Sparse Updates Degrade Generalization More Mildly Than Optimization: Stability Analysis for Randomized Subspace Descent

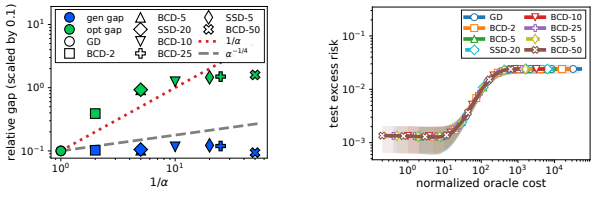
Yifei Liang[1,2] **Yan Sun**[3] **Yifei Cheng**[1] **Haobo Fu**[4] **Xiaochun Cao**[1] **Li Shen**[1,2]*
[1]Sun Yat-sen University, Shenzhen Campus, Shenzhen, China
[2]Shenzhen Loop Area Institute, Shenzhen, China
[3]Generative AI Lab, Nanyang Technological University, Singapore
[4]Tencent AI Platform Department, Shenzhen, China
liangyf65@mail2.sysu.edu.cn sun.yan.95@outlook.com
yfcheng.ifc@gmail.com haobofu@tencent.com
caoxiaochun@mail.sysu.edu.cn shenli6@mail.sysu.edu.cn

Abstract

Large-scale learning increasingly relies on updating only a coordinate, a block, or a low-dimensional subspace to reduce memory and communication costs, yet the statistical consequence of reduced update coverage remains unclear. Existing coordinate-level and common-step analyses do not jointly characterize how subspace coverage, blockwise step sizes, and output randomness affect population risk. We analyze randomized subspace descent (RSD), which recovers GD, RCD, BCD, and isotropic SSD. Under fixed, data-independent joint isotropy, expected projection energy reduces to a single coverage parameter α. For all $N \geq 1$ and $\alpha \in (0, 1]$, the convex excess-risk bound is $O(\min\{\alpha^{-1/4}N^{-1/2}, 1\})$, improving to $O((N\sqrt{\alpha})^{-1})$ under low noise, both milder than the $O(1/\alpha)$ optimization slowdown. In the nonconvex local Polyak–Łojasiewicz setting, smaller α lengthens the horizon and affects the bound through stopped-output moments. The analysis extends to the canonical blockwise step sizes $\eta_t = 1/L_{i_t}$ through a curvature-matched Lyapunov geometry. Hölder's inequality separates gradient and inverse-curvature moments without a uniform positive lower bound on neighboring curvature. Experiments on convex and nonconvex benchmarks and block-sparse GPT-2 fine-tuning support the predicted coverage–horizon trade-off.

1 Introduction

Efficient updates are central to large-scale learning. As models grow, the cost of maintaining optimizer states, communicating updates, or activating all parameters can rival the cost of computing the gradient of the loss.



(a) Opt. Gap vs. Gen. Gap (b) Oracle-cost comparison

Figure 1: (a) Optimization varies more with α than generalization. (b) Qualitative oracle-cost comparison (Appendix A).

A common remedy is to update only a restricted portion of the parameter space: a coordinate, a block, a low-dimensional subspace, or a selected module. This principle underlies classical coordinate and block coordinate descent [40, 47, 55], randomized subspace methods [23, 16, 9], and recent memory-efficient training of large language models [60, 61, 44, 37, 18].

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Table 1: Early-stopped excess-risk bounds for coordinate and subspace descent. Our convex BCD bounds assume $L_f < 2L_{\min}$; our nonconvex bounds use local Polyak–Łojasiewicz conditions. Bounds are in expectation; † marks a high-probability generalization-gap bound for deep networks.

Convexity

Algorithm	Step size	Coverage α	Training horizon	General convex case	Low-noise case
GD [27]	$1/L_f$	1	$\sqrt{N} \mid N$	$O(N^{-1/2})$	$O(N^{-1})$
RCD [54]	$\le 2/L_{\max}$	$1/d$	$d^{3/4}\sqrt{N} \mid \sqrt{d}\,N$	$O(d^{1/4}N^{-1/2})$	$O(\sqrt{d}\,N^{-1})$
RCD [56]	$\le 1/L_f$	$1/d$	$d\sqrt{N} \mid dN$	$O(N^{-1/2})^{\ddagger}$	$O(N^{-1})^{\ddagger}$
RSD (ours)	$1/L_f$	α	$\alpha^{-3/4}\sqrt{N} \mid N/\sqrt{\alpha}$	$O(\alpha^{-1/4}N^{-1/2})^{\star}$	$O((\sqrt{\alpha}\,N)^{-1})$
Randomized BCD (ours)	$1/L_i$	$1/m$	$m^{3/4}\sqrt{N} \mid \sqrt{m}\,N$	$O((1+K)m^{1/4}N^{-1/2})$	$O((1+K)\sqrt{m}\,N^{-1})$

Non-convexity

Algorithm	Step size	Coverage α	Excess-risk bound	Training horizon
GD [26]	$1/L_f$	1	$O(N^{-1})$	$\frac{L_f}{\beta_{\mathcal{S}}^{\star}} \log_+(N^2 L_f R_0)$
Cyclic BCD [2]	—	—	$\widetilde{O}(N^{-1/2})^{\dagger}$	—
RSD (ours)	$1/L_{\max}$	α	$\mathcal{E}_{\mathrm{nc}}$	$\frac{pL_{\max}}{\alpha\beta_{\mathcal{S}}^{\star}} \log_+(N^2 L_f R_0)$
Randomized BCD (ours)	$1/L_i$	$1/m$	$\mathcal{E}_{\mathrm{nc}}^{\mathcal{L}}$	$\frac{pm}{\beta_{\mathcal{S}}^{\mathcal{L}}} \log_+(N^2 L_f R_0)$

Horizons are rounded up; convex pairs list the general and low-noise cases. For RCD, $L_{\max}$ is coordinate smoothness; fixed smoothness constants are omitted. Here $K = L_f^2/[L_{\min}(2L_{\min} - L_f)]$. $\star$ Equation (7) gives uniform and two-term bounds. ‡ For pairwise losses, the stated common step size is sufficient. † This result has no α dependence. General-convex rates require $m = O(N^2)$ for BCD and $d = O(N^2)$ for RCD [54]; $\mathcal{E}_{\mathrm{nc}}$ and $\mathcal{E}_{\mathrm{nc}}^{\mathcal{L}}$ are defined in (12) and (22), respectively.

These restricted updates change how training data shape the learned model through the update directions available at each iteration, raising a learning-theoretic question:

How does reduced update coverage affect population risk?

Optimization theory addresses only one side of this question: convergence analyses quantify how sparse updates slow empirical optimization [40, 47, 55, 23, 9]. Population risk, however, depends jointly on optimization and generalization. Algorithmic stability studies the latter by measuring how the output changes when a single training sample is replaced [6, 17], yet three obstacles remain: *(i)* coordinate-level analyses exploit the orthogonality of coordinate projectors [54, 56], whereas general subspaces may overlap or be non-axis-aligned; *(ii)* applying algorithm-dependent nonconvex stability [26] to RSD requires curvature control at an output that remains random even for a fixed dataset; *(iii)* common-step analyses do not track the blockwise step size $1/L_i$ standard in block coordinate descent (BCD), which couples update scale to local curvature. These results leave open a joint stability analysis of randomized subspace coverage and nonuniform blockwise step sizes.

We study this question through randomized subspace descent (RSD), which recovers GD, RCD, BCD, and isotropic stochastic subspace descent (SSD) as special cases. Under joint isotropy, the expected projection energy of any subspace family reduces to a single coverage parameter $\alpha = \sum_{i=1}^{m} \rho_i \frac{r_i}{d}$, the average fraction of the ambient space explored per update. This scalar replaces coordinate-level orthogonality, making stability analysis tractable for general subspace families and resolving obstacle (i). The central finding, illustrated qualitatively in Figure 1 and summarized in Table 1, is that the statistical cost of sparse updates, as measured by our convex excess-risk bounds, is milder than their optimization cost. In the convex case, optimization slows as $1/\alpha$, whereas the stability bound carries no inverse-coverage factor; early stopping gives the uniform excess-risk bound $O(\min\{\alpha^{-1/4}N^{-1/2}, 1\})$, which improves to $O((N\sqrt{\alpha})^{-1})$ under low noise. In the nonconvex local Polyak–Łojasiewicz setting, smaller α lengthens the early-stopping horizon but does not enter the excess-risk bound as an explicit multiplier; the bound is instead governed by the sample-gradient moment $\mathfrak{G}_p(T)$ and the inverse-curvature moment $\mathfrak{I}_q(T)$ of the stopped output.

We use on-average model stability [12, 27] to connect sample replacement to population risk. In the convex case, neighboring runs share the subspace sequence, and joint isotropy expresses the projection energy in the perturbation recursion through α.

For BCD with the blockwise step size $\eta_t = 1/L_{i_t}$, a block Lyapunov geometry $D_{\mathcal{L}}$ matches the curvature-dependent update scale, yielding curvature-weighted bounds and exposing the sufficient balance condition $L_f < 2L_{\min}$, thereby addressing obstacle (iii). The nonconvex residual and inverse

curvature both depend on the output $\boldsymbol{x}^T$, which is random given $\mathcal{S}$. Hölder's inequality separates their effects through $\mathfrak{G}_p(T)$ and $\mathfrak{I}_q(T)$, controlling stability without a uniform positive curvature lower bound (obstacle (ii)). Experiments on convex and nonconvex benchmarks and block-sparse GPT-2 fine-tuning illustrate the trade-off (Section 5). Our contributions are as follows.

- **Sparse coverage incurs a fractional penalty in convex excess-risk bounds.** We establish on-average stability and excess-risk bounds for RSD under joint isotropy, unifying GD ($\alpha=1$), RCD ($\alpha = 1/d$), BCD ($\alpha = 1/m$), and isotropic SSD ($\alpha = \bar{r}/d$) within a single coverage-based analysis. The general-convex bound is $O(\min\{\alpha^{-1/4}N^{-1/2}, 1\})$ for all N and α, and Corollary 4.13 also gives the two-term bound; under low noise it improves to $O((N\sqrt{\alpha})^{-1})$. At $\alpha = 1/d$ this recovers the fractional dimension dependence known for RCD [54], extended here to general subspace families with overlapping or non-axis-aligned update directions.
- **Curvature balance governs the excess-risk bounds for BCD with blockwise step sizes.** A block Lyapunov geometry $D_{\mathcal{L}}$ replaces worst-case Euclidean control with a curvature-weighted potential tailored to the step size $1/L_i$. Under the sufficient condition $L_f < 2L_{\min}$, it yields stability and excess-risk bounds with explicit dependence on the curvature margin.
- **Nonconvex stability follows from randomized-output moment control.** Hölder's inequality bounds the coupled gradient and inverse-curvature terms by separate moments, without requiring a uniform positive curvature lower bound at the output. Coverage α affects both the sufficient horizon and these moments. Uniform moment control gives an excess-risk bound uniform in α.

2 Related work

Randomized Subspace Descent. RCD updates one randomly selected coordinate per iteration. Nesterov [40] established global convergence rates for large-scale problems, and Wright [55] provides a comprehensive survey. BCD extends RCD by updating blocks of variables, with foundational complexity results due to Richtárik and Takáč [47]. RSD allows updates in possibly non-coordinate low-dimensional subspaces. Kozak et al. [23] studied stochastic subspace descent for PDE-constrained problems, and Chen et al. [9] analyzed randomized fast subspace descent for convex objectives. A separate line of work has shown empirically that fine-tuning pre-trained language models often takes place in low-dimensional intrinsic subspaces [1, 60], motivating low-rank gradient projection methods such as GaLore [61] and online subspace descent [30, 10]. LoRA [19] instead parameterizes task-specific low-rank adaptations. He et al. [18] provided convergence guarantees for subspace optimization in LLMs. Appendix B reviews additional BCD variants, memory-efficient LLM training methods, and subspace ideas in federated learning and other domains. Our analysis is closest in algorithmic scope to stochastic and randomized subspace descent [23, 9, 18], but examines how restricting the update subspace affects sensitivity to training data and, in turn, population risk. We use RSD as a unified model for BCD and RCD and derive stability-based generalization and excess-risk bounds for isotropic orthogonal-projector updates beyond coordinate-level settings. Coverage provides a common scale for comparing these update families by measuring the expected fraction of the ambient space available at each step, including jointly isotropic sampling laws that mix subspaces of different dimensions or assign different probabilities to each subspace.

Stability-based Generalization Analysis. Bousquet and Elisseeff [6] introduced uniform stability and derived generalization bounds from leave-one-out perturbations. Hardt et al. [17] showed that SGD becomes more stable with early stopping under Lipschitz and smooth losses, leading to many refinements; see Appendix B for a longer discussion. Lei and Ying [27] developed fine-grained on-average stability bounds for SGD, which form the starting point of our convex analysis. For coordinate-type methods, Wang et al. [54] established the first stability bounds for RCD, later extended to pairwise learning by Wu et al. [56]. Akiyama [2] obtained Rademacher-complexity-based bounds for BCD in neural networks. Huang et al. [20] studied stability of randomly compressed gradients. Lei et al. [26] developed stability bounds for nonconvex algorithms using an algorithm-dependent relaxation of the Polyak–Łojasiewicz condition, with explicit dependence on optimization error and local curvature at the learned output. For RSD, including BCD and RCD, we combine subspace coupling with randomized-output moments to quantify how coverage governs the balance between optimization progress and sensitivity to the training sample. The BCD bounds also retain the dependence on individual block step sizes and curvature differences, linking the statistical guarantees to the same quantities that determine each update.

3 Preliminaries

Our population-risk analysis combines the optimization properties of RSD with its sensitivity to the training sample. We first define the statistical learning problem and randomized subspace update, then introduce the stability and risk quantities. Appendix H summarizes the notation.

3.1 Problem formulation

Let $f(\boldsymbol{x};\xi)$ denote the per-sample loss, where $\boldsymbol{x} \in \mathbb{R}^d$ is the optimization variable and ξ is a data sample drawn from an unknown distribution $\mathcal{D}$. Our ultimate goal is to minimize the population risk

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} F(\boldsymbol{x}) := \mathbb{E}_{\xi\sim\mathcal{D}}[f(\boldsymbol{x};\xi)], \qquad \boldsymbol{x}^\star \in \arg\min_{\boldsymbol{x}} F(\boldsymbol{x}). \tag{1}$$

Since $\mathcal{D}$ is unknown, we work with an i.i.d. sample $\mathcal{S} = \{\xi_1, \dots, \xi_N\}$ and define the empirical risk $F_{\mathcal{S}}(\boldsymbol{x}) := \frac{1}{N}\sum_{n=1}^{N} f(\boldsymbol{x};\xi_n)$. The corresponding empirical risk minimization (ERM) problem is

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} F_{\mathcal{S}}(\boldsymbol{x}), \qquad \boldsymbol{x}^\star_{\mathcal{S}} \in \arg\min_{\boldsymbol{x}} F_{\mathcal{S}}(\boldsymbol{x}). \tag{2}$$

Throughout the paper, we assume that the minimizer $\boldsymbol{x}^\star_{\mathcal{S}}$ exists for every dataset $\mathcal{S}$. We next introduce the optimization algorithm used to solve Eq. (2) through randomized subspace updates.

3.2 Algorithm

We consider a family of subspaces indexed by $i \in \{1, \dots, m\}$, where type i is represented by a matrix $P_i \in \mathbb{R}^{d\times r_i}$ with $r_i \le d$ the subspace dimension. P_i may be deterministic, as in block coordinate methods, or random, as in isotropic random subspace methods. We write ρ_i for the probability of sampling type i, with $\rho_i > 0$ and $\sum_{i=1}^m \rho_i = 1$. If P_i is random, we draw a fresh realization after selecting type i, independently of the current iterate and history.

For each type i, we associate a constant $L_i > 0$, interpreted as the smoothness constant of $F_{\mathcal{S}}$ restricted to the subspace generated by P_i. Throughout, we impose the normalization

$$P_i^\top P_i = I_{r_i}, \tag{3}$$

so that the columns of P_i form an orthonormal basis of the sampled subspace. Under Eq. (3), $P_i P_i^\top$ is the orthogonal projector onto $\mathrm{span}(P_i)$. If $F_{\mathcal{S}}$ is globally L_f-smooth, we choose valid subspace smoothness constants $L_i \le L_f$, since L_f is a valid bound for every orthonormal subspace.

Algorithm 1 gives the basic RSD scheme. At iteration t, the algorithm samples a subspace type $i_t \sim (\rho_1, \dots, \rho_m)$ and, if the corresponding matrix is random, draws a fresh realization conditionally on i_t. To define the update, let $\boldsymbol{z} \in \mathbb{R}^{r_{i_t}}$ denote the local coordinate in $\mathrm{span}(P_{i_t})$, and consider the restricted function $\boldsymbol{z} \mapsto F_{\mathcal{S}}(\boldsymbol{x}^t + P_{i_t}\boldsymbol{z})$. By the chain rule, its gradient at $\boldsymbol{z} = \boldsymbol{0}$ is $P_{i_t}^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}^t)$, which is the sampled subspace gradient in local coordinates. The RSD update therefore takes one gradient step in this subspace and maps the step back to the full ambient space.

Algorithm 1 Randomized Subspace Descent (RSD)

Require: Initial point $\boldsymbol{x}^0 \in \mathbb{R}^d$; empirical risk $F_{\mathcal{S}}$; subspace family $\{(P_i, L_i, \rho_i)\}_{i=1}^m$; step sizes $\{\eta_t\}$; iteration budget T

1: **for** $t = 0, 1, \dots, T-1$ **do**
2: Sample a type $i_t \sim (\rho_1, \dots, \rho_m)$
3: If P_{i_t} is random, draw a fresh realization of P_{i_t}
4: $\boldsymbol{g}^t_{i_t} \leftarrow P_{i_t}^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}^t)$
5: Update $\boldsymbol{x}^{t+1} \leftarrow \boldsymbol{x}^t - \eta_t P_{i_t} \boldsymbol{g}^t_{i_t}$
6: **end for**
7: **return** $\boldsymbol{x}^T$

Equivalently, substituting $\boldsymbol{g}^t_{i_t} = P_{i_t}^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}^t)$ into the update yields the projected-gradient form

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t\, P_{i_t} P_{i_t}^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}^t). \tag{4}$$

We summarize the average projection geometry through the following joint isotropy relation

$$\sum_{i=1}^m \rho_i\, \mathbb{E}\big[P_i P_i^\top\big] = \alpha I_d,\ \alpha := \sum_{i=1}^m \rho_i \frac{r_i}{d}. \tag{5}$$

The scalar $\alpha \in (0, 1]$ measures the average fraction of the ambient space covered per randomized subspace step. Write $\Pi_t := P_{i_t} P_{i_t}^\top$ for the projector used at iteration t. Because the type and the realization are drawn afresh and independently of the past, (5) gives $\mathbb{E}[\Pi_t \mid \boldsymbol{x}^t] = \alpha I_d$. For a common step-size schedule, $\alpha \sum_t \eta_t$ is the cumulative expected update scale, called the optimization clock below. We also use the projection-energy identity to control sensitivity to sample replacement. Throughout, the sampling law is fixed and independent of the data and the iterates. Coupled runs share a deterministic, data-independent initialization and a predetermined common step-size schedule, except for BCD with the prescribed sampled-block step $1/L_{i_t}$. Algorithm 1 contains several standard first-order methods as special cases under particular subspace sampling distributions.

Random BCD. Partition coordinates into m disjoint blocks with dimensions $r_1, \ldots, r_m$ satisfying $\sum_{i=1}^m r_i = d$. Let $U_i \in \mathbb{R}^{d \times r_i}$ be the block-selection matrix with columns the standard basis vectors of block i. Setting $P_i = U_i$ and $\rho_i = 1/m$ yields $\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t U_{i_t} U_{i_t}^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}^t)$, the standard randomized BCD update. In this case, $\alpha = 1/m$ even when the block sizes differ, so independent uniform sampling from a fixed complete partition into layers or modules is also covered.

Random RCD. Set $m = d$, $r_i = 1$, $P_i = e_i$, and $\rho_i = 1/d$. Then $P_i P_i^\top = e_i e_i^\top, \alpha = \frac{1}{d}$. Equation (4) reduces to the standard randomized single-coordinate gradient descent update.

Isotropic SSD. Suppose each P_i is random, has orthonormal columns, and satisfies $P_i^\top P_i = I_{r_i}, \mathbb{E}[P_i P_i^\top] = \frac{r_i}{d} I_d$. Under uniform sampling, $\rho_i = 1/m$, this gives $\alpha = \frac{\bar{r}}{d}, \bar{r} := \frac{1}{m} \sum_{i=1}^m r_i$. The equal-dimension isotropic RSD scheme corresponds to the special case $r_i \equiv r$, giving $\alpha = r/d$.

3.3 Stability and generalization

Applying the learning algorithm $\mathcal{A}$ to $\mathcal{S}$ produces an output $\mathcal{A}(\mathcal{S}) \in \mathbb{R}^d$. To control the excess population risk, we separate the statistical and optimization contributions of the algorithmic output.

Definition 3.1. Given a dataset $\mathcal{S}$ and a randomized learning algorithm $\mathcal{A}$, we define:

(i) Generalization error is $\epsilon_{\text{gen}} := \mathbb{E}[F(\mathcal{A}(\mathcal{S})) - F_{\mathcal{S}}(\mathcal{A}(\mathcal{S}))]$, i.e. the expected gap between population and empirical risk at the output.

(ii) Optimization error is $\epsilon_{\text{opt}} := \mathbb{E}[F_{\mathcal{S}}(\mathcal{A}(\mathcal{S})) - F_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}}^\star)]$, i.e. the expected empirical risk gap relative to an ERM solution on the same training sample.

Both expectations are over $\mathcal{S}$ and the internal randomness of $\mathcal{A}$.

Furthermore, the *expected excess population risk* can be decomposed as

$$\mathbb{E}[F(\mathcal{A}(\mathcal{S}))] - F(\boldsymbol{x}^\star) = \epsilon_{\text{gen}} + \epsilon_{\text{opt}} + \underbrace{\mathbb{E}[F_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}}^\star) - F(\boldsymbol{x}^\star)]}_{\leq 0}. \tag{6}$$

Hence, it suffices to control the generalization error and the optimization error of the algorithm.

To study the generalization term, we compare the outputs produced from neighboring datasets formed by independently replacing one training observation. Let $\mathcal{S}' = \{\xi_1', \ldots, \xi_N'\}$ be an independent copy of $\mathcal{S}$. For each $n \in [N]$, define $\mathcal{S}^{(n)} := \{\xi_1, \ldots, \xi_{n-1}, \xi_n', \xi_{n+1}, \ldots, \xi_N\}$. We write $\{\boldsymbol{x}^t\}_{t=0}^T$ and $\{\boldsymbol{x}^{(n),t}\}_{t=0}^T$ for the iterates generated on $\mathcal{S}$ and $\mathcal{S}^{(n)}$, respectively, using the same subspace draws. Sharing these draws isolates the effect of replacing one training example.

