|/\sqrt{n}$. Substitution in (D.42) proves the displayed main-text bound. Only the decaying transient avoids inverse-curvature coefficients; the persistent noise retains them.

The transient vanishes for fixed γ, giving

$$\begin{aligned} \limsup_{t\to\infty} S_t^{\Pi} \le \limsup_{t\to\infty} T_t^{\Pi} &\le \left(2 + \frac{4\lambda_{\max}(P)L}{\mu_P}\right) R_\infty + \frac{\gamma\lambda_{\max}(P)\nu_\star}{\sqrt{1-\rho^2}} \\ &= \frac{\lambda_{\max}(P)\sqrt{\gamma}}{\sqrt{\mu_P(2-\gamma\mu_P)}}\left[\left(2 + \frac{4\lambda_{\max}(P)L}{\mu_P}\right) G + \nu_\star\right]. \end{aligned} \tag{D.43}$$

Since $\gamma\mu_P \le 1$ and $\nu_\star \le 2G(L-\mu)/(\mu n) + 2G\sqrt{(1-1/n)/n}$, this is $O(\sqrt{\gamma})$ uniformly in n. Combining with (D.33) proves

$$\limsup_{t\to\infty} \mathbb{E}\left\|C_t^{\Pi}\right\| \le \limsup_{t\to\infty} S_t^{\Pi} = O\left(\min\left\{\frac{1}{n}, \sqrt{\gamma}\right\}\right). \tag{D.44}$$

The iterated large-time/small-step limit is zero; no mean-zero projection residual or fixed-step convergence to the ERM is asserted. □

Residual size is not residual Lipschitzness. For $K = [0, 1]$, $P = 1$, $h(x) = x^2/2 + x$, $0 < \gamma < 1$, $x = 0$, and $y = \gamma/2$, both preprojection points are negative. Thus

$$r_h(x) = \gamma, \qquad r_h(y) = \gamma - (1-\gamma)y, \qquad |r_h(x) - r_h(y)| = (1-\gamma)|x - y|.$$

Residuals can be $O(\gamma)$ in size without being $O(\gamma)$-Lipschitz. The bounds above use signed telescoping and fixed-point centering instead.

D.5.1 Comparison with total stability

Coupling and the comparison through the minimizers give

$$\mathbb{E}\,\|\Delta_t\| \le \min\{2\lambda_{\max}(P)G(1-\rho^t)/(\mu_P n),\ \|\Delta_\star\| + D_\Sigma \rho^t + 2R_t\}. \tag{D.45}$$

Write $\mathcal{G}_t = W_t^\delta + S_t^\delta + S_t^H + W_t^H$ and $\beta_t^{\mathrm{elem}} = \min\{a_{1,t}, \|\Delta_\star\| + D_\Sigma\rho^t + 2R_t\}$. Since $N_t^\Pi \ge a_{1,t}$,

$$\|\mathcal{B}_t\| + \mathcal{G}_t + N_t^\Pi \ge a_{1,t}. \tag{D.46}$$

Also $T_t^\Pi \ge D_\Sigma\rho^t + 2R_t + \chi(P)\rho^t\,\|\Delta_\star\|$, and $\|\Delta_\star\| \le \|\mathcal{B}_t\| + \|A^t\Delta_\star\| \le \|\mathcal{B}_t\| + \chi(P)\rho^t\,\|\Delta_\star\|$. Hence

$$\|\mathcal{B}_t\| + \mathcal{G}_t + T_t^\Pi \ge \|\Delta_\star\| + D_\Sigma\rho^t + 2R_t. \tag{D.47}$$

Taking the minimum gives

$$\|\mathcal{B}_t\| + \mathcal{G}_t + \min\{N_t^\Pi, T_t^\Pi\} \ge \beta_t^{\mathrm{elem}}. \tag{D.48}$$