Definition 3.2 (On-average loss stability)**.** Let $\mathcal{S}, \mathcal{S}'$ be two datasets of size N drawn independently from $\mathcal{D}$, and let $\mathcal{S}^{(n)}$ be defined above. We say that $\mathcal{A}$ is *on-average $\epsilon_{\mathcal{A}}$-stable* if

$$\frac{1}{N} \sum_{n=1}^N \mathbb{E}_{\mathcal{A},\mathcal{S},\mathcal{S}'} \left[f(\mathcal{A}(\mathcal{S}^{(n)}); \xi_n) - f(\mathcal{A}(\mathcal{S}); \xi_n) \right] \leq \epsilon_{\mathcal{A}}.$$

We first bound the root-mean-square distance between outputs on neighboring datasets (on-average argument stability), then apply the following lemma to control generalization in expectation.

Lemma 3.3 (Stability and generalization [27])**.** *Let $\mathcal{A}$ be any learning algorithm. If, for every ξ, the function $f(\cdot; \xi)$ is nonnegative and L_f-smooth, then for any $\gamma > 0$,*

$$\epsilon_{\text{gen}} \leq \frac{1}{\gamma} \mathbb{E}[F_{\mathcal{S}}(\mathcal{A}(\mathcal{S}))] + \frac{L_f(1+\gamma)}{2N} \sum_{n=1}^N \mathbb{E}_{\mathcal{A},\mathcal{S},\mathcal{S}'} \left[\|\mathcal{A}(\mathcal{S}) - \mathcal{A}(\mathcal{S}^{(n)})\|_2^2 \right].$$

4 Theoretical results

Convexity controls the separation of neighboring trajectories; local PŁ and output moments connect nonconvex optimization to population risk. We first state the shared regularity assumptions.

4.1 Basic assumptions

Throughout this section, the subspace family satisfies the orthonormal-columns normalization (3) and the joint-isotropy condition (5). The remaining assumptions specify the nonnegativity and smoothness conditions used in the empirical-risk descent and algorithmic stability estimates below.

Assumption 4.1 (Nonnegative Loss)**.** For every sample ξ, $f(\boldsymbol{x};\xi) \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^d$.

Assumption 4.2 (L_f-Smooth)**.** For every sample ξ, the per-sample loss $f(\cdot;\xi)$ is continuously differentiable and L_f-smooth, with $L_f > 0$: for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, $\|\nabla f(\boldsymbol{y};\xi) - \nabla f(\boldsymbol{x};\xi)\|_2 \leq L_f\|\boldsymbol{y}-\boldsymbol{x}\|_2$. In particular, $f(\boldsymbol{y};\xi) \leq f(\boldsymbol{x};\xi) + \langle \nabla f(\boldsymbol{x};\xi), \boldsymbol{y}-\boldsymbol{x}\rangle + \frac{L_f}{2}\|\boldsymbol{y}-\boldsymbol{x}\|_2^2, \forall\, \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$.

Remark 4.3 (Smoothness of the empirical risk). Under Assumption 4.2, averaging the per-sample gradients preserves L_f-Lipschitz continuity. Thus $F_{\mathcal{S}}$ is L_f-smooth for every dataset.

Assumption 4.4 (Subspace-wise Smoothness)**.** For every dataset $\mathcal{S}$ and subspace type $i \in \{1,\dots,m\}$, the empirical risk $F_{\mathcal{S}}$ is L_i-smooth along the subspace generated by P_i. That is, for any realization of P_i (if random), any $\boldsymbol{x} \in \mathbb{R}^d$, and any $\boldsymbol{z} \in \mathbb{R}^{r_i}$, $F_{\mathcal{S}}(\boldsymbol{x}+P_i\boldsymbol{z}) \leq F_{\mathcal{S}}(\boldsymbol{x}) + \langle P_i^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}), \boldsymbol{z}\rangle + \frac{L_i}{2}\|\boldsymbol{z}\|_2^2$. Define $L_{\max} := \max_{1\leq i\leq m} L_i$ and $L_{\min} := \min_{1\leq i\leq m} L_i$.

Remark 4.5 (Global and subspace smoothness). Assumptions 4.2–4.4 are standard in stability and subspace descent [17, 26, 40, 4, 23, 56]. Orthonormal columns let us choose $L_i \leq L_f$, with smaller constants when a subspace is smoother. These assumptions require smoothness without a separate bounded-gradient condition. Bounded covariates and smooth nonnegative losses give examples in generalized linear models. In smooth neural networks, global smoothness requires a uniformly bounded parameter Hessian, while bounded inputs and parameters can yield regional bounds.

4.2 Convex case

Definition 4.6 (Convexity)**.** For every sample ξ, we say that the differentiable loss $f(\cdot;\xi)$ is convex on $\mathbb{R}^d$ if, for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, $f(\boldsymbol{y};\xi) \geq f(\boldsymbol{x};\xi) + \langle \nabla f(\boldsymbol{x};\xi), \boldsymbol{y}-\boldsymbol{x}\rangle$.

Theorem 4.7 (On-average argument stability in the convex case)**.** *Suppose Definition 4.6, Assumptions 4.1–4.4, and the structural conditions (3)–(5) hold. Let $\boldsymbol{x}^0 \in \mathbb{R}^d$ be deterministic and let $0 < \eta_{t+1} \leq \eta_t \leq 1/L_f$ for all $t \geq 0$. Then, for every $T \geq 1$ and every deterministic comparator $\boldsymbol{x} \in \mathbb{R}^d$ independent of $\mathcal{S}$, $\mathcal{S}'$, and the algorithmic randomness,*

$$\left(\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\|\boldsymbol{x}^T - \boldsymbol{x}^{(n),T}\|_2^2\Big]\right)^{1/2} \leq \frac{2\sqrt{L_f T}}{N}\left[\eta_0\|\boldsymbol{x}^0-\boldsymbol{x}\|_2^2 + 2\eta_0^2 F(\boldsymbol{x}^0) + 2\alpha\Big(\sum_{t=0}^{T-1}\eta_t^2\Big)F(\boldsymbol{x})\right]^{1/2}.$$

Remark 4.8 (Coverage in convex optimization and stability). At fixed horizons with common steps, the optimization bound carries $1/\alpha$ (Theorem C.1), while the stability bound has no such factor (Theorem 4.7). Sharing the subspace draw isolates sample replacement, and joint isotropy scales its squared perturbation by α, canceling $1/\alpha$ from accumulated empirical risk. Expected nonexpansiveness controls the shared update (Lemma F.7). Coverage remains only in the comparator-risk term, which is small at $\boldsymbol{x} = \boldsymbol{x}^\star$ in realizable or low-noise settings. Setting $\alpha = 1/d$ recovers the RCD coverage dependence [54, 56]. See Appendix G.1.3.

The BCD refinement uses a curvature-weighted stability analysis for blockwise step sizes. For randomized coordinate and block coordinate descent, the standard step for block i is the inverse block smoothness constant [40, 47], that is, $\eta_t = 1/L_{i_t}$. Compared with a uniform step set by $L_{\max}$, this rule takes larger steps on smoother blocks and smaller steps on sharper blocks. The analysis uses the block-weighted squared distance $D_{\mathcal{L}}(\boldsymbol{y},\boldsymbol{z}) := \frac{1}{2}\sum_{i=1}^m L_i\|U_i^\top(\boldsymbol{y}-\boldsymbol{z})\|_2^2$. The next corollary bounds output sensitivity to replacing one training observation under a sufficient curvature-balance condition for these blockwise step sizes and makes the resulting curvature dependence explicit.

Corollary 4.9 (Stability under BCD with blockwise step sizes)**.** *Suppose Definition 4.6, Assumptions 4.1–4.4, and Eq. (3) hold. Consider BCD with $P_i = U_i$, $\rho_i = 1/m$, $\sum_{i=1}^m U_i U_i^\top = I_d$, and blockwise step sizes $\eta_t = 1/L_{i_t}$, with $\boldsymbol{x}^0 \in \mathbb{R}^d$ deterministic. Assume $L_f < 2L_{\min}$. For any $T \geq 1$ and deterministic $\boldsymbol{x} \in \mathbb{R}^d$ independent of $\mathcal{S}$, $\mathcal{S}'$, and algorithmic randomness,*

$$\left(\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\|\boldsymbol{x}^T - \boldsymbol{x}^{(n),T}\|_2^2\Big]\right)^{1/2} \leq \frac{4}{N}\sqrt{\frac{L_f T}{L_{\min}(2L_{\min} - L_f)}\left[D_{\mathcal{L}}(\boldsymbol{x}^0, \boldsymbol{x}) + F(\boldsymbol{x}^0) + \tfrac{T}{m}F(\boldsymbol{x})\right]}.$$

Remark 4.10 (Block geometry and curvature balance). In Corollary 4.9, $D_{\mathcal{L}}$ weights the initial comparator error by block curvature. The fixed orthogonal partition makes this recursion telescope under $\eta_t = 1/L_{i_t}$. The bound weakens as the sufficient curvature margin $2L_{\min} - L_f$ narrows. Figure 8 illustrates behavior beyond this regime. The derivation of the margin and the proof of the stability bound appear in Appendices D.1 and G.1.4, respectively.

Theorem 4.11 (Generalization error in the convex case)**.** *Suppose Definition 4.6, Assumptions 4.1–4.4, and the structural conditions (3)–(5) hold. Let $\boldsymbol{x}^0 \in \mathbb{R}^d$ be deterministic and let $0 < \eta_{t+1} \leq \eta_t \leq 1/L_f$ for all $t \geq 0$. Then, for every $T \geq 1$ and every $\gamma > 0$,*

$$\begin{aligned}\epsilon_{\text{gen}} \leq \frac{1}{\gamma}&\left[F(\boldsymbol{x}^\star) + \frac{\eta_0\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|_2^2 + 2\eta_0^2 F(\boldsymbol{x}^0)}{2\alpha\sum_{t=0}^{T-1}\eta_t^2}\right] \\ &+ \frac{2L_f^2(1+\gamma)T}{N^2}\left[\eta_0\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|_2^2 + 2\eta_0^2 F(\boldsymbol{x}^0) + 2\alpha\Big(\sum_{t=0}^{T-1}\eta_t^2\Big)F(\boldsymbol{x}^\star)\right],\end{aligned}$$

where the expectation in ϵ_{gen} is over $\mathcal{S}$, $\mathcal{S}'$, and all algorithmic randomness.

Remark 4.12 (Balancing optimization and generalization). The weights $1/\gamma$ and $(1+\gamma)T/N^2$ balance empirical risk and sample sensitivity in Theorem 4.11. As T grows, the leading optimization term $1/(\alpha T)$ falls and the generalization term $\sqrt{\alpha}\,T/N$ rises. Their balance gives the $\alpha^{-1/4}$ coverage cost for $a \leq 1$ in Corollary 4.13. Under low noise, balancing $1/(\alpha T)$ against T/N^2 permits a longer horizon and gives the $\alpha^{-1/2}$ cost. See Appendix G.1.5.

Corollary 4.13 (Excess-risk scaling under early stopping)**.** *Suppose the conditions of Theorem 4.11 hold with constant step size $\eta_t \equiv 1/L_f$. Assume that there exists $C_0 > 0$ such that $L_f\,\mathbb{E}[\|\boldsymbol{x}^0 - \boldsymbol{x}^\star_{\mathcal{S}}\|_2^2] + L_f\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|_2^2 + F(\boldsymbol{x}^0) + F(\boldsymbol{x}^\star) \leq C_0$. Let $N \geq 1$, $\alpha \in (0,1]$, and $a = \alpha^{-1/4}N^{-1/2}$. Setting $T = \lceil \alpha^{-3/4}\sqrt{N}\rceil$ and $\gamma = N/(\sqrt{\alpha}\,T)$ gives*

$$\mathbb{E}[F(\boldsymbol{x}^T)] - F(\boldsymbol{x}^\star) \leq 40C_0\min\{a, 1\} \leq 40C_0\left[\frac{1}{\alpha^{1/4}\sqrt{N}} + \frac{1}{\sqrt{\alpha}\,N}\right]. \tag{7}$$

If additionally $F(\boldsymbol{x}^\star) \leq C_0/N$ (low-noise regime), then setting $T = \lceil N/\sqrt{\alpha}\rceil$ and $\gamma = 1$ yields, again for all $N \geq 1$ and $\alpha \in (0,1]$,

$$\mathbb{E}\big[F(\boldsymbol{x}^T)\big] - F(\boldsymbol{x}^\star) = O\left(\frac{C_0}{N\sqrt{\alpha}}\right).$$

All numerical constants in these bounds are universal.

Remark 4.14 (Coverage scaling and algorithm specializations). In the regime $a \leq 1$, Equation (7) gives a coverage penalty of $\alpha^{-1/4}$; under low noise, the penalty is $\alpha^{-1/2}$. Both are milder than the $1/\alpha$ optimization slowdown. The uniform form sharpens the two-term estimate and stays bounded when coverage is very small. Consistency follows when $a \to 0$ with uniformly bounded C_0, using the general-convex horizon $T \asymp \alpha^{-3/4}\sqrt{N}$. Appendix G.1.7 proves both bounds.

Substituting $\alpha = 1, 1/d, 1/m, \bar{r}/d$ gives GD, RCD, uniform BCD, and isotropic SSD. The RCD comparison with [54] concerns coverage scaling under $\eta_t \leq 1/L_f$. BCD with $\eta_t = 1/L_{i_t}$ retains the same order in m, with $D_{\mathcal{L}}$ replacing Euclidean quantities. Its curvature factor is $1 + K$, where $K = L_f^2/[L_{\min}(2L_{\min} - L_f)]$; the general-convex rate requires $m = O(N^2)$ (Appendix D).

4.3 Nonconvex case

In nonconvex RSD, the residual and curvature share the output $\boldsymbol{x}^T$, which is random conditional on $\mathcal{S}$. For $\eta_t = 1/L_{\max}$, smoothness and joint isotropy give the stationarity bound in Lemma F.3.

$$\min_{0\le t<T} \mathbb{E}_{\mathcal{A}|\mathcal{S}} \|\nabla F_{\mathcal{S}}(\boldsymbol{x}^t)\|_2^2 \le \frac{2L_{\max}\big(F_{\mathcal{S}}(\boldsymbol{x}^0) - \inf F_{\mathcal{S}}\big)}{\alpha T}. \tag{8}$$

Equation (8) controls the best expected squared gradient norm along the trajectory. The last-iterate residual bound uses local Polyak–Łojasiewicz (PŁ) to relate gradients to the empirical residual.

Definition 4.15 (Local PŁ condition)**.** Let $\mathcal{U}_0 = \{\boldsymbol{x} \in \mathbb{R}^d : F_{\mathcal{S}}(\boldsymbol{x}) \le F_{\mathcal{S}}(\boldsymbol{x}^0)\}$. We say that $F_{\mathcal{S}}$ satisfies the local PŁ condition on the entire initial sublevel set $\mathcal{U}_0$ with fixed parameter $\beta_{\mathcal{S}} > 0$ if

$$F_{\mathcal{S}}(\boldsymbol{x}) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}}) \le \frac{1}{2\beta_{\mathcal{S}}} \|\nabla F_{\mathcal{S}}(\boldsymbol{x})\|_2^2, \qquad \forall\, \boldsymbol{x} \in \mathcal{U}_0.$$

Larger $\beta_{\mathcal{S}}$ tightens the residual bound at a fixed gradient norm. For blockwise BCD, $\beta^{\mathcal{L}}_{\mathcal{S}} > 0$ relates the residual to the block-scaled gradient energy.

$$\textstyle\sum_{i=1}^m \frac{1}{L_i} \|U_i^\top \nabla F_{\mathcal{S}}(\boldsymbol{x})\|_2^2 \ge 2\beta^{\mathcal{L}}_{\mathcal{S}} \big(F_{\mathcal{S}}(\boldsymbol{x}) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\big), \qquad \forall\, \boldsymbol{x} \in \mathcal{U}_0. \tag{9}$$

Remark 4.16 (Local PŁ on the initial sublevel set). The prescribed steps keep empirical risk nonincreasing and all iterates in $\mathcal{U}_0$, where local PŁ is imposed [22, 23, 26]. Appendix C.2 gives certificates on the full set for rank-deficient least squares, a nonconvex generalized linear model, and interpolating models with a uniformly nondegenerate tangent Gram matrix.

For positive empirical residual, define the pointwise ratio $\beta^{\mathrm{pt}}_{\mathcal{S}}(\boldsymbol{x}) := \|\nabla F_{\mathcal{S}}(\boldsymbol{x})\|_2^2 / \big[2(F_{\mathcal{S}}(\boldsymbol{x}) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}}))\big]$, and set $\beta^{\mathrm{pt}}_{\mathcal{S}}(\boldsymbol{x}) = +\infty$ at empirical minimizers. Local PŁ means $\beta^{\mathrm{pt}}_{\mathcal{S}}(\boldsymbol{x}) \ge \beta_{\mathcal{S}}$ on $\mathcal{U}_0$. The expectations must be well-defined. If the neighboring ratio vanishes with positive probability, the bound in Lemma 4.17 is infinite.

Lemma 4.17 (Nonconvex stability [26])**.** *Suppose Assumption 4.2 holds. Let $\boldsymbol{x}^T = \mathcal{A}(\mathcal{S})$ be the output of a symmetric algorithm, and let $\mathcal{S}^{(1)}$ denote the neighboring dataset obtained by replacing ξ_1 with an independent copy ξ_1'. If $\mathcal{A}$ is randomized, assume that each realization of its randomness yields a deterministic symmetric map. Then the generalization error satisfies $\epsilon_{\mathrm{gen}} \le \epsilon_{\mathcal{A}}$, where*

$$\epsilon_{\mathcal{A}} = \mathbb{E}\left[\frac{2\|\nabla f(\boldsymbol{x}^T;\xi_1)\|_2^2 + 2\|\nabla f(\boldsymbol{x}^T;\xi_1')\|_2^2}{N\,\beta^{\mathrm{pt}}_{\mathcal{S}^{(1)}}(\boldsymbol{x}^T)}\right] + \mathbb{E}\left[\frac{2L_f N\,\big(F_{\mathcal{S}}(\boldsymbol{x}^T) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\big)}{\beta^{\mathrm{pt}}_{\mathcal{S}^{(1)}}(\boldsymbol{x}^T)}\right].$$

The denominator $\beta^{\mathrm{pt}}_{\mathcal{S}^{(1)}}(\boldsymbol{x}^T)$ evaluates the neighboring ratio at the output trained on $\mathcal{S}$. The training PŁ condition alone gives no positive lower bound on it. Fixing RSD's sample-independent draws gives a deterministic symmetric map (Appendix G.2.3). Lemma 4.17 and Theorem C.5 give the risk estimate in Theorem D.7. The residual and inverse ratio share $\boldsymbol{x}^T$, so Hölder's inequality controls their product by separate moments without independence. For $p > 1$ and $q = p/(p-1)$, define

$$\mathfrak{G}_p(T) := \big\|\|\nabla f(\boldsymbol{x}^T;\xi_1)\|_2^2 + \|\nabla f(\boldsymbol{x}^T;\xi_1')\|_2^2\big\|_{L^p}, \qquad \mathfrak{I}_q(T) := \mathbb{E}\Big[\big(\beta^{\mathrm{pt}}_{\mathcal{S}^{(1)}}(\boldsymbol{x}^T)\big)^{-q}\Big], \tag{10}$$

The L^p norm measures sample-gradient size; $\mathfrak{I}_q(T)$ is a raw inverse moment sensitive to small neighboring ratios. Expectations include the data, replacement sample, and subspace draws.

Corollary 4.18 (Stopped excess risk under local PŁ)**.** *Suppose the assumptions of Theorem C.5 and Lemma 4.17 hold, with $L_f > 0$. Write the initial empirical residual as $R_0 := F_{\mathcal{S}}(\boldsymbol{x}^0) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})$ and the accumulated PŁ contraction as $\mathfrak{C}_{\mathcal{S}}(T) := 2\alpha\beta_{\mathcal{S}} \sum_{t<T} \eta_t(1 - L_{\max}\eta_t/2)$. Let $T = T(\mathcal{S}) \ge 1$ be an integer-valued, measurable, permutation-invariant budget, fixed conditional on $\mathcal{S}$ before the subspace draws, with $\mathfrak{G}_p(T) < \infty$ and $\mathfrak{I}_q(T) < \infty$. With $\log_+(u) := \log(u \vee 1)$, if, almost surely,*

$$\mathfrak{C}_{\mathcal{S}}(T) \ge p \log_+\big(N^2 L_f R_0\big), \tag{11}$$

then the excess risk satisfies

$$\mathbb{E}[F(\boldsymbol{x}^T)] - F(\boldsymbol{x}^\star) \le \frac{1}{N^2 L_f} + \frac{2}{N}\mathfrak{G}_p(T)\,\mathfrak{I}_q(T)^{1/q} + \frac{2}{N}\mathfrak{I}_q(T)^{1/q}. \tag{12}$$

For constant steps $\eta_t = 1/L_{\max}$, condition (11) reduces to $T \ge pL_{\max}/(\alpha\beta_{\mathcal{S}}) \cdot \log_+(N^2 L_f R_0)$. For uniformly sampled BCD with standard blockwise step sizes $\eta_t = 1/L_{i_t}$, the clock $\mathfrak{C}_{\mathcal{S}}(T)$ is replaced by $\beta^{\mathcal{L}}_{\mathcal{S}} T/m$ and the right-hand side of (12) is unchanged (Corollary D.10 in Appendix D).

Remark 4.19 (Stopping horizons and output moments). Condition (11) reduces the optimization term to $1/(N^2 L_f)$. The remaining terms scale with $(\mathfrak{G}_p(T)+1)\mathfrak{I}_q(T)^{1/q}/N$ in (12). The gradient moment captures sample-gradient size, and the inverse-curvature moment captures amplification at small neighboring ratios. An N-independent bound on $\sup_\alpha(\mathfrak{G}_p(T_\alpha)+1)\mathfrak{I}_q(T_\alpha)^{1/q}$ thus gives $O(1/N)$ uniformly over coverage. Coverage affects both the horizon and these output moments. Appendix D.2 gives lower-tail conditions allowing arbitrarily small ratios, and Appendix A.3 compares moments at matched αT. The budget factor p offsets the $1/p$ exponent from Hölder's inequality as shown in Lemma F.10 and the proof in Appendix G.2.5.

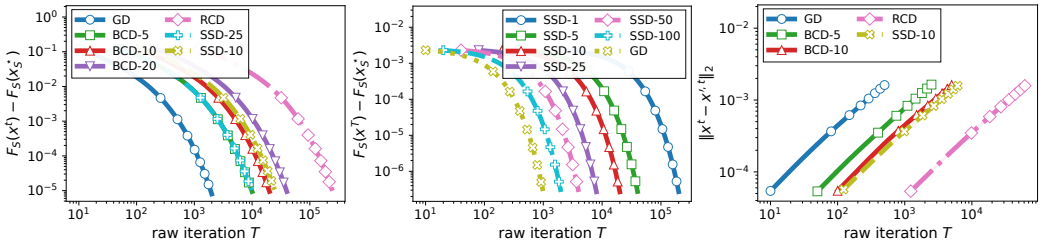
4.4 Practical interpretation

Sampling geometry and coverage. The expected projection operator determines coverage. Uniform sampling from a fixed, data-independent complete partition into m coordinate, layer, or module blocks gives $\alpha = 1/m$: each parameter has equal selection probability regardless of block size. An independently resampled isotropic right projection of rank r in the input space has coverage r/d_{in}. Standard LoRA learns two low-rank factors whose induced update is generally not an orthogonal projection, so its trainable parameter count alone does not determine coverage (Appendix E.2).

For isotropic SSD with fixed rank r in d dimensions, $\alpha = r/d$. Corollary 4.13 then prescribes horizons of order $(d/r)^{3/4}\sqrt{N}$ in the general convex case and $(d/r)^{1/2}N$ under low noise. These horizons give sufficient rank conditions for staying within a fixed iteration budget (Appendix D.1). Equal coverage makes the expected projection energy comparable across sampling schemes. In BCD with blockwise step sizes, the stability guarantee also depends on the curvature margin.

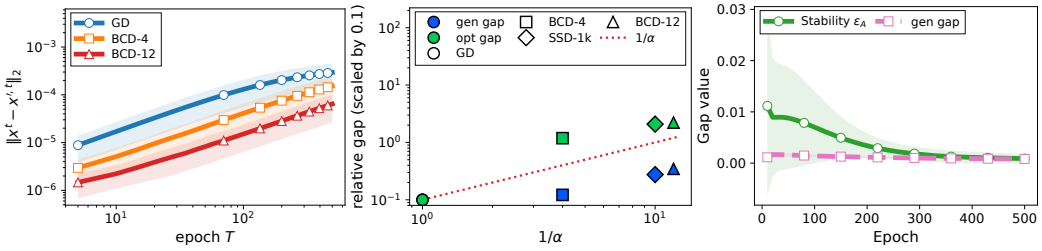
5 Experimental verification

We examine how coverage and horizon affect optimization, sensitivity to sample replacement, and generalization. We use regularized logistic and ridge regression for convex benchmarks and shallow networks with mean-squared loss for nonconvex benchmarks. GPT-2 fine-tuning tests longer horizons at reduced coverage. Appendix A gives details, further results, and reporting limits.



(a) ϵ_{opt} by α (b) ϵ_{opt} by r (c) Iterate discrepancy by α

Figure 2: Optimization error (a,b) and coupled-iterate discrepancy (c) on convex objectives, plotted against raw iterations T. Panels (a,c) vary coverage α; panel (b) varies subspace rank r.



(a) Iterate discrepancy by α (b) ϵ_{opt} versus ϵ_{gen} (c) $\epsilon_{\mathcal{A}}$ versus ϵ_{gen}

Figure 3: Nonconvex diagnostics. (a) Coupled-iterate discrepancy across coverages. (b) Relative optimization and generalization gaps. (c) Estimated stability quantity and generalization gap.

Effect of coverage α. We compare optimization gaps and coupled-iterate discrepancies across T and α. Figures 2 (a–c) and 3 (a) show faster convergence and faster growth of trajectory discrepancies at larger α. Appendices A.1.1 and A.2 report both views for other datasets and sampling schemes.

Optimization and generalization under reduced coverage. As α decreases, convergence is affected more strongly than generalization. In Figures 1 (a) and 3 (b), the optimization gap lies close to the α^{-1} guide, while the generalization gap grows more slowly. Figure 1 (a) also displays an $\alpha^{-1/4}$ guide. These lines are visual references, not fitted exponents. Figure 1 (b) is a qualitative cost illustration; Appendix A records the missing normalization and reference-risk definitions.

Estimated stability and generalization gaps. In Figure 3 (c), the estimated stability quantity $\epsilon_{\mathcal{A}}$ lies above the empirical generalization gap along training and follows a similar trend. Appendix A.2 reports the same comparison for GD and BCD on two datasets, together with gradient norms.

GPT-2 fine-tuning. We fine-tune the GPT-2 Small model [45] on WikiText-103 [39], uniformly sampling a subset of its 12 Transformer blocks for AdamW updates while freezing the embeddings, output head, and final layer normalization. At 20,000 steps, perplexity worsens as coverage decreases. Across coverage levels, all Transformer blocks remain eligible for updates. Reduced coverage therefore changes how often parameters are updated, without restricting which parameters can be trained over the full run. Extending training to $T_\alpha = 20{,}000/\alpha$ matches expected updates per block, narrowing perplexity gaps to 0.09, 0.31, and 1.05 relative to updating all 12 blocks (Table 2). The extended runs process more data, and AdamW differs from the full-gradient update analyzed here. The comparison illustrates the coverage–horizon relation; Appendix A.2 describes the setup and reporting limits for these runs.

Table 2: GPT-2 Small on WikiText-103: perplexity by coverage and horizon.

α	T_α	**Perplexity** ↓	
		20,000 steps	T_α steps
1	20,000	27.91	27.91
1/2	40,000	29.21	28.00
1/4	80,000	32.06	28.22
1/12	240,000	37.84	28.96

Output-dependent moments. A realizable tanh teacher–student study ($N = 64$, 16 hidden neurons, 24 datasets) examines how coverage affects moments (Remark 4.19). At $T = 40$, test excess rises from 0.0185 to 0.1015 as coverage falls from 1 to 1/8. At $\alpha T = 40$, it remains approximately in $[0.0185, 0.0187]$, while $(\widehat{\mathfrak{G}}_2 + 1)\widehat{\mathfrak{I}}_2^{1/2}$ ranges from 64.0 to 71.8 without monotone growth. Across four coverages, paired comparisons detect small differences, while a few runs strongly influence inverse-moment estimates (Appendix A.3). Small neighboring pointwise ratios contribute more to inverse moments, making the estimated stability contribution sensitive to a few datasets despite small mean test excess.

Curvature balance for convex BCD. Across regularization levels in the a9a logistic-regression sweep, $L_f < 2L_{\min}$ holds for all 100 sampled balanced partitions into 2 groups, for 65–100% with 4, for 0–18% with 8, and for none with 16 (Appendix A.1.2). Finer partitions pass less often because every block smoothness constant must exceed half the global bound. The partition thus affects both computation and the stability guarantee. Reducing block size changes coverage and the curvature margin governing the bound on sensitivity to sample replacement.

6 Conclusion

RSD unifies the coverage analysis of GD, RCD, BCD, and isotropic subspace descent. The convex excess-risk bound is uniform in sample size and coverage under the stated assumptions, with a fractional coverage cost despite the $1/\alpha$ optimization slowdown. For BCD with blockwise step sizes, the curvature-weighted analysis makes the role of the curvature margin explicit in both the stability and excess-risk bounds. In the nonconvex local PŁ setting, coverage acts through the horizon and the sample-gradient and inverse-curvature moments in the risk bound.

Limitations and future work. The theory assumes uniform smoothness, fixed joint isotropy, and full gradients. The nonconvex bounds also require local PŁ, output-moment control, and a horizon fixed before subspace sampling. Adaptive methods and validation-based stopping rules require further theoretical analysis. Sharper curvature certificates and high-probability guarantees are natural extensions. Appendix A details reporting gaps.

Acknowledgments and Disclosure of Funding

This work was supported by the National Key R&D Program of China (2024YFC3307100), the National Natural Science Foundation of China (62576364 and 62625602), the Guangdong Basic and Applied Basic Research Foundation (2026B1515020071), the Shenzhen Science and Technology Program (JCYJ20241202124430041 and SYSRD20250529113401002), the CCF-Tencent Rhino-Bird Open Research Fund (CCF-Tencent RAGR20250114), Tencent (JR2025TEG002), and the CAAI-Tencent Rhino-Bird Open Research Fund (2026CAAI-Tencent-08).

References

[1] Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 7319–7328, 2021.

[2] Shunta Akiyama. Block coordinate descent for neural networks provably finds global minima. *arXiv preprint arXiv:2510.22667*, 2025.

[3] Peter L Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of machine learning research*, 3(Nov):463–482, 2002.

[4] Amir Beck and Luba Tetruashvili. On the convergence of block coordinate descent type methods. *SIAM journal on Optimization*, 23(4):2037–2060, 2013.

[5] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989.

[6] Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

[7] Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, pages 610–626. PMLR, 2020.

[8] Long Chen, Xiaozhe Hu, and Steven Wise. Convergence analysis of the fast subspace descent method for convex optimization problems. *Mathematics of Computation*, 89(325):2249–2282, 2020.

[9] Long Chen, Xiaozhe Hu, and Huiwen Wu. Randomized fast subspace descent methods. *arXiv preprint arXiv:2006.06589*, 2020.

[10] Yiming Chen, Yuan Zhang, Yin Liu, Kun Yuan, and Zaiwen Wen. A memory efficient randomized subspace optimization method for training large language models. In *International Conference on Machine Learning*, volume 267, pages 9747–9763. PMLR, 2025.

[11] Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

[12] André Elisseeff, Theodoros Evgeniou, and Massimiliano Pontil. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6:55–79, 2005.

[13] Vitaly Feldman and Jan Vondrak. Generalization bounds for uniformly stable algorithms. *Advances in Neural Information Processing Systems*, 31, 2018.

[14] Vitaly Feldman and Jan Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on learning theory*, pages 1270–1279. PMLR, 2019.

[15] Rafael Frongillo and Mark D Reid. Convergence analysis of prediction markets via randomized subspace descent. *Advances in Neural Information Processing Systems*, 28, 2015.

[16] Robert Gower, Dmitry Kovalev, Felix Lieder, and Peter Richtárik. Rsn: randomized subspace newton. *Advances in Neural Information Processing Systems*, 32, 2019.

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

[18] Yutong He, Pengrui Li, Yipeng Hu, Chuyan Chen, and Kun Yuan. Subspace optimization for large language models with convergence guarantees. In *International Conference on Machine Learning*, volume 267, pages 22468–22522. PMLR, 2025.

[19] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[20] Zhanliang Huang, Yunwen Lei, and Ata Kabán. Optimization and learning with randomly compressed gradient updates. *Neural Computation*, 35(7):1234–1287, 2023.

[21] Li-Li Jin and Hou-Biao Li. Greedy double subspaces coordinate descent method via orthogonalization. *arXiv preprint arXiv:2203.02153*, 2022.

[22] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak–Łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 795–811. Springer, 2016.

[23] David Kozak, Stephen Becker, Alireza Doostan, and Luis Tenorio. Stochastic subspace descent. *arXiv preprint arXiv:1904.01145*, 2019.

[24] Ilja Kuzborskij and Christoph Lampert. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 2815–2824. PMLR, 2018.

[25] Yunwen Lei, Tao Sun, and Mingrui Liu. Minibatch and local SGD: Algorithmic stability and linear speedup in generalization. *Applied and Computational Harmonic Analysis*, 79:101795, 2025.

[26] Yunwen Lei, Zimeng Wang, and Xiaoming Yuan. From convergence to generalization: Stability of stationary-point learning algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(9):10598–10609, 2026.

[27] Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, pages 5809–5819. PMLR, 2020.

[28] Qinglun Li, Yingqi Liu, Miao Zhang, Xiaochun Cao, Quanjun Yin, and Li Shen. Unveiling the power of multiple gossip steps: A stability-based generalization analysis in decentralized training. In *Advances in Neural Information Processing Systems*, volume 38, pages 73808–73827, 2025.

[29] Xingguo Li, Tuo Zhao, Raman Arora, Han Liu, and Mingyi Hong. On faster convergence of cyclic block coordinate descent-type methods for strongly convex minimization. *Journal of Machine Learning Research*, 18(184):1–24, 2018.

[30] Kaizhao Liang, Bo Liu, Lizhang Chen, and Qiang Liu. Memory-efficient LLM training with online subspace descent. *Advances in Neural Information Processing Systems*, 37:64412–64432, 2024.

[31] Junkang Liu, Fanhua Shang, Yuanyuan Liu, Hongying Liu, Yuangang Li, and YunXiang Gong. Fedbcgd: Communication-efficient accelerated block coordinate gradient descent for federated learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2955–2963, 2024.

[32] Yingqi Liu, Qinglun Li, Jie Tan, Yifan Shi, Li Shen, and Xiaochun Cao. Understanding the stability-based generalization of personalized federated learning. In *International Conference on Learning Representations*, 2025.

[33] Yuxi Liu, Renjia Deng, Yutong He, Xue Wang, Tao Yao, and Kun Yuan. Misa: Memory-efficient LLMs optimization with module-wise importance sampling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[34] Zeyu Liu, Yan Li, Yunquan Zhang, Boyang Zhang, Guoyong Jiang, Xin Zhang, Limin Xiao, Weifeng Zhang, and Daning Cheng. Exploiting block coordinate descent for cost-effective LLM model training. *arXiv preprint arXiv:2506.12037*, 2025.

[35] Jackie Lok and Elizaveta Rebrova. Subspace-constrained randomized coordinate descent for linear systems with good low-rank matrix approximations. *arXiv preprint arXiv:2506.09394*, 2025.

[36] Qijun Luo, Yifei Shen, Liangzu Peng, Dongsheng Li, and Xiao Li. Accelerating block coordinate descent for LLM finetuning via landscape expansion. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[37] Qijun Luo, Hengxu Yu, and Xiao Li. Badam: A memory efficient full parameter optimization method for large language models. *Advances in Neural Information Processing Systems*, 37:24926–24958, 2024.

[38] Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1):157–178, 1993.

[39] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

[40] Yu Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

[41] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[42] Julie Nutini, Issam Laradji, and Mark Schmidt. Let's make block coordinate descent converge faster: faster greedy rules, message-passing, active-set complexity, and superlinear convergence. *Journal of Machine Learning Research*, 23(131):1–74, 2022.

[43] Gaku Omiya, Pierre-Louis Poirion, and Akiko Takeda. Convergence analysis of randomized subspace normalized SGD under heavy-tailed noise. *arXiv preprint arXiv:2601.20399*, 2026.

[44] Rui Pan, Xiang Liu, Shizhe Diao, Renjie Pi, Jipeng Zhang, Chi Han, and Tong Zhang. Lisa: Layerwise importance sampling for memory-efficient large language model fine-tuning. *Advances in Neural Information Processing Systems*, 37:57018–57049, 2024.

[45] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.

[46] Amrutha Varshini Ramesh, Vignesh Ganapathiraman, Issam H Laradji, and Mark Schmidt. Blockllm: Memory-efficient adaptation of LLMs by selecting and optimizing the right coordinate blocks. In *OPT 2024: Optimization for Machine Learning*, 2024.

[47] Peter Richtárik and Martin Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1):1–38, 2014.

[48] Gili Rosenberg, J Kyle Brubaker, Martin JA Schuetz, Elton Yechao Zhu, Serdar Kadıoğlu, Sima E Borujeni, and Helmut G Katzgraber. Scalable iterative pruning of large language and vision models using block coordinate descent. *arXiv preprint arXiv:2411.17796*, 2024.

[49] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

[50] Yan Sun, Li Shen, and Dacheng Tao. Toward understanding generalization and stability gaps between centralized and decentralized federated learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(4):4744–4755, 2026.

[51] Ty Trusty. Coordinate condensation: Subspace-accelerated coordinate descent for physics-based simulation. *arXiv preprint arXiv:2510.12053*, 2025.

[52] Paul Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494, 2001.

[53] Peng Wang, Li Shen, Zerui Tao, Shuaida He, and Dacheng Tao. Generalization analysis of stochastic weight averaging with general sampling. In *International Conference on Machine Learning*, volume 235, pages 51442–51464. PMLR, 2024.

[54] Puyu Wang, Liang Wu, and Yunwen Lei. Stability and generalization for randomized coordinate descent. In *International Joint Conference on Artificial Intelligence*, pages 3104–3110, 2021.

[55] Stephen J Wright. Coordinate descent algorithms. *Mathematical programming*, 151(1):3–34, 2015.

[56] Liang Wu, Ruixi Hu, and Yunwen Lei. Stability-based generalization analysis of randomized coordinate descent for pairwise learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(20):21545–21553, 2025.

[57] Yangyang Xu and Wotao Yin. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on imaging sciences*, 6(3):1758–1789, 2013.

[58] Jinshan Zeng, Tim Tsz-Kit Lau, Shaobo Lin, and Yuan Yao. Global convergence of block coordinate descent in deep learning. In *International conference on machine learning*, pages 7313–7323. PMLR, 2019.

[59] Haochen Zhang, Junze Yin, Guanchu Wang, Zirui Liu, Lin F Yang, Tianyi Zhang, Anshumali Shrivastava, and Vladimir Braverman. Breaking the frozen subspace: Importance sampling for low-rank optimization in LLM pretraining. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[60] Zhong Zhang, Bang Liu, and Junming Shao. Fine-tuning happens in tiny subspaces: Exploring intrinsic task-specific subspaces of pre-trained language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1701–1713, 2023.

[61] Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. GaLore: Memory-efficient LLM training by gradient low-rank projection. In *International Conference on Machine Learning*, volume 235, pages 61121–61143. PMLR, 2024.

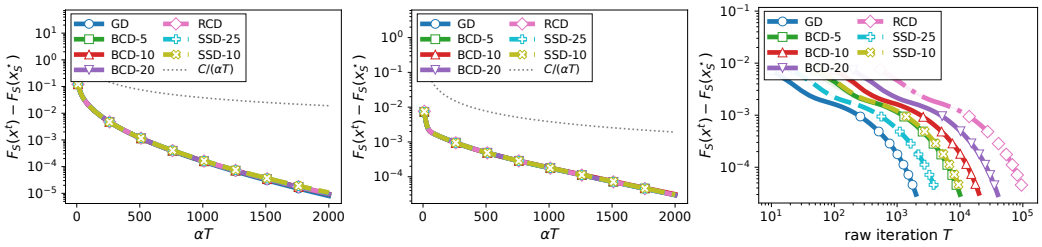
A Additional experiments

We examine how coverage, step size, block curvature, and noise affect optimization and stability, then study the output moments in the nonconvex risk bound. The benchmark plots provide qualitative comparisons: their generating code, run-level records, reference-solution definitions, and uncertainty-band metadata are unavailable. Normalized oracle cost in Figure 1 is distinct from measured runtime; its exact normalization and the displayed reference risks could not be reconstructed. These records therefore do not support estimates of scaling exponents or statistical significance.

A.1 Convex experiments

A.1.1 Coverage, training horizons, and sensitivity

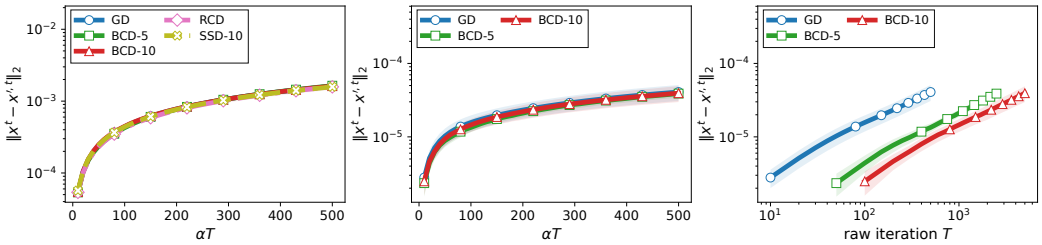
We compare optimization against the coverage-adjusted clock αT and the raw iteration count T. The mean projector has trace αd, so α is the expected sampled dimension divided by d.



(a) a9a (by αT) (b) covtype (by αT) (c) covtype (by T)

Figure 4: Optimization error on raw and coverage-adjusted time axes.

Rescaling by αT brings several curves closer in Figure 4, consistent with coverage entering the expected descent rate. The realized updates still depend on the sampled subspaces, so joint isotropy alone does not imply identical trajectories or exact time rescaling. To relate this optimization effect to stability, we compare the propagation of single-sample perturbations across coverage levels α.



(a) a9a (by αT) (b) covtype (by αT) (c) covtype (by T)

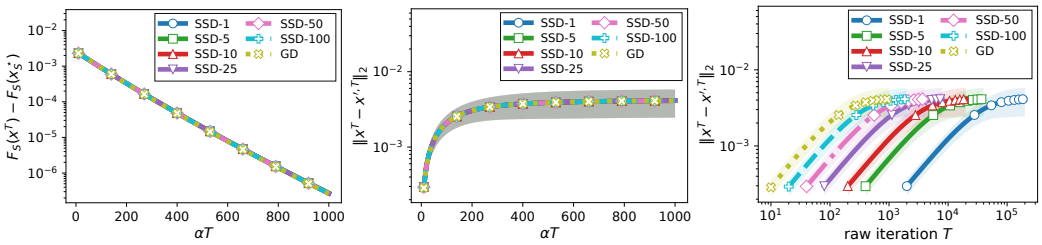
Figure 5: Neighboring-data trajectory discrepancy across coverages.

In Figure 5, the two clocks distinguish per-step perturbation growth from sensitivity at equal coverage-adjusted horizons. For isotropic SSD, $\alpha = r/d$ links subspace dimension to coverage. Varying r at fixed ambient dimension compares optimization and sensitivity to a single-sample replacement from rank-one projections to full-gradient updates, using both T and αT.

In Figure 6, rescaling by αT brings both suboptimality and trajectory discrepancy closer across SSD dimensions. On the raw-time axis, discrepancies grow faster at larger ranks.

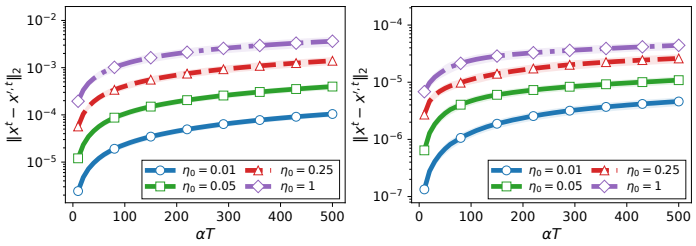
A.1.2 Step sizes and block curvature

We vary a common step size, then examine blockwise step sizes. Theorem 4.7 and Corollary 4.9 link these choices to perturbation growth and block curvature.



(a) Optimization (by αT) (b) Stability (by αT) (c) Stability (by T)

Figure 6: SSD suboptimality and trajectory discrepancy across subspace dimensions.

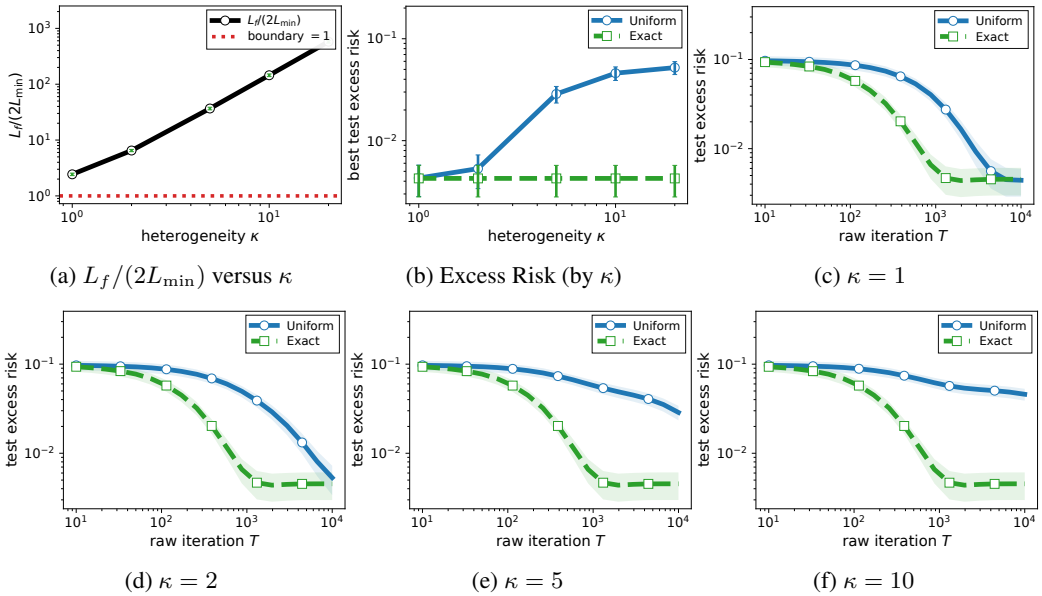


(a) a9a (by αT) (b) covtype (by αT)

Figure 7: Neighboring-data trajectory discrepancy across step sizes for BCD-10.