Thus N_t^Π, T_t^Π do not improve this elementary total bound; this conclusion does not concern the other candidates. Indeed, if $\mathcal{C}_t$ is the signed sum of the four gradient convolutions, the exact identity $C_t^\Pi = e_t + A^t\Delta_\star - \mathcal{C}_t$ reduces the decomposition to $\Delta_t = \Delta_\star + e_t$. Separate norm bounds can lose this cancellation. Finally $\Delta_t - \mathcal{B}_t = e_t + A^t\Delta_\star$ and (D.38) give

$$\mathbb{E}\,\|\Delta_t - \mathcal{B}_t\| \le (D_\Sigma + \chi(P)\,\|\Delta_\star\|)\rho^t + 2R_t. \tag{D.49}$$

This needs no Hessian Lipschitzness. Its deterministic counterpart is in Appendix C.4.

The five expected convolution norms in (4.6) are bounded, in order, by $W_t^\delta, S_t^\delta, S_t^H, W_t^H, S_t^\Pi$. Consequently,

$$\mathbb{E}\,\|\Delta_t - \mathcal{B}_t\| \le W_t^\delta + S_t^\delta + S_t^H + W_t^H + S_t^\Pi, \qquad \mathbb{E}\,\|\Delta_t\| \le \|\mathcal{B}_t\| + W_t^\delta + S_t^\delta + S_t^H + W_t^H + S_t^\Pi. \tag{D.50}$$

D.6 Proof of Theorem 4.6

Proof. We prove the four gradient bounds, then assemble the decomposition.

Proving W_t^δ. Since $\mathbb{E}|\zeta_k| = 2(1-1/n)/n$ and $\mathbb{E}\zeta_k^2 = (1-1/n)/n$, the triangle and martingale inequalities give the two entries of (D.12). Using $d_\star \le 2G$ gives its explicit version in (D.16).

Proving S_t^δ. Smoothness and independence of J_k from $\mathcal{F}_k$ give

$$\mathbb{E}[J_k\,\|\delta_k - \delta_\star\|] \le \frac{2L}{n}\mathbb{E}\,\|\widetilde{x}_k - \widetilde{x}_\star\| \le \frac{2L}{n}(\widetilde{D}_0\rho^k + R_k).$$

Convolution proves (D.13). Its explicit form follows from $\sum_{k<t}\rho^{t-1-k}\rho^k = t\rho^{t-1}$ and $R_k \le R_\infty$.

Proving S_t^H. The segment comparison gives $\left\|\bar{H}_{i,k} - H_i\right\| \le (L_H/2)(\|x_k - x_\star\| + \|\widetilde{x}_k - \widetilde{x}_\star\|)$. By Minkowski and (D.11), the L^2 norm of the sum of distances is at most $D_\Sigma\rho^k + 2R_k$. Cauchy–Schwarz then yields

$$\mathbb{E}\,\|U_k\| \le \frac{L_H}{2}(D_\Sigma\rho^k + 2R_k)a_{2,k}.$$

Independently, the spectral interval $[\mu, L]$ gives $\left\|\bar{H}_{i,k} - H_i\right\| \le \Lambda$, hence $\mathbb{E}\,\|U_k\| \le \Lambda a_{1,k}$. Convolving their minimum proves (D.14). Using $a_{2,k} \le a_{2,\infty}$ and $R_k \le R_\infty$ gives its explicit version in (D.16).

Proving the alternative S_t^H bound. The second-moment formula and $s_k, v_k \le k$ give

$$a_{2,k} \le 2\gamma\lambda_{\max}(P)G\left(\frac{k}{n} + \sqrt{\frac{1-1/n}{n}}\sqrt{k}\right).$$

Retain both geometric factors in the initial-distance term:

$$\begin{aligned}\sum_{k<t}\rho^{t-1-k}\rho^k a_{2,k} &\le 2\gamma\lambda_{\max}(P)G\rho^{t-1}\left[\frac{t(t-1)}{2n} + \sqrt{\frac{1-1/n}{n}}\sum_{k<t}\sqrt{k}\right] \\ &\le 2\gamma\lambda_{\max}(P)G\rho^{t-1}\left[\frac{t^2}{2n} + \sqrt{\frac{1-1/n}{n}}t^{3/2}\right].\end{aligned}$$

For noise, use $\sum_{k<t}\rho^{t-1-k}R_k a_{2,k} \le R_\infty a_{2,\infty} s_t$. Multiply these bounds by $\gamma\lambda_{\max}(P)L_H D_\Sigma/2$ and $\gamma\lambda_{\max}(P)L_H$, respectively, to obtain (D.17).