For BCD-10 at $\eta_0 \in \{0.01, 0.05, 0.25, 1.0\}$, Figure 7 shows larger trajectory discrepancies at larger step sizes on both datasets, indicating greater sensitivity to sample replacement.

We study the canonical blockwise rule $\eta_t = 1/L_{i_t}$. It adapts the update scale to each block's curvature, and the corresponding guarantee uses $D_{\mathcal{L}}$ and the margin $2L_{\min} - L_f$. Heterogeneous block curvatures can worsen the bound even when the method performs well empirically.



(a) $L_f/(2L_{\min})$ versus κ (b) Excess Risk (by κ) (c) $\kappa = 1$

(d) $\kappa = 2$ (e) $\kappa = 5$ (f) $\kappa = 10$

Figure 8: Excess risk across block-heterogeneity settings κ.

Figure 8 compares common (Uniform) and blockwise (Exact) step sizes on ten 10-dimensional standard Gaussian feature blocks, with the last five scaled by κ and no feature normalization. Stable empirical behavior may extend beyond the sufficient-condition boundary $L_f/(2L_{\min}) = 1$.

To examine the same condition on a9a, we use unit-normalized nonzero features, 4,096 training observations, and 1,024 replacement observations. Each regularization value and block count is

evaluated on 100 random balanced partitions to estimate the condition's empirical pass frequency. These 1,600 configurations are certificate calculations, with a pass defined by $L_f/(2L_{\min}) < 1$.

For regularized logistic loss $\log(1 + \exp(-y\boldsymbol{x}^\top \boldsymbol{w})) + \lambda\|\boldsymbol{w}\|_2^2/2$, valid smoothness constants on a finite feature pool $\mathcal{P}$ are

$$L_f = \lambda + \tfrac{1}{4}\max_{\boldsymbol{x}\in\mathcal{P}} \|\boldsymbol{x}\|_2^2, \qquad L_i = \lambda + \tfrac{1}{4}\max_{\boldsymbol{x}\in\mathcal{P}} \|U_i^\top \boldsymbol{x}\|_2^2,$$

with binary labels $y \in \{-1, 1\}$. The logistic second derivative is at most $1/4$, so these bounds hold at every parameter value for observations in $\mathcal{P}$. Population guarantees require corresponding bounds over the population support. These are valid analytic constants; the constants used to generate Table 3 cannot be checked from the retained records.

Table 3: a9a curvature certificate over 100 partitions per setting. Each block-count group reports the pass percentage and median ratio $L_f/(2L_{\min})$. Ratios are rounded to three decimals.

	2 blocks		4 blocks		8 blocks		16 blocks	
λ	Pass (%)	Ratio	Pass (%)	Ratio	Pass (%)	Ratio	Pass (%)	Ratio
10^{-4}	100	0.636	65	1.000	0	1.748	0	3.243
10^{-3}	100	0.636	71	0.996	0	1.733	0	3.180
10^{-2}	100	0.630	72	0.963	0	1.596	0	2.516
10^{-1}	100	0.590	100	0.778	18	1.021	0	1.264

Finer partitions satisfy the sufficient condition less often, whereas stronger regularization improves the pass frequency across the sampled partitions. Independent reconstruction requires the original partition seeds, unrounded ratios, and executed constant-estimation code, which are unavailable.

A.1.3 Low-noise regime

We compare test excess across sample sizes under high- and low-noise settings, motivated by the two regimes in Corollary 4.13. The corollary's low-noise condition is $F(\boldsymbol{x}^\star) = O(N^{-1})$. Through the self-bounding property, smaller risk reduces the gradient contribution to stability and permits a longer stopping horizon. For fixed coverage and problem constants, the resulting bound scales as N^{-1} rather than $N^{-1/2}$.

Figure 9 shows the comparison for GD and BCD-10. Lower plotted test excess under low noise is consistent with the sharper bound. The original noise-generation settings and uncertainty metadata were not retained, so the curves provide a qualitative comparison rather than an estimate of the sample-size scaling exponent in either noise regime.

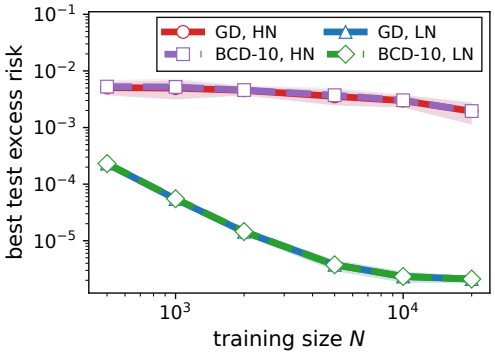
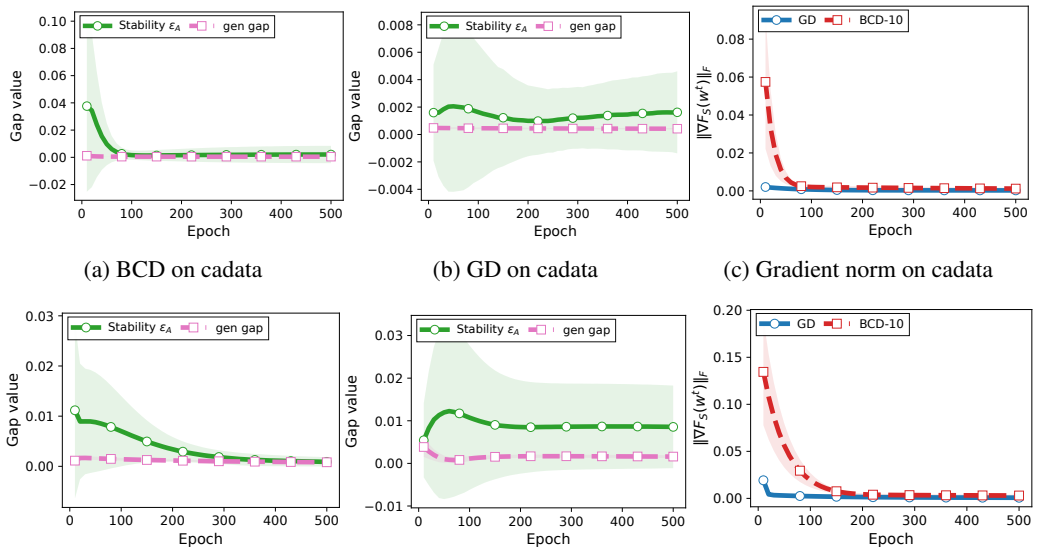


Figure 9: Test excess under high and low noise.

A.2 Nonconvex experiments

For overparameterized shallow neural networks, Figure 10 compares the estimated stability quantity $\epsilon_{\mathcal{A}}$ with the empirical train–test gap and separately plots the gradient norm. The decreasing gradient norms indicate progress toward stationarity. The population-risk bound also involves the pointwise ratio on a neighboring dataset, evaluated at the output trained on the original sample. Small training gradients alone do not control this ratio or the stability contribution from its inverse moments. The plotted $\epsilon_{\mathcal{A}}$ is computed from its definition; the presentation follows Lei et al. [26]. Appendix A.3 examines the neighboring ratios and inverse moments directly using run-level data.



(a) BCD on cadata (b) GD on cadata (c) Gradient norm on cadata

(d) GD on cpusmall (e) BCD on cpusmall (f) Gradient norm on cpusmall

Figure 10: Nonconvex stability diagnostics on cadata and cpusmall: estimated stability quantity, generalization gap, and gradient norm for GD and BCD.

Figure 11 compares distances between coupled GD and BCD trajectories against αT. Figure 3(a) shows the comparison at raw T. The two views show how sensitivity varies with the budget. These distances measure parameter changes after replacing one observation, whereas $\epsilon_{\mathcal{A}}$ in Lemma 4.17 bounds the average loss change. Smoothness alone gives no uniform ordering across coverages. Neighboring curvature and output moments also enter the nonconvex bound. Appendix A.3 examines both quantities at the stopped outputs.

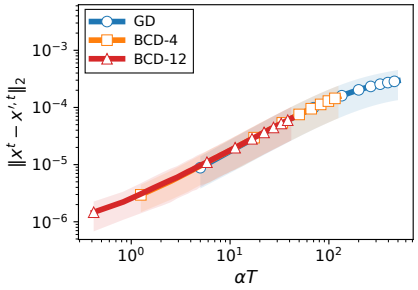


Figure 11: Iterate discrepancy against αT.

Block-sparse GPT-2 fine-tuning. Table 2 compares perplexity when fewer Transformer blocks are updated at each step. The GPT-2 Small/WikiText-103 experiment uses AdamW, sequence length 512, effective batch size 64, and learning rate 5×10^{-5}. Embeddings, the language-model head, and final layer normalization are frozen. A uniform subset of 12, 6, 3, or 1 of the twelve Transformer blocks is updated per step; coverage refers to this eligible parameter space.

At $T_\alpha = 20000/\alpha$, the expected number of updates per eligible block matches full coverage, and the perplexity gaps narrow. Smaller coverage processes more minibatches and tokens. Step times relative to full coverage are 1.00, 0.86, 0.78, and 0.73. Multiplying by the relative horizons gives 1.00, 1.72, 3.12, and 8.76: the lower per-step cost is offset by longer training. These products are cost extrapolations, rather than measured total runtimes or matched-compute comparisons.

A.3 Output-moment diagnostics

The following experiments examine how coverage and training time affect the output moments in Corollary 4.18. A realizable tanh model gives exact neighboring residuals; least squares provides an exact-curvature reference, and a sigmoid network illustrates an inverse-ratio surrogate.

Controlled tanh comparison. We optimize $W \in \mathbb{R}^{16\times 4}$ in $h_W(x) = v^\top \tanh(Wx)$ under half-squared loss, with x denoting the input. A unit-normalized Gaussian vector v and a teacher matrix with independent $\mathcal{N}(0, 0.8^2)$ entries are fixed across runs. Inputs are uniform on $[-1, 1]^4$, with noiseless teacher labels, so the population and neighboring empirical minima are zero.

Each of 24 outer runs draws 64 training and 4,096 test observations. Initialization is the teacher plus independent $\mathcal{N}(0, 0.9^2)$ noise; data and initialization are shared across coverages. For $\alpha \in \{1, 1/2, 1/4, 1/8\}$, the neurons form $1/\alpha$ equal contiguous blocks. Each full-batch update selects one block uniformly with step size 0.5. Outputs are evaluated at $\alpha T \in \{5, 10, 20, 40\}$ and, separately, at $T = 40$. Each iteration computes the full gradient before applying the block mask, so no computational savings are measured. The step size and budgets are chosen for moment comparisons; they have not been certified against the sufficient conditions of Corollary 4.18.

At each output, we evaluate 24 independent one-sample replacements with uniformly chosen indices, without retraining. Following Eq. (10), $\widehat{\mathfrak{G}}_2$ is the root mean square of the sum of two squared sample-gradient norms, and $\widehat{\mathfrak{I}}_2^{1/2}$ is the root mean square of the inverse pointwise ratio, pooled over neighbors and runs. Realizability gives exact neighboring residuals. The screening rule excludes losses at most 10^{-12}, NaN ratios, or nonpositive ratios, without clipping. None of the 11,520 evaluations across 480 outputs was flagged; the 24 outer runs are the independent sampling units.

Matching αT removes most of the increase in test excess seen at fixed T (Table 4). Here test excess is mean test loss, and the generalization gap subtracts training loss. Marginal 95% percentile intervals use 2,000 outer-run bootstrap resamples; fixed-T intervals were not retained.

Table 4: Tanh estimates and marginal 95% confidence intervals in the displayed units.

	Fixed $T = 40$	Matched $\alpha T = 40$			
α	Test excess (10^{-2})	Test excess (10^{-2})	95% CI	Gap (10^{-3})	95% CI
1	1.846	1.846	[1.599, 2.109]	3.217	[2.219, 4.271]
1/2	3.697	1.869	[1.603, 2.139]	3.219	[2.208, 4.218]
1/4	6.783	1.857	[1.612, 2.124]	3.207	[2.194, 4.199]
1/8	10.151	1.872	[1.602, 2.158]	3.205	[2.159, 4.288]

The output moments show a similar pattern at $\alpha T = 40$: the factor $(\widehat{\mathfrak{G}}_2 + 1)\widehat{\mathfrak{I}}_2^{1/2}$ ranges from 64.01 to 71.76, without monotone growth as coverage decreases (Table 5). Multiplication by $2/N$ gives 2.00–2.24, far above the observed test excess.

Table 5: Tanh output moments at $\alpha T = 40$, with marginal 95% confidence intervals.

	$10^2\widehat{\mathfrak{G}}_2$		$\widehat{\mathfrak{I}}_2^{1/2}$		Moment factor	
α	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
1	4.834	[4.019, 5.721]	65.21	[48.53, 82.72]	68.37	[50.78, 86.94]
1/2	5.321	[4.460, 6.239]	64.57	[48.51, 83.33]	68.01	[50.98, 87.77]
1/4	5.173	[4.390, 5.965]	60.86	[45.90, 77.85]	64.01	[48.40, 81.76]
1/8	5.493	[4.432, 6.563]	68.02	[47.11, 91.15]	71.76	[49.63, 96.98]

Paired comparisons reveal small differences within this overall agreement. At $\alpha = 1/8$, test excess exceeds full coverage by 2.585×10^{-4}, with interval $[0.152, 5.062] \times 10^{-4}$. At $\alpha = 1/4$, the moment-factor difference is -4.358, with interval $[-6.696, -1.907]$. Paired percentile intervals use 10,000 resamples and seed 20261025, resampling matched outer runs and recomputing nonlinear moments. These intervals are unadjusted for multiple comparisons; overlapping marginal intervals in Tables 4–5 are not evidence of equivalence.

The inverse-moment estimates are sensitive to individual datasets. Run 23 contributes 27.2–38.0% of the squared inverse-ratio sums across coverages. At $\alpha = 1/8$, omitting it changes the factor from 71.76 to 57.45; all runs remain in the primary estimates. Inverse moments amplify small pointwise ratios, so a few datasets can strongly affect this factor even when mean test excess changes little.

Least squares. The rank-40 study provides a comparison with exact curvature, using $N = 512$, ambient dimension 200, and eight coordinate groups. Its horizons satisfy $T_\alpha/T_1 = 1/\alpha$; absolute horizons were not retained, so the stopping condition cannot be checked. Across $\alpha = 1, 1/2, 1/4, 1/8$, the point estimates are $\widehat{\mathfrak{G}}_2 = 0.0020677$ and $\widehat{\mathfrak{I}}_2^{1/2} = 41.6859$, with nearly overlapping cluster-bootstrap intervals. The moment contribution is 0.163176, well above the population excess 3.34472×10^{-5}. Full summaries and intervals are in the supplementary CSV.

Sigmoid network. Nonnegative loss provides an inverse-ratio surrogate when the neighboring optimum is unknown. The study uses $N = 256$, input dimension 10, widths $10, 20, 40$, 3,000 steps, step size 0.5, and ten runs with twenty neighbors each. For $F_{\mathcal{S}^{(1)}}(\boldsymbol{x}) > 0$,

$$\frac{\|\nabla F_{\mathcal{S}^{(1)}}(\boldsymbol{x})\|_2^2}{2F_{\mathcal{S}^{(1)}}(\boldsymbol{x})} \leq \beta^{\mathrm{pt}}_{\mathcal{S}^{(1)}}(\boldsymbol{x}),$$

and taking inverses gives the surrogate. Root inverse-moment surrogates range from 1.50×10^6 to 3.53×10^6, and gradient moments from 1.80×10^{-5} to 2.76×10^{-5}. Width-specific values are in the supplementary CSV. All sampled inverses were finite; integrability and uniform coverage control require population lower-tail bounds.

Implementation details and records. The supplementary tanh code transcription uses float64 and base seed 20260728, with run-level data, seed mappings, screening rules, and influence summaries. Re-execution reproduced saved outputs to floating-point precision and paired reanalysis reproduced the summaries. Original hardware and package versions were not recorded. Least squares and sigmoid retain aggregate summaries and partial run records. For GPT-2, exact model/data revisions, evaluation split, loss aggregation, seed counts, scheduler, inactive optimizer-state behavior, hardware, and timing protocol were not retained.

B Additional related work

Convergence Theory for Coordinate and Block Coordinate Descent. Early error-bound analyses for coordinate descent can be traced to Luo and Tseng [38]. Tseng [52] extended BCD convergence theory to nondifferentiable objectives. Beck and Tetruashvili [4] established sublinear and linear rates under convexity and strong convexity, respectively. Xu and Yin [57] studied multiconvex problems using the Kurdyka–Łojasiewicz property, while Li et al. [29] tightened the convergence gap for cyclic BCD. Zeng et al. [58] proved global convergence of BCD in specific deep-learning settings, and Nutini et al. [42] studied greedy and superlinear block-selection rules.

Subspace Descent Methods. Frongillo and Reid [15] connected randomized subspace descent to prediction-market convergence. Gower et al. [16] proposed a randomized subspace Newton method with linear convergence, and Chen et al. [8] analyzed fast subspace descent for convex problems. Other variants include greedy double-subspace methods [21], subspace-constrained linear solvers [35], and randomized subspace normalized SGD under heavy-tailed noise [43].

Memory-Efficient LLM Training via Subspace and Block Updates. LISA [44] and MISA [33] sample layers or modules to reduce training memory. Zhang et al. [59] instead use importance sampling to select low-rank gradient subspaces during LLM pretraining. Block-coordinate families targeting full-parameter tuning on resource-constrained GPUs include BAdam [37], BlockLLM [46], landscape-expansion-accelerated BCD [36], and the BCD exploitation strategy of Liu et al. [34]. Similar block or subspace ideas have also appeared in federated learning [31], scalable iterative pruning [48], and physics simulation [51].

Stability and Generalization Literature. Elisseeff et al. [12] extended the uniform-stability framework of Bousquet and Elisseeff [6] to randomized algorithms. Shalev-Shwartz et al. [49] showed that stability characterizes learnability and can apply beyond uniform-convergence regimes. Beyond the foundational SGD stability analysis of Hardt et al. [17], subsequent refinements include data-dependent bounds [24], high-probability bounds [13, 14], sharper bounds for uniformly stable algorithms [7], and mini-batch and local SGD analyses [25].

Stability analyses also cover stochastic weight averaging under general sampling [53], partial model personalization in federated learning [32], multiple gossip steps in decentralized training [28], and centralized–decentralized federated learning comparisons [50]. Huang et al. [20] analyzed Gaussian-sketch updates projected onto compressed feasible sets and lifted back, including full-batch gradient

descent. Their convex risk bounds concern averaged outputs; we study last iterates of isotropic orthogonal-projector updates and their dependence on coverage. Complexity-based generalization analyses relying on measures such as VC dimension [5], Rademacher complexity [3], and PAC-Bayes arguments [11] provide alternative routes to generalization bounds. For RSD, directly controlling these complexity terms can be impractical. Our stability analysis tracks how coverage shapes the propagation of single-sample perturbations.

C Optimization error bounds

C.1 Convex case

Theorem C.1 (Optimization error in the convex case)**.** *Suppose Definition 4.6, Assumptions 4.1–4.4, and the standing structural conditions (3)–(5) hold. Let the step sizes satisfy* $0 < \eta_{t+1} \le \eta_t \le \frac{1}{L_{\max}}, t \ge 0$. *Then, for every* $T \ge 1$, *the output of Algorithm 1 satisfies*

$$\epsilon_{\mathrm{opt}} := \mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^T) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\big] \le \frac{(1+\eta_0 L_f)\,\mathbb{E}\big[\|\boldsymbol{x}^0 - \boldsymbol{x}^\star_{\mathcal{S}}\|_2^2\big]}{2\alpha\sum_{t=0}^{T-1}\eta_t}.$$

Remark C.2 (Subspace coverage as the optimization clock). With constant admissible step sizes, Theorem C.1 gives the effective-clock rate $O(1/(\alpha T))$: reduced coverage slows optimization by $1/\alpha$. The choices $\alpha = 1$, $\alpha = 1/d$, and $\alpha = 1/m$ recover the standard rates for GD, RCD, and uniform m-block BCD, respectively [41, 40, 47]. Joint isotropy expresses this classical optimization cost through a common coverage parameter, providing the baseline for comparison with the stability and excess-risk bounds. See Appendix G.1.1 for the proof. The next bound uses blockwise steps and matches the descent potential to each block's smoothness constant and update scale.

Corollary C.3 (Optimization under BCD with blockwise step sizes)**.** *Suppose Definition 4.6, Assumptions 4.1–4.4, and Eq. (3) hold. Consider the BCD specialization of Algorithm 1, namely* $P_i = U_i$, $\rho_i = 1/m$, *and* $\sum_{i=1}^m U_i U_i^\top = I_d$. *Suppose that, at iteration* t, *the sampled block* i_t *is updated with the blockwise step size* $\eta_t = \frac{1}{L_{i_t}}$. *Then, for every* $T \ge 1$,

$$\mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^T) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\big] \le \frac{m}{T}\left(\mathbb{E}\big[D_{\mathcal{L}}(\boldsymbol{x}^0, \boldsymbol{x}^\star_{\mathcal{S}})\big] + \frac{L_f}{2}\,\mathbb{E}\big[\|\boldsymbol{x}^0 - \boldsymbol{x}^\star_{\mathcal{S}}\|_2^2\big]\right).$$

Remark C.4 (Block Lyapunov geometry for blockwise step sizes). In Corollary C.3, the steps $1/L_{i_t}$ lead to the potential $D_{\mathcal{L}}(\boldsymbol{x}^0, \boldsymbol{x}^\star_{\mathcal{S}})$, together with the Euclidean initial-distance term weighted by $L_f/2$, without requiring curvature balance. The potential weights the initial error by each block's smoothness constant, so its size depends on where the error lies. The fixed orthogonal decomposition $\{U_i U_i^\top\}_{i=1}^m$ lets the weighted recursion telescope: each update changes one component, and the L_i-weight cancels the step size $1/L_i$ in the Lyapunov recursion. Random or overlapping subspaces generally lack this decomposition, so the same cancellation is unavailable; the general RSD bound instead uses uniform steps controlled by $L_{\max}$ or L_f. See Appendix G.1.2 for the proof.

C.2 Nonconvex case

Theorem C.5 (Last-iterate optimization under local PŁ)**.** *Suppose Definition 4.15, Assumptions 4.1–4.4, and the standing structural conditions (3)–(5) hold. Let* $0 < \eta_t \le 1/L_{\max}$ *for all* $t \ge 0$, *and define* $\mathfrak{C}_{\mathcal{S}}(T) \equiv 2\alpha\beta_{\mathcal{S}}\sum_{t=0}^{T-1}\eta_t\left(1 - \frac{L_{\max}\eta_t}{2}\right)$. *Then, for every* $T \ge 1$, *Algorithm 1 satisfies*

$$\begin{aligned}\mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^T) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\big] &\le \mathbb{E}\Big[e^{-\mathfrak{C}_{\mathcal{S}}(T)}\left(F_{\mathcal{S}}(\boldsymbol{x}^0) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\right)\Big] \\ &\le \frac{L_f}{2}\mathbb{E}\Big[e^{-\mathfrak{C}_{\mathcal{S}}(T)}\|\boldsymbol{x}^0 - \boldsymbol{x}^\star_{\mathcal{S}}\|_2^2\Big]. \end{aligned} \tag{13}$$

Proof. See Appendix G.2.1 for the proof. □

Remark C.6 (Optimization clock). For $\eta_t = 1/L_{\max}$, Theorem C.5 gives $\mathfrak{C}_{\mathcal{S}}(T) = \frac{\alpha\beta_{\mathcal{S}}}{L_{\max}}T$, so a sufficient budget for constant-factor residual reduction is $T \asymp \frac{L_{\max}}{\alpha\beta_{\mathcal{S}}}$. The choices $\alpha = 1$, $\alpha = 1/d$, and $\alpha = 1/m$ recover the standard GD, RCD, and uniform m-block BCD clocks under PŁ-type geometry [22, 40, 47, 23]. Thus the local-PŁ and convex bounds share the $1/\alpha$ coverage slowdown.

For BCD with steps $1/L_{i_t}$, descent is governed by the block-weighted gradient energy. We recall the blockwise PŁ condition from Definition 4.15; the contraction rate combines this curvature with the probability that each block is updated.

Definition C.7 (Blockwise local PŁ condition)**.** Consider the BCD specialization with block matrices $\{U_i\}_{i=1}^m$ and positive smoothness constants $\{L_i\}_{i=1}^m$. Let $\mathcal{U}_0 = \{\boldsymbol{x} \in \mathbb{R}^d : F_{\mathcal{S}}(\boldsymbol{x}) \le F_{\mathcal{S}}(\boldsymbol{x}^0)\}$. We say that $F_{\mathcal{S}}$ satisfies the blockwise local PŁ condition on $\mathcal{U}_0$ with parameter $\beta_{\mathcal{S}}^{\mathcal{L}} > 0$ if

$$\sum_{i=1}^m \frac{1}{L_i}\|U_i^\top \nabla F_{\mathcal{S}}(\boldsymbol{x})\|_2^2 \ge 2\beta_{\mathcal{S}}^{\mathcal{L}}\left(F_{\mathcal{S}}(\boldsymbol{x}) - F_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}}^\star)\right), \qquad \forall\, \boldsymbol{x} \in \mathcal{U}_0. \tag{14}$$

Corollary C.8 (Last-iterate optimization under BCD with blockwise step sizes)**.** *Suppose Definition C.7, Assumptions 4.1–4.4, and Eq. (3) hold. Consider the BCD specialization of Algorithm 1, where $P_i = U_i$, $\rho_i = 1/m$, and $\sum_{i=1}^m U_i U_i^\top = I_d$. Suppose that the sampled block i_t is updated with the blockwise step size $\eta_t = 1/L_{i_t}$. Then, for every $T \ge 1$,*

$$\begin{aligned}\mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^T) - F_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}}^\star)\big] &\le \mathbb{E}\Big[e^{-\beta_{\mathcal{S}}^{\mathcal{L}} T/m}\left(F_{\mathcal{S}}(\boldsymbol{x}^0) - F_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}}^\star)\right)\Big] \\ &\le \frac{L_f}{2}\mathbb{E}\Big[e^{-\beta_{\mathcal{S}}^{\mathcal{L}} T/m}\|\boldsymbol{x}^0 - \boldsymbol{x}_{\mathcal{S}}^\star\|_2^2\Big]. \end{aligned} \tag{15}$$

Proof. See Appendix G.2.2 for the proof. □

Remark C.9 (Optimization clock in block geometry). Corollary C.8 shows how blockwise step sizes change the optimization clock. Uniform-step BCD uses the Euclidean local PŁ constant through $\frac{\beta_{\mathcal{S}}}{mL_{\max}}T$, using the largest smoothness constant $L_{\max}$ for every block. BCD with blockwise step sizes instead uses $\frac{\beta_{\mathcal{S}}^{\mathcal{L}}}{m}T$, where $\beta_{\mathcal{S}}^{\mathcal{L}}$ measures the residual-gradient relation in the block metric induced by the steps $1/L_i$. Since $\sum_{i=1}^m U_i U_i^\top = I_d$ and $L_i \le L_{\max}$, the Euclidean local PŁ condition always makes $\frac{\beta_{\mathcal{S}}}{L_{\max}}$ an admissible blockwise PŁ parameter. Thus the blockwise formulation retains the uniform-step guarantee and can yield a sharper contraction bound when the gradient is concentrated in smoother blocks. This blockwise contraction determines the sufficient horizon in Corollary D.10.

Local PŁ can hold without strong convexity, or even convexity. The following examples give explicit constants, which also yield blockwise constants through Remark C.9.

Rank-deficient least squares. For $F_{\mathcal{S}}(\boldsymbol{x}) = \|X\boldsymbol{x} - \boldsymbol{y}\|_2^2/(2N)$, set $H = X^\top X/N$. Stationarity at $\boldsymbol{x}_{\mathcal{S}}^\star$ gives exact expressions for both the empirical residual and the squared gradient norm:

$$\begin{aligned} F_{\mathcal{S}}(\boldsymbol{x}) - F_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}}^\star) &= \tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_{\mathcal{S}}^\star)^\top H(\boldsymbol{x} - \boldsymbol{x}_{\mathcal{S}}^\star), \\ \|\nabla F_{\mathcal{S}}(\boldsymbol{x})\|_2^2 &= (\boldsymbol{x} - \boldsymbol{x}_{\mathcal{S}}^\star)^\top H^2(\boldsymbol{x} - \boldsymbol{x}_{\mathcal{S}}^\star). \end{aligned}$$

If H has a positive eigenvalue, $H^2 \succeq \lambda_{\min}^+(H)H$ yields PŁ with constant $\lambda_{\min}^+(H)$, even when H is singular. If $H = 0$, the objective is constant.

A nonconvex interpolating example. Consider $f(\boldsymbol{w}; (\boldsymbol{a}, y)) = \frac{1}{2}(\phi(\boldsymbol{a}^\top \boldsymbol{w}) - y)^2$ with $\phi(u) = u + 2\tanh u$. For scalar $a = 1, y = 0$, its second derivative is negative at $w = 1.5$. Nevertheless, if the design matrix A has full row rank, $1 \le \phi' \le 3$ implies

$$\|\nabla F_{\mathcal{S}}(\boldsymbol{w})\|_2^2 \ge 2\lambda_{\min}(AA^\top/N)F_{\mathcal{S}}(\boldsymbol{w}).$$

Full row rank and the surjectivity of ϕ give interpolation, so $F_{\mathcal{S}}^\star = 0$. The same argument controls the neighboring pointwise ratio when the rank condition holds for each relevant neighboring design. This establishes the curvature condition; smoothness and output-moment control remain separate requirements for the excess-risk bound, especially on an unbounded parameter space.

The same calculation extends to nonlinear predictors through their Jacobian. For an interpolating prediction vector $\boldsymbol{h}_\theta = (h_\theta(\boldsymbol{x}_n))_{n=1}^N$, let J_θ be its Jacobian and $K_\theta = J_\theta J_\theta^\top/N$. Squared loss gives

$$\|\nabla F_{\mathcal{S}}(\boldsymbol{\theta})\|_2^2 = \frac{1}{N}(\boldsymbol{h}_\theta - \boldsymbol{y})^\top K_\theta(\boldsymbol{h}_\theta - \boldsymbol{y}) \ge 2\lambda_{\min}(K_\theta)F_{\mathcal{S}}(\boldsymbol{\theta}).$$

A positive lower bound on this eigenvalue throughout the initial sublevel set gives local PŁ; checking observed iterates covers only part of that set. If $\|\nabla h_\theta(\boldsymbol{x})\|_2 \le B_1$, then $\mathrm{tr}(K_\theta) \le B_1^2$ and $\lambda_{\min}(K_\theta) \le B_1^2/N$. This certificate therefore supports coverage comparisons at fixed sample size, while its explicit N dependence prevents a uniform positive lower bound as N grows with fixed B_1.

D Generalization and excess risk bounds

D.1 Convex case

Combining stability and optimization gives convex excess-risk bounds for common-step RSD and blockwise BCD. The stopping choices separate block-count dependence from curvature factors by balancing empirical suboptimality against accumulated sensitivity to replacing one observation.

Theorem D.1 (Generalization error under BCD with blockwise step sizes)**.** *Suppose Definition 4.6, Assumptions 4.1–4.4, and Eq. (3) hold. Consider the BCD specialization of Algorithm 1, namely $P_i = U_i$, $\rho_i = 1/m$, and $\sum_{i=1}^m U_i U_i^\top = I_d$. Suppose that, at iteration t, the sampled block i_t is updated with the blockwise step size $\eta_t = \frac{1}{L_{i_t}}$. Assume additionally that $L_f < 2L_{\min}$. Then, for every $T \geq 1$ and every $\gamma > 0$, the generalization error of Algorithm 1 satisfies*

$$\epsilon_{\text{gen}} \leq \frac{1}{\gamma}\left[F(\boldsymbol{x}^\star) + \frac{m}{T}\Big(D_{\mathcal{L}}(\boldsymbol{x}^0, \boldsymbol{x}^\star) + F(\boldsymbol{x}^0)\Big)\right] + \frac{8L_f^2(1+\gamma)T}{N^2 L_{\min}(2L_{\min} - L_f)}\left[D_{\mathcal{L}}(\boldsymbol{x}^0, \boldsymbol{x}^\star) + F(\boldsymbol{x}^0) + \frac{T}{m}F(\boldsymbol{x}^\star)\right]. \tag{16}$$

Proof. See Appendix G.1.9 for the proof. □

Theorem D.2 (Excess population risk in the convex case)**.** *Suppose Definition 4.6, Assumptions 4.1–4.4, and the standing structural conditions (3)–(5) hold. Let $\boldsymbol{x}^0 \in \mathbb{R}^d$ be deterministic, and let the step sizes satisfy $0 < \eta_{t+1} \leq \eta_t \leq \min\left\{\frac{1}{L_{\max}}, \frac{1}{L_f}\right\}, t \geq 0$. For any $T \geq 1$ and $\gamma > 0$,*

$$\mathbb{E}\big[F(\boldsymbol{x}^T)\big] - F(\boldsymbol{x}^\star) \leq \frac{(1+\eta_0 L_f)\,\mathbb{E}\big[\|\boldsymbol{x}^0 - \boldsymbol{x}^\star_{\mathcal{S}}\|_2^2\big]}{2\alpha\sum_{t=0}^{T-1}\eta_t} + \frac{1}{\gamma}\left[F(\boldsymbol{x}^\star) + \frac{\eta_0\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|_2^2 + 2\eta_0^2 F(\boldsymbol{x}^0)}{2\alpha\sum_{t=0}^{T-1}\eta_t^2}\right] + \frac{2L_f^2(1+\gamma)T}{N^2}\left[\eta_0\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|_2^2 + 2\eta_0^2 F(\boldsymbol{x}^0) + 2\alpha\Big(\sum_{t=0}^{T-1}\eta_t^2\Big)F(\boldsymbol{x}^\star)\right]. \tag{17}$$

Proof. See Appendix G.1.6 for the proof. □

Theorem D.3 (Excess population risk under BCD with blockwise step sizes)**.** *Suppose Definition 4.6, Assumptions 4.1–4.4, and Eq. (3) hold. Consider the BCD specialization of Algorithm 1, namely $P_i = U_i$, $\rho_i = 1/m$, and $\sum_{i=1}^m U_i U_i^\top = I_d$. Suppose that, at iteration t, the sampled block i_t is updated with the blockwise step size $\eta_t = \frac{1}{L_{i_t}}$. Assume additionally that $L_f < 2L_{\min}$. Then, for every $T \geq 1$ and every $\gamma > 0$,*

$$\mathbb{E}\big[F(\boldsymbol{x}^T)\big] - F(\boldsymbol{x}^\star) \leq \frac{m}{T}\left(\mathbb{E}\big[D_{\mathcal{L}}(\boldsymbol{x}^0, \boldsymbol{x}^\star_{\mathcal{S}})\big] + \frac{L_f}{2}\,\mathbb{E}\big[\|\boldsymbol{x}^0 - \boldsymbol{x}^\star_{\mathcal{S}}\|_2^2\big]\right) + \frac{1}{\gamma}\left[F(\boldsymbol{x}^\star) + \frac{m}{T}\Big(D_{\mathcal{L}}(\boldsymbol{x}^0, \boldsymbol{x}^\star) + F(\boldsymbol{x}^0)\Big)\right] + \frac{8L_f^2(1+\gamma)T}{N^2 L_{\min}(2L_{\min} - L_f)}\left[D_{\mathcal{L}}(\boldsymbol{x}^0, \boldsymbol{x}^\star) + F(\boldsymbol{x}^0) + \frac{T}{m}F(\boldsymbol{x}^\star)\right]. \tag{18}$$

Corollary D.4 (Early stopping under blockwise step sizes)**.** *Under the assumptions of Theorem D.3, assume in addition that there exists a constant $C_0 > 0$ such that $\mathbb{E}\big[D_{\mathcal{L}}(\boldsymbol{x}^0, \boldsymbol{x}^\star_{\mathcal{S}})\big] + L_f\,\mathbb{E}\big[\|\boldsymbol{x}^0 - \boldsymbol{x}^\star_{\mathcal{S}}\|_2^2\big] + D_{\mathcal{L}}(\boldsymbol{x}^0, \boldsymbol{x}^\star) + F(\boldsymbol{x}^0) + F(\boldsymbol{x}^\star) \leq C_0$. Let $K := L_f^2/[L_{\min}(2L_{\min} - L_f)]$ and assume moreover that $m \leq C_m N^2$ for a fixed constant $C_m > 0$. If $T = \lceil m^{3/4}\sqrt{N}\rceil$ and $\gamma = N\sqrt{m}/T$, then*

$$\mathbb{E}[F(\boldsymbol{x}^T)] - F(\boldsymbol{x}^\star) = O\left(\frac{C_0(1+K)m^{1/4}}{\sqrt{N}}\right),$$

where the constant depends only on C_m; see Appendix G.1.11 for the proof.

Corollary D.5 (Low-noise regime under blockwise step sizes)**.** *Under the assumptions of Corollary D.4 other than $m \le C_m N^2$, assume in addition that $F(\boldsymbol{x}^\star) \le C_0/N$. If $T = \lceil N\sqrt{m} \rceil$ and $\gamma = 1$, then, for all integers $N, m \ge 1$,*

$$\mathbb{E}[F(\boldsymbol{x}^T)] - F(\boldsymbol{x}^\star) \le (3 + 96K)\frac{C_0\sqrt{m}}{N}.$$

Proof. See Appendix G.1.12 for the proof. □

Remark D.6 (Block geometry in excess risk under blockwise step sizes). The rates in Corollaries D.4–D.5 have the same block-count dependence as common-step RSD at $\alpha = 1/m$. The refinement lies in the weighted initialization $D_{\mathcal{L}}$ and the curvature factor $K = L_f^2/[L_{\min}(2L_{\min} - L_f)]$. A smaller weighted potential can improve the bound, whereas a small margin $2L_{\min} - L_f$ increases K. Comparing blockwise and common steps therefore requires both the initialization geometry and the curvature constants, even when the coverage and the stopping orders are the same. The margin $2L_{\min} - L_f$ in the curvature factor above comes from the same-sample part of the BCD stability recursion. For the coupled difference $\boldsymbol{\delta} = \boldsymbol{x} - \boldsymbol{y}$ and shared-component gradient difference $\boldsymbol{h}$, cocoercivity and $L_i \ge L_{\min}$ give the following bound on the weighted shared-sample contribution:

$$-\frac{1}{m}\langle \boldsymbol{h}, \boldsymbol{\delta}\rangle + \frac{1}{2m}\sum_i \frac{\|U_i^\top \boldsymbol{h}\|_2^2}{L_i} \le -\frac{2L_{\min} - L_f}{2mL_f L_{\min}}\|\boldsymbol{h}\|_2^2.$$

When $L_f < 2L_{\min}$, the negative coefficient provides the margin used to absorb the cross term from the replaced sample. When $L_f \ge 2L_{\min}$, the bound loses its margin, although the unreduced expression can remain negative if $2\langle \boldsymbol{h}, \boldsymbol{\delta}\rangle > \sum_i \|U_i^\top \boldsymbol{h}\|_2^2/L_i$. A general replacement-stability bound based on this directional condition remains open.

Coverage, horizon, and step size. For fixed problem constants, smaller α requires a longer stopping horizon. The common-step bound uses $\eta_t \le 1/L_f$; the convex stability bound for blockwise steps uses $L_f < 2L_{\min}$. For isotropic SSD with an iteration budget B, the general-convex horizon requires $r \gtrsim d(\sqrt{N}/B)^{4/3}$, feasible only for $B \gtrsim \sqrt{N}$. Under low noise, it requires $r \gtrsim d(N/B)^2$ with $B \gtrsim N$. These rank conditions follow by substituting isotropic coverage into the prescribed stopping horizons and solving for the rank.

D.2 Nonconvex case

Theorem D.7 (Excess risk under local PŁ)**.** *Suppose the assumptions of Theorem C.5 and Lemma 4.17 hold. Then, for every $T \ge 1$,*

$$\begin{aligned}\mathbb{E}[F(\boldsymbol{x}^T)] - F(\boldsymbol{x}^\star) &\le \mathbb{E}\left[e^{-\mathfrak{C}_{\mathcal{S}}(T)}\left(F_{\mathcal{S}}(\boldsymbol{x}^0) - F_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}}^\star)\right)\right] \\ &+ \mathbb{E}\left[\frac{2\|\nabla f(\boldsymbol{x}^T;\xi_1)\|_2^2 + 2\|\nabla f(\boldsymbol{x}^T;\xi_1')\|_2^2}{N\,\beta_{\mathcal{S}^{(1)}}^{\mathrm{pt}}(\boldsymbol{x}^T)}\right] + \mathbb{E}\left[\frac{2L_f N\left(F_{\mathcal{S}}(\boldsymbol{x}^T) - F_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}}^\star)\right)}{\beta_{\mathcal{S}^{(1)}}^{\mathrm{pt}}(\boldsymbol{x}^T)}\right]. \end{aligned} \tag{19}$$

Proof. See Appendix G.2.4 for the proof. □

Remark D.8 (Structure of the nonconvex excess-risk bound). The first term in Theorem D.7 is the optimization residual, governed by the clock $\mathfrak{C}_{\mathcal{S}}(T)$; the second controls loss stability through sample gradients and neighboring curvature. The third couples the empirical residual with the inverse neighboring pointwise ratio. Both depend on the same random output, so the expectation need not factor. Training local PŁ does not ensure a positive uniform lower bound on the neighboring ratio. Corollary 4.18 instead controls the coupling through Hölder's inequality and output moments.

Corollary D.9 (Excess risk under BCD with blockwise step sizes)**.** *Suppose the assumptions of Corollary C.8 and Lemma 4.17 hold. Then, for every $T \ge 1$,*

$$\begin{aligned}\mathbb{E}[F(\boldsymbol{x}^T)] - F(\boldsymbol{x}^\star) &\le \mathbb{E}\left[e^{-\beta\xi_{\mathcal{S}}^{\mathcal{L}} T/m}\left(F_{\mathcal{S}}(\boldsymbol{x}^0) - F_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}}^\star)\right)\right] \\ &+ \mathbb{E}\left[\frac{2\|\nabla f(\boldsymbol{x}^T;\xi_1)\|_2^2 + 2\|\nabla f(\boldsymbol{x}^T;\xi_1')\|_2^2}{N\,\beta_{\mathcal{S}^{(1)}}^{\mathrm{pt}}(\boldsymbol{x}^T)}\right] + \mathbb{E}\left[\frac{2L_f N\left(F_{\mathcal{S}}(\boldsymbol{x}^T) - F_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}}^\star)\right)}{\beta_{\mathcal{S}^{(1)}}^{\mathrm{pt}}(\boldsymbol{x}^T)}\right]. \end{aligned} \tag{20}$$

Proof. See Appendix G.2.6 for detailed proofs. □

Corollary D.10 (Stopped excess risk for BCD with blockwise step sizes)**.** *Suppose the assumptions of Corollary C.8 and Lemma 4.17 hold, with $L_f > 0$. Let $p > 1$ and $q = p/(p-1)$. Define $R_0 := F_{\mathcal{S}}(\boldsymbol{x}^0) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})$, $\log_+(u) := \log(u \vee 1)$, $u \geq 0$. Use the moment quantities $\mathfrak{G}_p(T)$ and $\mathfrak{I}_q(T)$ from Eq. (10). Let $T = T(\mathcal{S}) \geq 1$ be integer-valued, measurable, and permutation-invariant, fixed before the block draws, and let $\boldsymbol{x}^T = \boldsymbol{x}^{T(\mathcal{S})}$ denote the stopped BCD output. Assume $\mathfrak{G}_p(T) < \infty$ and $\mathfrak{I}_q(T) < \infty$. If*

$$\frac{\beta^{\mathcal{L}}_{\mathcal{S}}}{m} T \geq p \log_+\left(N^2 L_f R_0\right), \tag{21}$$

then

$$\mathbb{E}[F(\boldsymbol{x}^T)] - F(\boldsymbol{x}^\star) \leq \frac{1}{N^2 L_f} + \frac{2}{N}\mathfrak{G}_p(T)\mathfrak{I}_q(T)^{1/q} + \frac{2}{N}\mathfrak{I}_q(T)^{1/q}. \tag{22}$$

Proof. See Appendix G.2.7. □

Remark D.11 (Scaling with blockwise step sizes). Compared with common steps $1/L_{\max}$, the blockwise stopping condition uses $\beta^{\mathcal{L}}_{\mathcal{S}}$ in place of the conservative curvature scale $\beta_{\mathcal{S}}/L_{\max}$. As explained in Remark C.9, this parameter uses the same weighted gradient energy as the update. With the same p and residual target, the sufficient horizon can therefore shorten when $\beta^{\mathcal{L}}_{\mathcal{S}} > \beta_{\mathcal{S}}/L_{\max}$. The sample-gradient and inverse-curvature terms retain their functional form, but their values depend on the joint law of the data, replacement sample, and stopped output generated by the chosen steps.