Proving W_t^H. The spectral bound $\|H_i - H\| \le \Lambda$ gives $\mathbb{E}\|V_k\| \le \Lambda a_{1,k}$ and $\mathbb{E}\|V_k\|^2 \le \Lambda^2 a_{2,k}^2$. Triangle and martingale inequalities prove (D.15). Substituting $a_{1,k} \le 2\lambda_{\max}(P)G/(\mu_P n)$ and $a_{2,k} \le a_{2,\infty}$ gives (D.16).

Projection, assembly, and rates. Appendix D.5 proves $\mathbb{E}\left\|C_t^\Pi\right\| \le S_t^\Pi$ and (4.8). The triangle inequality in (4.6) now proves (D.50). To obtain the compact bound (4.7), apply the aggregation in Appendix B.3 with replacement smoothness L and fluctuation width $L - \mu \le L$. Replacing both widths by L yields exactly $\mathcal{E}_t^{\mathrm{p}}$. For fixed γ, the first-moment entries, $a_{1,k} = O(1/n)$ and $D_\Sigma \le 2\chi(P)\operatorname{diam}(K)$, give uniform-in-time $O(1/n)$ bounds. For the small-step rates, use

$$a_{2,\infty} = O(n^{-1} + \sqrt{\gamma/n}), \quad R_\infty = O(\sqrt{\gamma}), \quad \gamma s_\infty = \mu_P^{-1}, \quad \gamma\sqrt{v_\infty} = \sqrt{\frac{\gamma}{\mu_P(2-\gamma\mu_P)}}.$$

In (D.16), the transient $t\rho^{t-1}$ vanishes. The remaining entries give (D.18); the independent first-moment bound supplies the $O(1/n)$ minimum for S_t^H. The projection rates follow from (D.33) and (D.43). □

D.7 Stochastic projected generalization and excess risk

For Theorem 5.1, set

$$\beta_t^{\mathrm{s}} = \min\left\{\frac{2\lambda_{\max}(P)G}{\mu_P n}(1-\rho^t), \frac{2G}{\mu n} + 2\chi(P)\operatorname{diam}(K)\rho^t + 2R_t\right\}. \tag{D.51}$$

For $t \ge 0$, the bound is

$$\mathbb{E}\mathcal{R}(x_t) - \mathcal{R}(x_{\mathrm{opt}}) \le G\beta_t^{\mathrm{s}} + G\sqrt{\chi(P)^2\operatorname{diam}(K)^2\rho^{2t} + R_t^2}. \tag{D.52}$$

Proof. Synchronous coupling, Theorem D.1, and the replace-one identity prove

$$|\mathbb{E}[\mathcal{R}(x_t^{q,a}) - F_S(x_t^{q,a})]| \le G\beta_t^{q,a}. \tag{D.53}$$

Lipschitzness and (D.11) give

$$\mathbb{E}[F_S(x_t) - F_S(x_\star)] \le G\sqrt{D_0^2\rho^{2t} + R_t^2}. \tag{D.54}$$

Insert both estimates into (C.47) and use $D_0 \le \chi(P)\operatorname{diam}(K)$ to prove (D.52). Then take $t \to \infty$ followed by $\gamma \downarrow 0$. □

For the averaged output, the explicit bound is

$$\mathbb{E}\mathcal{R}(\bar{x}_T) - \mathcal{R}(x_{\mathrm{opt}}) \le \frac{\operatorname{diam}(K)^2}{2\gamma\lambda_{\min}(P)T} + \frac{\gamma\lambda_{\max}(P)G^2}{2} + \frac{G}{T}\sum_{t=0}^{T-1}\beta_t^{\mathrm{s}}. \tag{D.55}$$