Remark D.12 (Sufficient conditions for uniform output moments). Within a fixed family of data-independent projector laws, consider the following sufficient conditions for uniform moment bounds. For almost every pair $(\mathcal{S}, \mathcal{S}')$, suppose all sample-gradient norms, including replacements, are bounded by G on the training sublevel set $\mathcal{U}_0$. On the same set, assume $\beta^{\mathrm{pt}}_{\mathcal{S}^{(1)}}(\boldsymbol{x}) \geq \mu > 0$ for every $\boldsymbol{x}$. Pathwise descent gives $\mathfrak{G}_p(T_\alpha) \leq 2G^2$ and $\mathfrak{I}_q(T_\alpha)^{1/q} \leq 1/\mu$. Substituting these moment bounds in Eq. (12) gives the uniform excess-risk estimate:

$$\mathbb{E}[F(\boldsymbol{x}^{T_\alpha})] - F(\boldsymbol{x}^\star) \leq \frac{1}{N^2 L_f} + \frac{4G^2 + 2}{N\mu}.$$

The bound is uniform over coverages for which the same constants and required budgets apply. An almost-sure positive margin is sufficient but not necessary. Suppose $\Pr(\beta^{\mathrm{pt}}_{\mathcal{S}^{(1)}}(\boldsymbol{x}^{T_\alpha}) \leq u) \leq C_{\mathrm{tail}} u^\nu$ uniformly over coverage, for $0 < u \leq u_0$ and $\nu > q$. Integrating this lower-tail bound gives

$$\mathfrak{I}_q(T_\alpha) \leq u_0^{-q} + qC_{\mathrm{tail}} u_0^{\nu-q}/(\nu - q).$$

Together with uniform $2p$-moment bounds on sample-gradient norms, this controls the product in Remark 4.19. The lower-tail condition is additional to local PŁ of the training objective and concerns the neighboring ratio at the stopped output.

A deterministic sufficient horizon. If bounds $\beta_{\mathcal{S}} \geq \underline{\beta} > 0$ and $R_0 \leq \overline{R}_0$ are known, use common steps $1/L_{\max}$ and the fixed horizon

$$T = \max\left\{1, \left\lceil \frac{pL_{\max}}{\alpha\underline{\beta}} \log_+(N^2 L_f \overline{R}_0) \right\rceil\right\}$$

which satisfies (11). The horizon is fixed before the subspace draws; a trajectory-dependent first-hitting time requires a separate argument.

E Sampling models and projected updates

Uniform sampling from a fixed complete partition into coordinates, layers, or modules is jointly isotropic for any block sizes. The mean projector also provides a way to analyze fixed nonuniform sampling and distinguish RSD from learned low-rank updates. We retain full-gradient updates with a law independent of the data and iterates; adaptive selection requires separate analysis.

E.1 Fixed anisotropic sampling

Let the orthogonal projector $Q = Q^\top = Q^2$ have a fixed, data-independent law and be sampled independently at each iteration. Write $M = \mathbb{E}Q$, with $0 < cI \preceq M \preceq CI$, and use the weighted norm $\|\boldsymbol{z}\|_{M^{-1}}^2 = \boldsymbol{z}^\top M^{-1}\boldsymbol{z}$. Here $L_{\max}$ uniformly bounds the subspace smoothness of F_S over the support of Q. For a convex L_f-smooth function g, set $\boldsymbol{\delta} = \boldsymbol{x} - \boldsymbol{y}$ and $\boldsymbol{h} = \nabla g(\boldsymbol{x}) - \nabla g(\boldsymbol{y})$. Subspace smoothness and cocoercivity give, for $0 < \eta \le 1/L_{\max}$,

$$\mathbb{E}_t F_S(\boldsymbol{x} - \eta Q \nabla F_S(\boldsymbol{x})) \le F_S(\boldsymbol{x}) - c\eta(1 - L_{\max}\eta/2)\|\nabla F_S(\boldsymbol{x})\|_2^2,$$
$$\mathbb{E}_t\|\boldsymbol{\delta} - \eta Q\boldsymbol{h}\|_{M^{-1}}^2 \le \|\boldsymbol{\delta}\|_{M^{-1}}^2 - \eta(2/L_f - \eta C/c)\|\boldsymbol{h}\|_2^2.$$

The first inequality controls empirical descent. The second uses $\mathbb{E}[QM^{-1}Q] \preceq M/c \preceq (C/c)I$ and independence from the current pair. Thus $\eta \le 2c/(CL_f)$ ensures expected nonexpansiveness for the coupled update in this weighted norm.

Under both step-size restrictions, take $g = N^{-1}\sum_{j\neq n} f(\cdot;\xi_j)$ and $\boldsymbol{b} = \nabla f(\boldsymbol{x};\xi_n) - \nabla f(\boldsymbol{y};\xi_n')$ for one-sample replacement. The next difference is $\boldsymbol{\delta}^+ = \boldsymbol{\delta} - \eta Q\boldsymbol{h} - (\eta/N)Q\boldsymbol{b}$, so conditional Minkowski gives the one-step bound:

$$\left(\mathbb{E}_t\|\boldsymbol{\delta}^+\|_{M^{-1}}^2\right)^{1/2} \le \|\boldsymbol{\delta}\|_{M^{-1}} + \frac{\eta}{N}\sqrt{C/c}\,\|\boldsymbol{b}\|_2.$$

Since $C^{-1}\|\boldsymbol{u}\|_2^2 \le \|\boldsymbol{u}\|_{M^{-1}}^2 \le c^{-1}\|\boldsymbol{u}\|_2^2$, converting the recursion to Euclidean discrepancy introduces spectral constants. An excess-risk bound then requires controlling the replacement perturbations accumulated over training.

E.2 LoRA and projected updates

LoRA [19] writes $W = W_0 + BA$. With $G = \nabla_W F(W_0 + BA)$, simultaneous gradient steps on the two factors give

$$B^+A^+ - BA = -\eta(BB^\top G + GA^\top A) + \eta^2 GA^\top B^\top G.$$

The first-order operator depends on the learned factors and is generally not an orthogonal projector. The quadratic term also has no counterpart in an RSD update. Thus the rank-r manifold dimension $r(d_{\text{out}} + d_{\text{in}} - r)$ does not determine the coverage parameter α.

If $R_t \in \mathbb{R}^{d_{\text{in}}\times r}$ is independently resampled with orthonormal columns and $\mathbb{E}[R_tR_t^\top] = (r/d_{\text{in}})I$, the projected gradient update

$$W^{t+1} = W^t - \eta_t \nabla F(W^t) R_t R_t^\top$$

vectorizes to RSD with projector $(R_tR_t^\top) \otimes I_{d_{\text{out}}}$ and coverage $\alpha = r/d_{\text{in}}$.

F Technical lemmas and definitions

We collect the descent, nonexpansiveness, and moment estimates used in the proofs. The final three lemmas connect local-PŁ contraction to the stopped risk bound.

Definition F.1 (Symmetric algorithm)**.** A randomized learning algorithm $\mathcal{A}$ is said to be *symmetric* if its output distribution is invariant under permutations of the training samples. More precisely, for any dataset $\mathcal{S} = \{\xi_1, \dots, \xi_N\}$ and any permutation π of $[N]$, if $\pi(\mathcal{S}) := \{\xi_{\pi(1)}, \dots, \xi_{\pi(N)}\}$, then

$$\mathcal{A}(\mathcal{S}) \overset{d}{=} \mathcal{A}(\pi(\mathcal{S})),$$

where $\overset{d}{=}$ denotes equality in distribution.

Remark F.2. Algorithm 1 is symmetric because the empirical risk F_S is permutation-invariant and the subspace draws are independent of sample ordering. This applies to a fixed horizon T and to stopping rules based on permutation-invariant quantities.

Lemma F.3 (Averaged descent)**.** *Suppose Assumption 4.4 and Eqs. (3)–(5) hold. For positive step sizes and every $t \ge 0$,*

$$\mathbb{E}_t\left[F_S(\boldsymbol{x}^{t+1})\right] \le F_S(\boldsymbol{x}^t) - \alpha\eta_t\left(1 - \frac{L_{\max}\eta_t}{2}\right)\|\nabla F_S(\boldsymbol{x}^t)\|_2^2. \tag{23}$$

In particular, if $\eta_t \le 1/L_{\max}$, then

$$\mathbb{E}_t\big[F_{\mathcal{S}}(\boldsymbol{x}^{t+1})\big] \le F_{\mathcal{S}}(\boldsymbol{x}^t) - \frac{\alpha\eta_t}{2}\|\nabla F_{\mathcal{S}}(\boldsymbol{x}^t)\|_2^2 \le F_{\mathcal{S}}(\boldsymbol{x}^t). \tag{24}$$

Proof. Given $\mathcal{S}$ and all algorithmic randomness before iteration t, the iterate $\boldsymbol{x}^t$ is fixed. The expectation $\mathbb{E}_t[\cdot]$ averages only over the current choice of (i_t, P_{i_t}).

By Eq. (4), $\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t P_{i_t} P_{i_t}^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}^t)$. Applying Assumption 4.4 at the point $\boldsymbol{x}^t$ we obtain

$$\begin{aligned}
F_{\mathcal{S}}(\boldsymbol{x}^{t+1}) &= F_{\mathcal{S}}\big(\boldsymbol{x}^t - \eta_t P_{i_t} P_{i_t}^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}^t)\big) \\
&\le F_{\mathcal{S}}(\boldsymbol{x}^t) + \big\langle P_{i_t}^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}^t), -\eta_t P_{i_t}^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}^t)\big\rangle + \frac{L_{i_t}\eta_t^2}{2}\left\|P_{i_t}^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}^t)\right\|_2^2 \\
&= F_{\mathcal{S}}(\boldsymbol{x}^t) - \eta_t \left\|P_{i_t}^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}^t)\right\|_2^2 + \frac{L_{i_t}\eta_t^2}{2}\left\|P_{i_t}^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}^t)\right\|_2^2 \\
&\le F_{\mathcal{S}}(\boldsymbol{x}^t) - \eta_t\left(1 - \frac{L_{\max}\eta_t}{2}\right)\left\|P_{i_t}^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}^t)\right\|_2^2,
\end{aligned} \tag{25}$$

where the last step uses $L_{i_t} \le L_{\max}$.

Taking conditional expectation over the sampled subspace and using Eq. (5) shows how coverage scales the squared projected gradient norm:

$$\begin{aligned}
\mathbb{E}_t\left[\left\|P_{i_t}^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}^t)\right\|_2^2\right] &= \mathbb{E}_t\big[\nabla F_{\mathcal{S}}(\boldsymbol{x}^t)^\top P_{i_t} P_{i_t}^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}^t)\big] \\
&= \nabla F_{\mathcal{S}}(\boldsymbol{x}^t)^\top \mathbb{E}_t\big[P_{i_t} P_{i_t}^\top\big]\, \nabla F_{\mathcal{S}}(\boldsymbol{x}^t) \\
&= \alpha\|\nabla F_{\mathcal{S}}(\boldsymbol{x}^t)\|_2^2.
\end{aligned} \tag{26}$$

Combining Eqs. (25) and (26) yields Eq. (23).

If $\eta_t \le 1/L_{\max}$, then $1 - \frac{L_{\max}\eta_t}{2} \ge \frac{1}{2}$, and Eq. (24) follows immediately. □

Lemma F.4 (Convex optimization recursion)**.** *Suppose the losses are convex as in Definition 4.6, Assumptions 4.1–4.4 hold, and the standing structural conditions Eqs. (3)–(5) hold. Let the step sizes satisfy $0 < \eta_{t+1} \le \eta_t \le 1/L_{\max}$ for all $t \ge 0$.*

(i) For every $\boldsymbol{x} \in \mathbb{R}^d$ that is independent of the algorithmic randomness and satisfies $F_{\mathcal{S}}(\boldsymbol{x}) \le F_{\mathcal{S}}(\boldsymbol{x}^t)$ almost surely for all $t \ge 0$, and for every $T \ge 1$,

$$\mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^T) - F_{\mathcal{S}}(\boldsymbol{x})\big] \le \frac{\mathbb{E}\big[\|\boldsymbol{x}^0 - \boldsymbol{x}\|_2^2\big] + 2\eta_0\,\mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^0) - F_{\mathcal{S}}(\boldsymbol{x})\big]}{2\alpha\sum_{t=0}^{T-1}\eta_t}. \tag{27}$$

(ii) For every $\boldsymbol{x} \in \mathbb{R}^d$ that is independent of the algorithmic randomness, and for every $T \ge 1$,

$$2\alpha\sum_{t=0}^{T-1}\eta_t^2\,\mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x})\big] \le \eta_0\,\mathbb{E}\big[\|\boldsymbol{x}^0 - \boldsymbol{x}\|_2^2\big] + 2\eta_0^2\,\mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^0)\big]. \tag{28}$$

Proof. Fix a comparator $\boldsymbol{x} \in \mathbb{R}^d$ that is independent of the algorithmic randomness. Here $\mathbb{E}_t[\cdot]$ denotes conditional expectation over the current random choice (i_t, P_{i_t}), conditioned on $\mathcal{S}$ and on all randomness up to iteration t. Unconditional expectations are taken over the dataset, when random, and all algorithmic randomness. Expanding the squared distance to $\boldsymbol{x}$ using (4) gives

$$\begin{aligned}
\mathbb{E}_t\big[\|\boldsymbol{x}^{t+1} - \boldsymbol{x}\|_2^2\big] &= \|\boldsymbol{x}^t - \boldsymbol{x}\|_2^2 + \eta_t^2\mathbb{E}_t\big[\|P_{i_t}P_{i_t}^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}^t)\|_2^2\big] \\
&\quad + 2\eta_t\mathbb{E}_t\big[\big\langle \boldsymbol{x} - \boldsymbol{x}^t,\, P_{i_t}P_{i_t}^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}^t)\big\rangle\big].
\end{aligned} \tag{29}$$

By the subspace orthonormality condition (3), $P_i P_i^\top$ is an orthogonal projector, and hence $\|P_i P_i^\top v\|_2^2 = v^\top P_i P_i^\top v$. Together with the joint isotropy condition (5), this gives

$$\begin{aligned}
\mathbb{E}_t\big[\|\boldsymbol{x}^{t+1} - \boldsymbol{x}\|_2^2\big] &= \|\boldsymbol{x}^t - \boldsymbol{x}\|_2^2 + \alpha\eta_t^2\|\nabla F_{\mathcal{S}}(\boldsymbol{x}^t)\|_2^2 \\
&\quad + 2\alpha\eta_t\big\langle \boldsymbol{x} - \boldsymbol{x}^t,\, \nabla F_{\mathcal{S}}(\boldsymbol{x}^t)\big\rangle.
\end{aligned} \tag{30}$$

Since $F_{\mathcal{S}}$ is convex by Definition 4.6, rearranging therefore yields

$$2\alpha\eta_t\big(F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x})\big) \le \|\boldsymbol{x}^t - \boldsymbol{x}\|_2^2 - \mathbb{E}_t\big[\|\boldsymbol{x}^{t+1} - \boldsymbol{x}\|_2^2\big] + \alpha\eta_t^2\|\nabla F_{\mathcal{S}}(\boldsymbol{x}^t)\|_2^2. \tag{31}$$

Lemma F.3, applied with $\eta_t \le 1/L_{\max}$, implies

$$\alpha\eta_t^2\|\nabla F_{\mathcal{S}}(\boldsymbol{x}^t)\|_2^2 \le 2\eta_t\left(F_{\mathcal{S}}(\boldsymbol{x}^t) - \mathbb{E}_t[F_{\mathcal{S}}(\boldsymbol{x}^{t+1})]\right).$$

Substituting this bound into (31) and taking total expectation gives the basic recursion

$$\begin{aligned}2\alpha\eta_t\mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x})\big] \le{}& \mathbb{E}\big[\|\boldsymbol{x}^t - \boldsymbol{x}\|_2^2\big] - \mathbb{E}\big[\|\boldsymbol{x}^{t+1} - \boldsymbol{x}\|_2^2\big] \\ &+ 2\eta_t\mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x}^{t+1})\big].\end{aligned} \tag{32}$$

We first prove (27). Summing (32) over $t = 0, \dots, T-1$, the squared distance terms telescope and are bounded above by $\mathbb{E}[\|\boldsymbol{x}^0 - \boldsymbol{x}\|_2^2]$. For the function-value terms, since the comparator satisfies $F_{\mathcal{S}}(\boldsymbol{x}) \le F_{\mathcal{S}}(\boldsymbol{x}^t)$ for all t, Abel's summation formula gives

$$\begin{aligned}\sum_{t=0}^{T-1}\eta_t\mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x}^{t+1})\big] ={}& \eta_0\mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^0) - F_{\mathcal{S}}(\boldsymbol{x})\big] \\ &+ \sum_{t=1}^{T-1}(\eta_t - \eta_{t-1})\mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x})\big] \\ &- \eta_{T-1}\mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^T) - F_{\mathcal{S}}(\boldsymbol{x})\big].\end{aligned} \tag{33}$$

The last two terms are nonpositive because the step sizes are nonincreasing and, by the comparator condition, the displayed excess risks are nonnegative almost surely. Hence

$$2\alpha\sum_{t=0}^{T-1}\eta_t\mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x})\big] \le \mathbb{E}\big[\|\boldsymbol{x}^0 - \boldsymbol{x}\|_2^2\big] + 2\eta_0\mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^0) - F_{\mathcal{S}}(\boldsymbol{x})\big]. \tag{34}$$

Moreover, Lemma F.3 implies that $\mathbb{E}[F_{\mathcal{S}}(\boldsymbol{x}^t)]$ is nonincreasing in t. Therefore, for every $0 \le t \le T-1$,

$$\mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^T) - F_{\mathcal{S}}(\boldsymbol{x})\big] \le \mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x})\big].$$

Multiplying by η_t and summing over t yields

$$\mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^T) - F_{\mathcal{S}}(\boldsymbol{x})\big]\sum_{t=0}^{T-1}\eta_t \le \sum_{t=0}^{T-1}\eta_t\mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x})\big].$$

Combining this display with (34) proves (27).

It remains to prove (28). Multiplying (32) by η_t and summing over $t = 0, \dots, T-1$ gives

$$\begin{aligned}2\alpha\sum_{t=0}^{T-1}\eta_t^2\mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x})\big] \le{}& \sum_{t=0}^{T-1}\eta_t\left(\mathbb{E}\big[\|\boldsymbol{x}^t - \boldsymbol{x}\|_2^2\big] - \mathbb{E}\big[\|\boldsymbol{x}^{t+1} - \boldsymbol{x}\|_2^2\big]\right) \\ &+ 2\sum_{t=0}^{T-1}\eta_t^2\mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x}^{t+1})\big].\end{aligned} \tag{35}$$

Applying Abel's summation formula to the first sum and using $\eta_t \le \eta_{t-1}$ together with nonnegativity of squared norms gives

$$\sum_{t=0}^{T-1}\eta_t\left(\mathbb{E}\big[\|\boldsymbol{x}^t - \boldsymbol{x}\|_2^2\big] - \mathbb{E}\big[\|\boldsymbol{x}^{t+1} - \boldsymbol{x}\|_2^2\big]\right) \le \eta_0\mathbb{E}\big[\|\boldsymbol{x}^0 - \boldsymbol{x}\|_2^2\big].$$

Similarly, Abel's summation formula, the monotonicity of η_t^2, and Assumption 4.1 imply

$$\sum_{t=0}^{T-1}\eta_t^2\mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x}^{t+1})\big] \le \eta_0^2\mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^0)\big].$$

Substituting the last two displays into (35) proves (28). □

Lemma F.5 (Self-bounding property)**.** *Suppose Assumptions 4.1 and 4.2 hold. Then, for every sample ξ and every $\boldsymbol{x} \in \mathbb{R}^d$,*

$$\|\nabla f(\boldsymbol{x};\xi)\|_2^2 \le 2L_f\, f(\boldsymbol{x};\xi).$$

Proof. Fix any sample ξ and any point $\boldsymbol{x} \in \mathbb{R}^d$. By Assumption 4.2, the function $f(\cdot;\xi)$ is L_f-smooth, and therefore $f(\boldsymbol{y};\xi) \le f(\boldsymbol{x};\xi) + \langle \nabla f(\boldsymbol{x};\xi), \boldsymbol{y}-\boldsymbol{x}\rangle + \frac{L_f}{2}\|\boldsymbol{y}-\boldsymbol{x}\|_2^2, \forall\, \boldsymbol{y} \in \mathbb{R}^d$. Applying this inequality at $\boldsymbol{y} = \boldsymbol{x} - \frac{1}{L_f}\nabla f(\boldsymbol{x};\xi)$, we obtain

$$\begin{aligned} f\left(\boldsymbol{x} - \frac{1}{L_f}\nabla f(\boldsymbol{x};\xi);\xi\right) &\le f(\boldsymbol{x};\xi) - \frac{1}{L_f}\|\nabla f(\boldsymbol{x};\xi)\|_2^2 + \frac{L_f}{2}\left\|\frac{1}{L_f}\nabla f(\boldsymbol{x};\xi)\right\|_2^2 \\ &= f(\boldsymbol{x};\xi) - \frac{1}{2L_f}\|\nabla f(\boldsymbol{x};\xi)\|_2^2. \end{aligned} \tag{36}$$

On the other hand, Assumption 4.1 implies that $f\left(\boldsymbol{x} - \frac{1}{L_f}\nabla f(\boldsymbol{x};\xi);\xi\right) \ge 0$.