D.8 Proof of Theorem 5.2

Proof. Convexity gives $\bar{x}_T \in K$, and synchronous coupling gives

$$\mathbb{E}\left\|\bar{x}_T - \widetilde{\bar{x}}_T\right\| \le \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|x_t - \widetilde{x}_t\right\| \le \frac{1}{T}\sum_{t=0}^{T-1}\beta_t^{\mathrm{s}}. \tag{D.56}$$

The replace-one argument bounds its generalization gap by $GT^{-1}\sum_{t<T}\beta_t^{\mathrm{s}}$.

For its empirical optimization error, compare projection to the feasible point $x_\star = \Pi_K^{P^{-1}}(x_\star)$:

$$\|x_{t+1} - x_\star\|_{P^{-1}}^2 \le \|x_t - \gamma P\nabla f_{I_t}(x_t) - x_\star\|_{P^{-1}}^2 \tag{D.57}$$

$$\le \|x_t - x_\star\|_{P^{-1}}^2 - 2\gamma\left\langle \nabla f_{I_t}(x_t), x_t - x_\star\right\rangle + \gamma^2\lambda_{\max}(P)G^2. \tag{D.58}$$

Take conditional expectation over I_t and use convexity of F:

$$2\gamma\mathbb{E}[F(x_t) - F(x_\star)] \le \mathbb{E}\left\|x_t - x_\star\right\|_{P^{-1}}^2 - \mathbb{E}\left\|x_{t+1} - x_\star\right\|_{P^{-1}}^2 + \gamma^2\lambda_{\max}(P)G^2. \tag{D.59}$$

Summing, discarding the final squared distance, and applying Jensen gives, conditionally on S,

$$\mathbb{E}[F(\bar{x}_T) - F(x_\star)] \le \frac{\|x_0 - x_\star\|_{P^{-1}}^2}{2\gamma T} + \frac{\gamma\lambda_{\max}(P)G^2}{2}. \tag{D.60}$$

Combine this with the generalization bound in (C.47), and use $\|x_0 - x_\star\|_{P^{-1}}^2 \le \mathrm{diam}(K)^2/\lambda_{\min}(P)$. This proves (D.55) and its stated scale. □

E Numerical experiments: protocol and additional results

Objectives and data. We consider two-dimensional parameters and $n = 64$ observations. For binary logistic regression,

$$f_i(x) = \log(1 + \exp(a_i^\top x)) - y_i a_i^\top x + \frac{\lambda}{2}\|x\|^2, \qquad \lambda = 0.25.$$

Before normalization, the feature coordinates are independent Gaussians with mean $(1, 0.15)$ and standard deviations $(0.22, 0.65)$. Labels are Bernoulli with success probability 0.85. Observation zero is replaced by feature $(0.6, -0.8)$ and the opposite label. There is no additional linear tilt in this logistic experiment. These are controlled synthetic examples, not predictive-performance benchmarks.

Four settings and numerical constants. For each objective, the regularized algorithms operate on $\mathbb{R}^2$ and the constrained algorithms operate on $K = [-0.35, 0.35]^2$, retaining the same ridge term. This ensures sample strong convexity in both settings. We use $P = \mathrm{diag}(0.5, 1)$ and $x_0 = (-0.25, 0.20)$. The P^{-1}-metric projection onto this box is coordinatewise clipping. The constrained empirical minimizers lie on the face $x_1 = 0.35$, so projection is active.

All global constants include both datasets. Writing G_{data} for a bound on data-loss gradients, the constants are

	G_{data}	L_ℓ	L_H
logistic	$\max_i \|a_i\|$	$\frac{1}{4}\max_i \|a_i\|^2$	$\frac{1}{6\sqrt{3}}\max_i \|a_i\|^3$

and $\mu = \lambda$, $L = \lambda + L_\ell$. For constrained bounds we use $G = G_{\mathrm{data}} + \lambda\sqrt{2}(0.35)$; for regularized bounds we use $G = G_{\mathrm{data}}$. The default step is $\gamma = 0.01/[L\lambda_{\max}(P)]$, giving $\gamma = 0.02$ for logistic regression.

Computing and checking the decomposition. We solve the empirical minimization problems by L-BFGS-B, polish on the free coordinates, and check the metric gradient-mapping residual. The secant matrices

$$H_i = \int_0^1 \nabla^2 f_i(\widetilde{x}_\star + s\Delta_\star)\, ds, \qquad H = \frac{1}{n}\sum_i H_i, \qquad A = I - \gamma PH$$

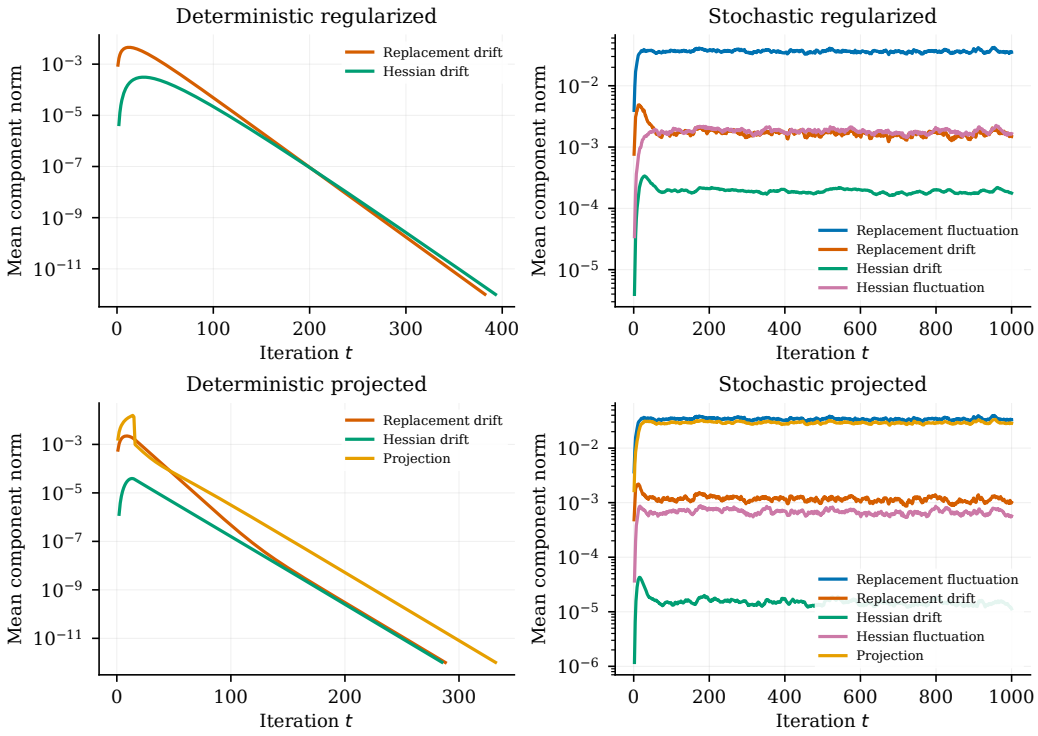


Figure 3: Binary logistic regression in the four settings. Comparison of the different components of this work's bounds. The represented values are the average over 128 runs of the experiment.

are evaluated by divided differences of $\tanh$ or the sigmoid, using their derivatives for nearly coincident arguments. The notebooks independently check gradients and Hessians by finite differences and check the secant identity.

Every convolution is propagated as a signed vector, $q_{t+1} = Aq_t + u_t$, before taking its norm. The projection forcing is centered at the difference of the two empirical fixed-point residuals. At every step and in every run, we check that the sum of the ERM displacement response and the applicable replacement drift, Hessian drift, fluctuation, and projection contributions reconstructs Δ_t. The default experiments give maximum absolute reconstruction errors below 2×10^{-15}. Separate component norms need not add to the norm of their sum.