Combining this with Eq. (36) yields

$$0 \le f(\boldsymbol{x};\xi) - \frac{1}{2L_f}\|\nabla f(\boldsymbol{x};\xi)\|_2^2.$$

□

Lemma F.6 (Self-bounding perturbation)**.** *Suppose Assumptions 4.1 and 4.2 hold, and Eqs. (3) and (5) hold. Then, for every $\boldsymbol{x} \in \mathbb{R}^d$ and every pair of samples ξ, ξ',*

$$\mathbb{E}_{i,P_i}\left[\left\|P_i P_i^\top\left(\nabla f(\boldsymbol{x};\xi) - \nabla f(\boldsymbol{x};\xi')\right)\right\|_2^2\right] \le 4\alpha L_f\left(f(\boldsymbol{x};\xi) + f(\boldsymbol{x};\xi')\right).$$

Proof. Fix any $\boldsymbol{x} \in \mathbb{R}^d$ and any samples ξ, ξ'. By Eq. (3), the matrix $P_i P_i^\top$ is an orthogonal projector, and hence $(P_i P_i^\top)^2 = P_i P_i^\top$, $\|P_i P_i^\top v\|_2^2 = v^\top P_i P_i^\top v$. Taking expectation with respect to the sampled type i and P_i, and using Eq. (5), we obtain

$$\begin{aligned} &\mathbb{E}_{i,P_i}\left[\left\|P_i P_i^\top\left(\nabla f(\boldsymbol{x};\xi) - \nabla f(\boldsymbol{x};\xi')\right)\right\|_2^2\right] \\ &\quad = \alpha\|\nabla f(\boldsymbol{x};\xi) - \nabla f(\boldsymbol{x};\xi')\|_2^2. \end{aligned} \tag{37}$$

Next, the elementary inequality $\|\boldsymbol{a}-\boldsymbol{b}\|_2^2 \le 2\|\boldsymbol{a}\|_2^2 + 2\|\boldsymbol{b}\|_2^2$ yields

$$\|\nabla f(\boldsymbol{x};\xi) - \nabla f(\boldsymbol{x};\xi')\|_2^2 \le 2\|\nabla f(\boldsymbol{x};\xi)\|_2^2 + 2\|\nabla f(\boldsymbol{x};\xi')\|_2^2.$$

Under Assumptions 4.1 and 4.2, Lemma F.5 gives

$$\|\nabla f(\boldsymbol{x};\xi) - \nabla f(\boldsymbol{x};\xi')\|_2^2 \le 4L_f\left(f(\boldsymbol{x};\xi) + f(\boldsymbol{x};\xi')\right).$$

Combining this estimate with Eq. (37) proves the lemma. □

Lemma F.7 (Expected nonexpansiveness in the convex case)**.** *Suppose the losses are convex as in Definition 4.6, Assumption 4.2 holds, and the standing structural conditions Eqs. (3)–(5) hold. Let $\mathcal{S}$ be any dataset, and let $\eta_t \in (0, 2/L_f]$. For any $\boldsymbol{x} \in \mathbb{R}^d$, define*

$$\Psi^{(t)}_{i_t,P_{i_t}}(\boldsymbol{x}) := \boldsymbol{x} - \eta_t\, P_{i_t} P_{i_t}^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}),$$

where $i_t \sim (\rho_1, \dots, \rho_m)$ and P_{i_t} is drawn according to its law. Then, for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,

$$\mathbb{E}_{i_t,P_{i_t}}\left[\|\Psi^{(t)}_{i_t,P_{i_t}}(\boldsymbol{x}) - \Psi^{(t)}_{i_t,P_{i_t}}(\boldsymbol{y})\|_2^2\right] \le \|\boldsymbol{x}-\boldsymbol{y}\|_2^2 - \alpha\eta_t\left(\frac{2}{L_f} - \eta_t\right)\|\nabla F_{\mathcal{S}}(\boldsymbol{x}) - \nabla F_{\mathcal{S}}(\boldsymbol{y})\|_2^2. \tag{38}$$

In particular,

$$\mathbb{E}_{i_t,P_{i_t}}\left[\|\Psi^{(t)}_{i_t,P_{i_t}}(\boldsymbol{x}) - \Psi^{(t)}_{i_t,P_{i_t}}(\boldsymbol{y})\|_2^2\right] \le \|\boldsymbol{x}-\boldsymbol{y}\|_2^2. \tag{39}$$

Proof. Fix any dataset $\mathcal{S}$. Subtracting the two update maps under the shared subspace draw gives:

$$\Psi^{(t)}_{i_t,P_{i_t}}(\boldsymbol{x}) - \Psi^{(t)}_{i_t,P_{i_t}}(\boldsymbol{y}) = \boldsymbol{x} - \boldsymbol{y} - \eta_t P_{i_t} P_{i_t}^\top \big(\nabla F_{\mathcal{S}}(\boldsymbol{x}) - \nabla F_{\mathcal{S}}(\boldsymbol{y})\big).$$

Expanding the squared norm of this difference separates the distance, mixed and gradient terms:

$$\begin{aligned}\|\Psi^{(t)}_{i_t,P_{i_t}}(\boldsymbol{x}) - \Psi^{(t)}_{i_t,P_{i_t}}(\boldsymbol{y})\|_2^2 = \|\boldsymbol{x}-\boldsymbol{y}\|_2^2 - 2\eta_t \Big\langle \boldsymbol{x}-\boldsymbol{y},\, P_{i_t}P_{i_t}^\top\big(\nabla F_{\mathcal{S}}(\boldsymbol{x}) - \nabla F_{\mathcal{S}}(\boldsymbol{y})\big)\Big\rangle \\ + \eta_t^2 \left\| P_{i_t}P_{i_t}^\top\big(\nabla F_{\mathcal{S}}(\boldsymbol{x}) - \nabla F_{\mathcal{S}}(\boldsymbol{y})\big)\right\|_2^2. \qquad (40)\end{aligned}$$

Since Eq. (3) implies that $P_{i_t}P_{i_t}^\top$ is an orthogonal projector, the last term in Eq. (40) can be written as

$$\big(\nabla F_{\mathcal{S}}(\boldsymbol{x}) - \nabla F_{\mathcal{S}}(\boldsymbol{y})\big)^\top P_{i_t}P_{i_t}^\top\big(\nabla F_{\mathcal{S}}(\boldsymbol{x}) - \nabla F_{\mathcal{S}}(\boldsymbol{y})\big).$$

Taking expectation with respect to i_t and P_{i_t} and using Eq. (5), we get

$$\begin{aligned}\mathbb{E}_{i_t,P_{i_t}}\left[\|\Psi^{(t)}_{i_t,P_{i_t}}(\boldsymbol{x}) - \Psi^{(t)}_{i_t,P_{i_t}}(\boldsymbol{y})\|_2^2\right] = \|\boldsymbol{x}-\boldsymbol{y}\|_2^2 - 2\alpha\eta_t \langle \boldsymbol{x}-\boldsymbol{y},\, \nabla F_{\mathcal{S}}(\boldsymbol{x}) - \nabla F_{\mathcal{S}}(\boldsymbol{y})\rangle \\ + \alpha\eta_t^2 \|\nabla F_{\mathcal{S}}(\boldsymbol{x}) - \nabla F_{\mathcal{S}}(\boldsymbol{y})\|_2^2. \qquad (41)\end{aligned}$$

Finally, by the conditions of Definition 4.6, $F_{\mathcal{S}}$ is convex. By Assumption 4.2 and Remark 4.3, $F_{\mathcal{S}}$ is L_f-smooth. The gradient of $F_{\mathcal{S}}$ is $1/L_f$-cocoercive, namely

$$\langle \nabla F_{\mathcal{S}}(\boldsymbol{x}) - \nabla F_{\mathcal{S}}(\boldsymbol{y}),\, \boldsymbol{x}-\boldsymbol{y}\rangle \geq \frac{1}{L_f}\|\nabla F_{\mathcal{S}}(\boldsymbol{x}) - \nabla F_{\mathcal{S}}(\boldsymbol{y})\|_2^2.$$

Substituting this estimate into Eq. (41) yields Eq. (38). Since $\eta_t \in (0, 2/L_f]$, the coefficient $\frac{2}{L_f} - \eta_t$ is nonnegative, and Eq. (39) follows immediately. □

Lemma F.8 (One-step local PŁ recursion for RSD)**.** *Suppose that $F_{\mathcal{S}}$ satisfies the local PŁ condition on a set $\mathcal{U}_0$ with parameter $\beta_{\mathcal{S}} > 0$, and suppose that $\{\boldsymbol{x} \in \mathbb{R}^d : F_{\mathcal{S}}(\boldsymbol{x}) \leq F_{\mathcal{S}}(\boldsymbol{x}^0)\} \subseteq \mathcal{U}_0$. Suppose also that Assumption 4.4 and Eqs. (3)–(5) hold. If $0 < \eta_t \leq 1/L_{\max}$ for all $t \geq 0$, then $\boldsymbol{x}^t \in \mathcal{U}_0$ for every realization and every $t \geq 0$, and*

$$\mathbb{E}_t\big[F_{\mathcal{S}}(\boldsymbol{x}^{t+1}) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\big] \leq \left[1 - 2\alpha\beta_{\mathcal{S}}\eta_t\left(1 - \frac{L_{\max}\eta_t}{2}\right)\right]_+ \big(F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\big), \qquad (42)$$

where $[a]_+ := \max\{a, 0\}$, and $\mathbb{E}_t[\cdot]$ denotes conditional expectation over the current draw.

Proof. We first verify that the iterates remain in the region where the local PŁ condition applies. By the update rule (4) and Assumption 4.4 with $\boldsymbol{z} = -\eta_t P_{i_t}^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}^t)$, we have

$$F_{\mathcal{S}}(\boldsymbol{x}^{t+1}) \leq F_{\mathcal{S}}(\boldsymbol{x}^t) - \eta_t\left(1 - \frac{L_{i_t}\eta_t}{2}\right)\left\|P_{i_t}^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}^t)\right\|_2^2. \qquad (43)$$

Since $L_{i_t} \leq L_{\max}$ and $\eta_t \leq 1/L_{\max}$, the coefficient in front of the squared norm is nonnegative. Hence $F_{\mathcal{S}}(\boldsymbol{x}^{t+1}) \leq F_{\mathcal{S}}(\boldsymbol{x}^t)$ pathwise. By induction, $F_{\mathcal{S}}(\boldsymbol{x}^t) \leq F_{\mathcal{S}}(\boldsymbol{x}^0), t \geq 0$. The assumed sublevel-set inclusion therefore gives $\boldsymbol{x}^t \in \mathcal{U}_0$ for all $t \geq 0$.

We now derive the empirical-residual recursion. Lemma F.3 gives

$$\mathbb{E}_t\big[F_{\mathcal{S}}(\boldsymbol{x}^{t+1})\big] \leq F_{\mathcal{S}}(\boldsymbol{x}^t) - \alpha\eta_t\left(1 - \frac{L_{\max}\eta_t}{2}\right)\left\|\nabla F_{\mathcal{S}}(\boldsymbol{x}^t)\right\|_2^2. \qquad (44)$$

Since $\boldsymbol{x}^t \in \mathcal{U}_0$, the local PŁ condition implies $\|\nabla F_{\mathcal{S}}(\boldsymbol{x}^t)\|_2^2 \geq 2\beta_{\mathcal{S}}\left(F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\right)$. Substituting this inequality into (44) and subtracting $F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})$ from both sides yields

$$\mathbb{E}_t\big[F_{\mathcal{S}}(\boldsymbol{x}^{t+1}) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\big] \leq \left(1 - 2\alpha\beta_{\mathcal{S}}\eta_t\left(1 - \frac{L_{\max}\eta_t}{2}\right)\right)\big(F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\big). \qquad (45)$$

Because the left-hand side and the empirical residual are nonnegative, the positive-part form in (42) follows immediately: if the multiplier in (45) is nonnegative, there is nothing to prove; if it is negative, then (45) forces both sides of the desired positive-part bound to be zero. □

Lemma F.9 (Stopping budget under local PŁ recursion)**.** *Suppose the assumptions of Lemma F.8 hold, and suppose Assumption 4.2 holds with $L_f > 0$. Let $T \geq 1$. If, for almost every dataset $\mathcal{S}$ satisfying $F_{\mathcal{S}}(\boldsymbol{x}^0) > F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})$, the horizon T satisfies*

$$2\alpha\beta_{\mathcal{S}} \sum_{t=0}^{T-1} \eta_t \left(1 - \frac{L_{\max}\eta_t}{2}\right) \geq \log\left(N^2 L_f \left[F_{\mathcal{S}}(\boldsymbol{x}^0) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\right]\right), \tag{46}$$

then, for almost every dataset $\mathcal{S}$,

$$\mathbb{E}_{\mathcal{A}|\mathcal{S}}\left[F_{\mathcal{S}}(\boldsymbol{x}^T) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\right] \leq \frac{1}{N^2 L_f}, \tag{47}$$

where $\mathbb{E}_{\mathcal{A}|\mathcal{S}}[\cdot]$ denotes expectation over the randomness of Algorithm 1, conditional on $\mathcal{S}$.

Proof. Fix a dataset $\mathcal{S}$. If $F_{\mathcal{S}}(\boldsymbol{x}^0) = F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})$, then Lemma F.8 and the nonnegativity of $F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})$ imply inductively that

$$F_{\mathcal{S}}(\boldsymbol{x}^t) = F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}}) \qquad \text{almost surely for all } t \geq 0.$$

This proves (47) when the initial point is an empirical minimizer. For the rest of the proof, we assume that the initial empirical residual is strictly positive.

Iterating Lemma F.8 and using the tower property, we obtain

$$\begin{aligned}\mathbb{E}_{\mathcal{A}|\mathcal{S}}\left[F_{\mathcal{S}}(\boldsymbol{x}^T) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\right] &\leq \prod_{t=0}^{T-1} \left[1 - 2\alpha\beta_{\mathcal{S}}\eta_t \left(1 - \frac{L_{\max}\eta_t}{2}\right)\right]_+ \left(F_{\mathcal{S}}(\boldsymbol{x}^0) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\right) \\ &\leq \exp\left(-2\alpha\beta_{\mathcal{S}} \sum_{t=0}^{T-1} \eta_t \left(1 - \frac{L_{\max}\eta_t}{2}\right)\right) \left(F_{\mathcal{S}}(\boldsymbol{x}^0) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\right), \end{aligned} \tag{48}$$

where the last step uses $[1-a]_+ \leq e^{-a}$ for $a \geq 0$. The budget condition (46) therefore yields

$$\mathbb{E}_{\mathcal{A}|\mathcal{S}}\left[F_{\mathcal{S}}(\boldsymbol{x}^T) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\right] \leq \frac{F_{\mathcal{S}}(\boldsymbol{x}^0) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})}{N^2 L_f \left[F_{\mathcal{S}}(\boldsymbol{x}^0) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\right]} = \frac{1}{N^2 L_f}.$$

This proves (47). □

Lemma F.10 (Hölder bound for inverse-PŁ weighted residuals)**.** *Consider the joint probability space of the dataset $\mathcal{S}$, an independent copy $\mathcal{S}'$, and the algorithmic randomness. Let $R_t := F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})$ be pathwise nonincreasing and suppose that, for some $\mathcal{S}$-measurable constants $c_{\mathcal{S},t} \geq 0$,*

$$\mathbb{E}_t[R_{t+1}] \leq [1 - c_{\mathcal{S},t}]_+ R_t, \qquad [a]_+ := \max\{a, 0\}. \tag{49}$$

Let $\beta^{\mathrm{pt}}_{\mathcal{S}^{(1)}}(\boldsymbol{x}^T)$ be a positive, possibly infinite, $(\mathcal{S}, \mathcal{S}', \mathcal{A})$-measurable random variable. Then, for every $p > 1$ with conjugate exponent $q = p/(p-1)$ and every $T \geq 1$,

$$\mathbb{E}\left[\frac{R_T}{\beta^{\mathrm{pt}}_{\mathcal{S}^{(1)}}(\boldsymbol{x}^T)}\right] \leq \left(\mathbb{E}\left[\exp\left(-\sum_{t=0}^{T-1} c_{\mathcal{S},t}\right) R_0^p\right]\right)^{1/p} \left(\mathbb{E}\left[\left(\beta^{\mathrm{pt}}_{\mathcal{S}^{(1)}}(\boldsymbol{x}^T)\right)^{-q}\right]\right)^{1/q}, \tag{50}$$

whenever the right-hand side is finite.

Proof. Since the residual is pathwise nonincreasing, $0 \leq R_{t+1} \leq R_t$ for every realized trajectory. Hence, for $p > 1$, $R_{t+1}^p \leq R_t^{p-1} R_{t+1}$. Taking conditional expectation and using (49), we obtain

$$\mathbb{E}_t[R_{t+1}^p] \leq R_t^{p-1} \mathbb{E}_t[R_{t+1}] \leq [1 - c_{\mathcal{S},t}]_+ R_t^p \leq e^{-c_{\mathcal{S},t}} R_t^p, \tag{51}$$

where the last step uses $[1-u]_+ \leq e^{-u}$ for all $u \geq 0$.

Since $c_{\mathcal{S},t}$ is $\mathcal{S}$-measurable, iterating (51) conditionally on $\mathcal{S}$ and applying the tower property gives

$$\mathbb{E}_{\mathcal{A}|\mathcal{S}}[R_T^p] \leq \exp\left(-\sum_{t=0}^{T-1} c_{\mathcal{S},t}\right) R_0^p. \tag{52}$$

Taking expectation over $\mathcal{S}$ gives the corresponding unconditional moment bound:

$$\mathbb{E}[R_T^p] \le \mathbb{E}\left[\exp\left(-\sum_{t=0}^{T-1} c_{\mathcal{S},t}\right) R_0^p\right]. \tag{53}$$

Finally, Hölder's inequality on the joint probability space of $(\mathcal{S}, \mathcal{S}', \mathcal{A})$ gives

$$\mathbb{E}\left[\frac{R_T}{\beta^{\mathrm{pt}}_{\mathcal{S}^{(1)}}(\boldsymbol{x}^T)}\right] \le \left(\mathbb{E}[R_T^p]\right)^{1/p} \left(\mathbb{E}\left[\left(\beta^{\mathrm{pt}}_{\mathcal{S}^{(1)}}(\boldsymbol{x}^T)\right)^{-q}\right]\right)^{1/q}. \tag{54}$$

Substituting (53) proves (50). □

G Detailed proofs of major theorems

G.1 Proofs for the convex case

G.1.1 Proof of Theorem C.1

Proof. Minimality gives $F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}}) \le F_{\mathcal{S}}(\boldsymbol{x}^t)$ for every $t \ge 0$. Since $\boldsymbol{x}^\star_{\mathcal{S}}$ depends only on $\mathcal{S}$, it is independent of the algorithmic randomness. Lemma F.4(i) applies with $\boldsymbol{x} = \boldsymbol{x}^\star_{\mathcal{S}}$, yielding an expected suboptimality bound in terms of the initial distance and empirical residual:

$$\mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^T) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\big] \le \frac{\mathbb{E}\big[\|\boldsymbol{x}^0 - \boldsymbol{x}^\star_{\mathcal{S}}\|_2^2\big] + 2\eta_0\, \mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^0) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\big]}{2\alpha \sum_{t=0}^{T-1} \eta_t}. \tag{55}$$

By Remark 4.3, $F_{\mathcal{S}}$ is L_f-smooth. Since $\boldsymbol{x}^\star_{\mathcal{S}}$ is an unconstrained minimizer, $\nabla F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}}) = 0$, and smoothness at $(\boldsymbol{x}^\star_{\mathcal{S}}, \boldsymbol{x}^0)$ bounds this initial residual by the squared distance to the minimizer:

$$F_{\mathcal{S}}(\boldsymbol{x}^0) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}}) \le \frac{L_f}{2}\|\boldsymbol{x}^0 - \boldsymbol{x}^\star_{\mathcal{S}}\|_2^2. \tag{56}$$

Substituting Eq. (56) into Eq. (55), we obtain

$$\begin{aligned}\mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^T) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\big] &\le \frac{\mathbb{E}\big[\|\boldsymbol{x}^0 - \boldsymbol{x}^\star_{\mathcal{S}}\|_2^2\big] + \eta_0 L_f\, \mathbb{E}\big[\|\boldsymbol{x}^0 - \boldsymbol{x}^\star_{\mathcal{S}}\|_2^2\big]}{2\alpha \sum_{t=0}^{T-1} \eta_t} \\ &= \frac{(1 + \eta_0 L_f)\, \mathbb{E}\big[\|\boldsymbol{x}^0 - \boldsymbol{x}^\star_{\mathcal{S}}\|_2^2\big]}{2\alpha \sum_{t=0}^{T-1} \eta_t},\end{aligned} \tag{57}$$

which is exactly the claimed bound. □

G.1.2 Proof of Corollary C.3

Proof. Theorem C.1 does not apply directly, since the step size now depends on the sampled block. For $\eta_t = 1/L_{i_t}$, write the existing weighted distance $D_{\mathcal{L}}(\boldsymbol{x}, \boldsymbol{x}^\star_{\mathcal{S}})$ as

$$\Phi(\boldsymbol{x}) := \frac{1}{2}\sum_{i=1}^{m} L_i \|U_i^\top(\boldsymbol{x} - \boldsymbol{x}^\star_{\mathcal{S}})\|_2^2.$$

Condition on $\mathcal{S}$ and on all algorithmic randomness up to iteration t, and let $\mathbb{E}_t[\cdot]$ denote conditional expectation over the current random block $i_t \sim \mathrm{Unif}([m])$. If $i_t = j$, then

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \frac{1}{L_j} U_j U_j^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}^t).$$

Block orthogonality implies that only the j-th block changes. Expanding Φ gives

$$\begin{aligned}\Phi(\boldsymbol{x}^{t+1}) &= \frac{1}{2}\sum_{i \ne j} L_i \|U_i^\top(\boldsymbol{x}^t - \boldsymbol{x}^\star_{\mathcal{S}})\|_2^2 + \frac{L_j}{2}\left\|U_j^\top(\boldsymbol{x}^t - \boldsymbol{x}^\star_{\mathcal{S}}) - \frac{1}{L_j} U_j^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}^t)\right\|_2^2 \\ &= \Phi(\boldsymbol{x}^t) - \left\langle U_j^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}^t), U_j^\top(\boldsymbol{x}^t - \boldsymbol{x}^\star_{\mathcal{S}})\right\rangle + \frac{1}{2L_j}\|U_j^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}^t)\|_2^2.\end{aligned} \tag{58}$$

Taking conditional expectation over the uniform choice of j and using $\sum_{j=1}^{m} U_j U_j^\top = I_d$, we obtain

$$\mathbb{E}_t[\Phi(\boldsymbol{x}^{t+1})] = \Phi(\boldsymbol{x}^t) - \frac{1}{m}\left\langle \nabla F_{\mathcal{S}}(\boldsymbol{x}^t), \boldsymbol{x}^t - \boldsymbol{x}_{\mathcal{S}}^\star \right\rangle + \frac{1}{2m}\sum_{j=1}^{m} \frac{1}{L_j}\|U_j^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}^t)\|_2^2. \tag{59}$$

For each block j, Assumption 4.4 with $P_j = U_j$, applied at $\boldsymbol{x}^t$ in the direction $-\frac{1}{L_j}U_j^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}^t)$ gives the one-step descent estimate for the sampled block:

$$\begin{aligned} F_{\mathcal{S}}\left(\boldsymbol{x}^t - \frac{1}{L_j}U_j U_j^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}^t)\right) &\le F_{\mathcal{S}}(\boldsymbol{x}^t) - \frac{1}{L_j}\|U_j^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}^t)\|_2^2 + \frac{L_j}{2}\left\|\frac{1}{L_j}U_j^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}^t)\right\|_2^2 \\ &= F_{\mathcal{S}}(\boldsymbol{x}^t) - \frac{1}{2L_j}\|U_j^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}^t)\|_2^2. \end{aligned} \tag{60}$$

When the current sampled block is $i_t = j$, the point on the left-hand side is precisely $\boldsymbol{x}^{t+1}$. Hence, after averaging the preceding inequality over the uniform block choice,

$$\mathbb{E}_t[F_{\mathcal{S}}(\boldsymbol{x}^{t+1})] \le F_{\mathcal{S}}(\boldsymbol{x}^t) - \frac{1}{2m}\sum_{j=1}^{m} \frac{1}{L_j}\|U_j^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}^t)\|_2^2, \tag{61}$$

or equivalently,

$$\frac{1}{2m}\sum_{j=1}^{m} \frac{1}{L_j}\|U_j^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}^t)\|_2^2 \le F_{\mathcal{S}}(\boldsymbol{x}^t) - \mathbb{E}_t[F_{\mathcal{S}}(\boldsymbol{x}^{t+1})]. \tag{62}$$

Moreover, by convexity of $F_{\mathcal{S}}$,

$$\left\langle \nabla F_{\mathcal{S}}(\boldsymbol{x}^t), \boldsymbol{x}^t - \boldsymbol{x}_{\mathcal{S}}^\star \right\rangle \ge F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}}^\star). \tag{63}$$

Substituting (62) and (63) into (59), we get

$$\mathbb{E}_t[\Phi(\boldsymbol{x}^{t+1})] \le \Phi(\boldsymbol{x}^t) - \frac{1}{m}\Big(F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}}^\star)\Big) + F_{\mathcal{S}}(\boldsymbol{x}^t) - \mathbb{E}_t[F_{\mathcal{S}}(\boldsymbol{x}^{t+1})]. \tag{64}$$

Rearranging this inequality yields the one-step Lyapunov recursion

$$\begin{aligned} \frac{1}{m}\Big(F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}}^\star)\Big) &\le \Phi(\boldsymbol{x}^t) + F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}}^\star) \\ &\quad - \mathbb{E}_t\big[\Phi(\boldsymbol{x}^{t+1}) + F_{\mathcal{S}}(\boldsymbol{x}^{t+1}) - F_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}}^\star)\big]. \end{aligned} \tag{65}$$

Taking total expectation in (65) and summing from $t = 0$ to $T - 1$, the right-hand side telescopes:

$$\begin{aligned} \frac{1}{m}\sum_{t=0}^{T-1} \mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}}^\star)\big] &\le \mathbb{E}\big[\Phi(\boldsymbol{x}^0) + F_{\mathcal{S}}(\boldsymbol{x}^0) - F_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}}^\star)\big] \\ &\quad - \mathbb{E}\big[\Phi(\boldsymbol{x}^T) + F_{\mathcal{S}}(\boldsymbol{x}^T) - F_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}}^\star)\big]. \end{aligned} \tag{66}$$

Since $\Phi(\boldsymbol{x}^T) \ge 0$ by definition and $F_{\mathcal{S}}(\boldsymbol{x}^T) - F_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}}^\star) \ge 0$ by optimality of $\boldsymbol{x}_{\mathcal{S}}^\star$, the terminal term is nonnegative. Thus

$$\frac{1}{m}\sum_{t=0}^{T-1} \mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}}^\star)\big] \le \mathbb{E}[\Phi(\boldsymbol{x}^0)] + \mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^0) - F_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}}^\star)\big]. \tag{67}$$

On the other hand, (60) shows that each block update is a pathwise descent step:

$$F_{\mathcal{S}}(\boldsymbol{x}^{t+1}) \le F_{\mathcal{S}}(\boldsymbol{x}^t) - \frac{1}{2L_{i_t}}\|U_{i_t}^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}^t)\|_2^2 \le F_{\mathcal{S}}(\boldsymbol{x}^t).$$

Therefore, for every realization, $F_{\mathcal{S}}(\boldsymbol{x}^T) - F_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}}^\star) \le F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}}^\star), t = 0, \ldots, T-1$. Taking expectation, summing over $t = 0, \ldots, T-1$, and using the preceding inequality gives

$$T\,\mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^T) - F_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}}^\star)\big] \le \sum_{t=0}^{T-1} \mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}}^\star)\big]. \tag{68}$$

Combining (67) and (68), we obtain

$$\mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^T) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\big] \le \frac{m}{T}\left(\mathbb{E}[\Phi(\boldsymbol{x}^0)] + \mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^0) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\big]\right). \tag{69}$$

Finally, by Remark 4.3, $F_{\mathcal{S}}$ is L_f-smooth. Since $\boldsymbol{x}^\star_{\mathcal{S}}$ is an unconstrained minimizer, $\nabla F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}}) = 0$. Smoothness at $\boldsymbol{x}^\star_{\mathcal{S}}$ bounds the initial excess risk by the squared distance to this minimizer:

$$\begin{aligned} F_{\mathcal{S}}(\boldsymbol{x}^0) &\le F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}}) + \big\langle \nabla F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}}), \boldsymbol{x}^0 - \boldsymbol{x}^\star_{\mathcal{S}}\big\rangle + \frac{L_f}{2}\|\boldsymbol{x}^0 - \boldsymbol{x}^\star_{\mathcal{S}}\|_2^2 \\ &= F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}}) + \frac{L_f}{2}\|\boldsymbol{x}^0 - \boldsymbol{x}^\star_{\mathcal{S}}\|_2^2, \end{aligned} \tag{70}$$

and thus

$$F_{\mathcal{S}}(\boldsymbol{x}^0) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}}) \le \frac{L_f}{2}\|\boldsymbol{x}^0 - \boldsymbol{x}^\star_{\mathcal{S}}\|_2^2. \tag{71}$$

Substituting (71) and the definition of $\Phi(\boldsymbol{x}^0)$ into (69) yields

$$\mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^T) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\big] \le \frac{m}{T}\left(\frac{1}{2}\sum_{i=1}^m L_i\,\mathbb{E}\big[\|U_i^\top(\boldsymbol{x}^0 - \boldsymbol{x}^\star_{\mathcal{S}})\|_2^2\big] + \frac{L_f}{2}\mathbb{E}\big[\|\boldsymbol{x}^0 - \boldsymbol{x}^\star_{\mathcal{S}}\|_2^2\big]\right), \tag{72}$$

which is the claimed bound. □

G.1.3 Proof of Theorem 4.7

Proof. Fix a deterministic comparator $\boldsymbol{x} \in \mathbb{R}^d$. For each $n \in [N]$, define

$$\Delta_t^{(n)} := \boldsymbol{x}^t - \boldsymbol{x}^{(n),t}, \qquad \widetilde{F}_{\mathcal{S},-n}(\boldsymbol{z}) := \frac{1}{N}\sum_{k\neq n} f(\boldsymbol{z};\xi_k).$$

Then

$$F_{\mathcal{S}}(\boldsymbol{z}) = \widetilde{F}_{\mathcal{S},-n}(\boldsymbol{z}) + \frac{1}{N}f(\boldsymbol{z};\xi_n), \qquad F_{\mathcal{S}^{(n)}}(\boldsymbol{z}) = \widetilde{F}_{\mathcal{S},-n}(\boldsymbol{z}) + \frac{1}{N}f(\boldsymbol{z};\xi'_n).$$

We couple the two runs on $\mathcal{S}$ and $\mathcal{S}^{(n)}$ by using the same sampled type i_t at each iteration and, whenever P_{i_t} is random, the same realization of P_{i_t}. By the RSD update (4),

$$\begin{aligned} \Delta_{t+1}^{(n)} = \Delta_t^{(n)} &- \eta_t P_{i_t} P_{i_t}^\top\Big(\nabla\widetilde{F}_{\mathcal{S},-n}(\boldsymbol{x}^t) - \nabla\widetilde{F}_{\mathcal{S},-n}(\boldsymbol{x}^{(n),t})\Big) \\ &- \frac{\eta_t}{N}P_{i_t}P_{i_t}^\top\Big(\nabla f(\boldsymbol{x}^t;\xi_n) - \nabla f(\boldsymbol{x}^{(n),t};\xi'_n)\Big). \end{aligned} \tag{73}$$

Let $\mathcal{F}_t$ be the sigma-field generated by $\mathcal{S}$, $\mathcal{S}'$, and all algorithmic randomness strictly before the current draw at iteration t, and write $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_t]$. Conditional on $\mathcal{F}_t$, the iterates $\boldsymbol{x}^t$ and $\boldsymbol{x}^{(n),t}$ are fixed, whereas the current random choice of (i_t, P_{i_t}) is independent of $\mathcal{F}_t$.

By Definition 4.6 and linearity of averaging, $\widetilde{F}_{\mathcal{S},-n}$ is convex. Moreover, Assumption 4.2 implies that $\widetilde{F}_{\mathcal{S},-n}$ is $(N-1)L_f/N$-smooth, and hence L_f-smooth. Therefore Lemma F.7 applies to the common empirical part in (73). Since $\eta_t \le 1/L_f \le 2/L_f$, we have

$$\mathbb{E}_t\left[\left\|\Delta_t^{(n)} - \eta_t P_{i_t}P_{i_t}^\top\Big(\nabla\widetilde{F}_{\mathcal{S},-n}(\boldsymbol{x}^t) - \nabla\widetilde{F}_{\mathcal{S},-n}(\boldsymbol{x}^{(n),t})\Big)\right\|_2^2\right] \le \|\Delta_t^{(n)}\|_2^2. \tag{74}$$

Combining (73) and (74), and using $\|\boldsymbol{u}-\boldsymbol{v}\|_2^2 \le \|\boldsymbol{u}\|_2^2 + \|\boldsymbol{v}\|_2^2 + 2\|\boldsymbol{u}\|_2\|\boldsymbol{v}\|_2$ together with conditional Cauchy–Schwarz for the mixed product, yields

$$\begin{aligned} \mathbb{E}_t\Big[\|\Delta_{t+1}^{(n)}\|_2^2\Big] \le \|\Delta_t^{(n)}\|_2^2 &+ \frac{\eta_t^2}{N^2}\mathbb{E}_t\left[\left\|P_{i_t}P_{i_t}^\top\Big(\nabla f(\boldsymbol{x}^t;\xi_n) - \nabla f(\boldsymbol{x}^{(n),t};\xi'_n)\Big)\right\|_2^2\right] \\ &+ \frac{2\eta_t}{N}\|\Delta_t^{(n)}\|_2\left(\mathbb{E}_t\left[\left\|P_{i_t}P_{i_t}^\top\Big(\nabla f(\boldsymbol{x}^t;\xi_n) - \nabla f(\boldsymbol{x}^{(n),t};\xi'_n)\Big)\right\|_2^2\right]\right)^{1/2}. \end{aligned} \tag{75}$$

Indeed, the mixed term is bounded by the square root of the conditional second moment of the common part, which is at most $\|\Delta_t^{(n)}\|_2$ by (74).

By (3), $P_{i_t}P_{i_t}^\top$ is an orthogonal projector, and hence $\|P_{i_t}P_{i_t}^\top v\|_2^2 = v^\top P_{i_t}P_{i_t}^\top v$. Together with (5), this gives

$$\mathbb{E}_t\left[\left\|P_{i_t}P_{i_t}^\top\Big(\nabla f(\boldsymbol{x}^t;\xi_n)-\nabla f(\boldsymbol{x}^{(n),t};\xi_n')\Big)\right\|_2^2\right] = \alpha\left\|\nabla f(\boldsymbol{x}^t;\xi_n)-\nabla f(\boldsymbol{x}^{(n),t};\xi_n')\right\|_2^2. \tag{76}$$

Moreover, by $\|\boldsymbol{a}-\boldsymbol{b}\|_2^2 \le 2\|\boldsymbol{a}\|_2^2 + 2\|\boldsymbol{b}\|_2^2$ and Lemma F.5,

$$\begin{aligned}\left\|\nabla f(\boldsymbol{x}^t;\xi_n)-\nabla f(\boldsymbol{x}^{(n),t};\xi_n')\right\|_2^2 &\le 2\|\nabla f(\boldsymbol{x}^t;\xi_n)\|_2^2 + 2\|\nabla f(\boldsymbol{x}^{(n),t};\xi_n')\|_2^2\\ &\le 4L_f\Big(f(\boldsymbol{x}^t;\xi_n)+f(\boldsymbol{x}^{(n),t};\xi_n')\Big).\end{aligned} \tag{77}$$

Substituting (76) and (77) into (75), taking full expectation, and applying Cauchy–Schwarz to the remaining mixed term yield

$$\begin{aligned}\mathbb{E}\left[\|\Delta_{t+1}^{(n)}\|_2^2\right] \le \mathbb{E}\left[\|\Delta_t^{(n)}\|_2^2\right] &+ \frac{4\alpha L_f\eta_t^2}{N^2}\mathbb{E}\Big[f(\boldsymbol{x}^t;\xi_n)+f(\boldsymbol{x}^{(n),t};\xi_n')\Big]\\ &+ \frac{4\eta_t\sqrt{\alpha L_f}}{N}\Big(\mathbb{E}\left[\|\Delta_t^{(n)}\|_2^2\right]\Big)^{1/2}\Big(\mathbb{E}\Big[f(\boldsymbol{x}^t;\xi_n)+f(\boldsymbol{x}^{(n),t};\xi_n')\Big]\Big)^{1/2}.\end{aligned} \tag{78}$$

Define

$$\bar\Delta_t^2 := \frac{1}{N}\sum_{n=1}^N \mathbb{E}\left[\|\Delta_t^{(n)}\|_2^2\right], \qquad \bar\Delta_t := (\bar\Delta_t^2)^{1/2}.$$

Averaging (78) over $n=1,\dots,N$ and using Cauchy–Schwarz for the finite sum gives

$$\begin{aligned}\bar\Delta_{t+1}^2 \le \bar\Delta_t^2 &+ \frac{4\alpha L_f\eta_t^2}{N^2}\frac{1}{N}\sum_{n=1}^N\mathbb{E}\Big[f(\boldsymbol{x}^t;\xi_n)+f(\boldsymbol{x}^{(n),t};\xi_n')\Big]\\ &+ \frac{4\eta_t\sqrt{\alpha L_f}}{N}\bar\Delta_t\left(\frac{1}{N}\sum_{n=1}^N\mathbb{E}\Big[f(\boldsymbol{x}^t;\xi_n)+f(\boldsymbol{x}^{(n),t};\xi_n')\Big]\right)^{1/2}.\end{aligned} \tag{79}$$

The first average satisfies $\frac{1}{N}\sum_{n=1}^N\mathbb{E}[f(\boldsymbol{x}^t;\xi_n)] = \mathbb{E}[F_{\mathcal{S}}(\boldsymbol{x}^t)]$. For the second average, by exchangeability of the i.i.d. samples, $(\mathcal{S}^{(n)},\xi_n')$ has the same distribution as $(\mathcal{S},\xi_n)$ for each fixed n. Since the algorithmic randomness has the same law in the two coupled runs, $\mathbb{E}[f(\boldsymbol{x}^{(n),t};\xi_n')] = \mathbb{E}[f(\boldsymbol{x}^t;\xi_n)]$.

$$\frac{1}{N}\sum_{n=1}^N\mathbb{E}\Big[f(\boldsymbol{x}^t;\xi_n)+f(\boldsymbol{x}^{(n),t};\xi_n')\Big] = 2\,\mathbb{E}[F_{\mathcal{S}}(\boldsymbol{x}^t)]. \tag{80}$$

Substituting (80) into (79), and using Assumption 4.1 to ensure $\mathbb{E}[F_{\mathcal{S}}(\boldsymbol{x}^t)]\ge 0$, gives

$$\begin{aligned}\bar\Delta_{t+1}^2 &\le \bar\Delta_t^2 + \frac{8\alpha L_f\eta_t^2}{N^2}\mathbb{E}[F_{\mathcal{S}}(\boldsymbol{x}^t)] + \frac{4\sqrt{2\alpha L_f}\,\eta_t}{N}\bar\Delta_t\Big(\mathbb{E}[F_{\mathcal{S}}(\boldsymbol{x}^t)]\Big)^{1/2}\\ &= \left(\bar\Delta_t + \frac{2\sqrt{2\alpha L_f}\,\eta_t}{N}\Big(\mathbb{E}[F_{\mathcal{S}}(\boldsymbol{x}^t)]\Big)^{1/2}\right)^2.\end{aligned} \tag{81}$$

Taking square roots and using $\bar\Delta_0 = 0$,

$$\bar\Delta_T \le \frac{2\sqrt{2\alpha L_f}}{N}\sum_{t=0}^{T-1}\eta_t\Big(\mathbb{E}[F_{\mathcal{S}}(\boldsymbol{x}^t)]\Big)^{1/2}. \tag{82}$$

By Cauchy–Schwarz,

$$\bar\Delta_T^2 \le \frac{8\alpha L_f T}{N^2}\sum_{t=0}^{T-1}\eta_t^2\,\mathbb{E}[F_{\mathcal{S}}(\boldsymbol{x}^t)]. \tag{83}$$

It remains to control the weighted empirical-risk sum. Since $\eta_t \le 1/L_f$ and, by Remark 4.5, $L_{\max} \le L_f$, we have $\eta_t \le 1/L_{\max}$. Together with the assumed monotonicity of the step sizes, part (ii) of Lemma F.4 applies to the fixed comparator $\boldsymbol{x}$ and gives

$$2\alpha \sum_{t=0}^{T-1} \eta_t^2 \mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x})\big] \le \eta_0 \|\boldsymbol{x}^0 - \boldsymbol{x}\|_2^2 + 2\eta_0^2\, \mathbb{E}[F_{\mathcal{S}}(\boldsymbol{x}^0)]. \tag{84}$$

Thus,

$$\begin{aligned}\sum_{t=0}^{T-1} \eta_t^2 \mathbb{E}[F_{\mathcal{S}}(\boldsymbol{x}^t)] &= \sum_{t=0}^{T-1} \eta_t^2 \mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x})\big] + \left(\sum_{t=0}^{T-1} \eta_t^2\right) \mathbb{E}[F_{\mathcal{S}}(\boldsymbol{x})] \\ &\le \frac{\eta_0 \|\boldsymbol{x}^0 - \boldsymbol{x}\|_2^2 + 2\eta_0^2\, \mathbb{E}[F_{\mathcal{S}}(\boldsymbol{x}^0)]}{2\alpha} + \left(\sum_{t=0}^{T-1} \eta_t^2\right) \mathbb{E}[F_{\mathcal{S}}(\boldsymbol{x})]. \end{aligned} \tag{85}$$

Since $\boldsymbol{x}^0$ and $\boldsymbol{x}$ are deterministic and independent of $\mathcal{S}$, $\mathbb{E}[F_{\mathcal{S}}(\boldsymbol{x}^0)] = F(\boldsymbol{x}^0)$, $\mathbb{E}[F_{\mathcal{S}}(\boldsymbol{x})] = F(\boldsymbol{x})$. Substituting (85) into (83) yields

$$\bar{\Delta}_T^2 \le \frac{4L_f T}{N^2} \left[\eta_0 \|\boldsymbol{x}^0 - \boldsymbol{x}\|_2^2 + 2\eta_0^2 F(\boldsymbol{x}^0) + 2\alpha \left(\sum_{t=0}^{T-1} \eta_t^2\right) F(\boldsymbol{x})\right]. \tag{86}$$

Taking square roots proves the claimed on-average argument stability bound. □

G.1.4 Proof of Corollary 4.9

Proof. We bound the weighted perturbation, use descent to sum the risk terms, and return to Euclidean distance. For each $n \in [N]$, let

$$\Delta_t^{(n)} := \boldsymbol{x}^t - \boldsymbol{x}^{(n),t}, \qquad \mathcal{E}_t^{(n)} := \frac{1}{2} \sum_{i=1}^{m} L_i \|U_i^\top \Delta_t^{(n)}\|_2^2,$$

under the same coupling as in Theorem 4.7, and define

$$\bar{\mathcal{E}}_t := \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}[\mathcal{E}_t^{(n)}].$$

Since the two coupled runs start from the same deterministic initialization, $\bar{\mathcal{E}}_0 = 0$. As in the proof of Theorem 4.7, define the common empirical part $\widetilde{F}_{\mathcal{S},-n}(\boldsymbol{z}) := \frac{1}{N} \sum_{k \ne n} f(\boldsymbol{z}; \xi_k)$, so that

$$F_{\mathcal{S}}(\boldsymbol{z}) = \widetilde{F}_{\mathcal{S},-n}(\boldsymbol{z}) + \frac{1}{N} f(\boldsymbol{z}; \xi_n), \qquad F_{\mathcal{S}^{(n)}}(\boldsymbol{z}) = \widetilde{F}_{\mathcal{S},-n}(\boldsymbol{z}) + \frac{1}{N} f(\boldsymbol{z}; \xi_n').$$

Condition on the sigma-field generated by $\mathcal{S}$, $\mathcal{S}'$, and all algorithmic randomness strictly before the current draw at iteration t. Let $\mathbb{E}_t[\cdot]$ denote conditional expectation over the current uniformly sampled block $i_t \in [m]$. On the event $\{i_t = j\}$, the coupled update with blockwise step sizes gives

$$\begin{aligned}\Delta_{t+1}^{(n)} = \Delta_t^{(n)} - \frac{1}{L_j} U_j U_j^\top \Big[&\nabla \widetilde{F}_{\mathcal{S},-n}(\boldsymbol{x}^t) - \nabla \widetilde{F}_{\mathcal{S},-n}(\boldsymbol{x}^{(n),t}) \\ &+ \frac{1}{N} \Big(\nabla f(\boldsymbol{x}^t; \xi_n) - \nabla f(\boldsymbol{x}^{(n),t}; \xi_n')\Big)\Big]. \end{aligned} \tag{87}$$

Expanding $\mathcal{E}_{t+1}^{(n)}$ blockwise, using the orthogonal block decomposition $\sum_{i=1}^{m} U_i U_i^\top = I_d$, and averaging over $j \sim \mathrm{Unif}([m])$, we obtain

$$\begin{aligned}\mathbb{E}_t[\mathcal{E}_{t+1}^{(n)}] = \mathcal{E}_t^{(n)} &- \frac{1}{m} \Big\langle \nabla \widetilde{F}_{\mathcal{S},-n}(\boldsymbol{x}^t) - \nabla \widetilde{F}_{\mathcal{S},-n}(\boldsymbol{x}^{(n),t}), \Delta_t^{(n)} \Big\rangle \\ &- \frac{1}{mN} \Big\langle \nabla f(\boldsymbol{x}^t; \xi_n) - \nabla f(\boldsymbol{x}^{(n),t}; \xi_n'), \Delta_t^{(n)} \Big\rangle \end{aligned}$$

$$+\frac{1}{2m}\sum_{j=1}^{m}\frac{1}{L_j}\Bigg\|U_j^\top\Bigg[\nabla\widetilde{F}_{\mathcal{S},-n}(\boldsymbol{x}^t)-\nabla\widetilde{F}_{\mathcal{S},-n}(\boldsymbol{x}^{(n),t}) + \frac{1}{N}\Big(\nabla f(\boldsymbol{x}^t;\xi_n)-\nabla f(\boldsymbol{x}^{(n),t};\xi_n')\Big)\Bigg]\Bigg\|_2^2. \tag{88}$$

By Definition 4.6 and linearity of averaging, $\widetilde{F}_{\mathcal{S},-n}$ is convex. Moreover, Assumption 4.2 implies that $\widetilde{F}_{\mathcal{S},-n}$ is $(N-1)L_f/N$-smooth, and hence L_f-smooth. By the co-coercivity of gradients,

$$\Big\langle\nabla\widetilde{F}_{\mathcal{S},-n}(\boldsymbol{x}^t)-\nabla\widetilde{F}_{\mathcal{S},-n}(\boldsymbol{x}^{(n),t}),\Delta_t^{(n)}\Big\rangle\geq\frac{1}{L_f}\Big\|\nabla\widetilde{F}_{\mathcal{S},-n}(\boldsymbol{x}^t)-\nabla\widetilde{F}_{\mathcal{S},-n}(\boldsymbol{x}^{(n),t})\Big\|_2^2. \tag{89}$$

Together with $L_j\geq L_{\min}$, this gives

$$\begin{aligned}&-\frac{1}{m}\Big\langle\nabla\widetilde{F}_{\mathcal{S},-n}(\boldsymbol{x}^t)-\nabla\widetilde{F}_{\mathcal{S},-n}(\boldsymbol{x}^{(n),t}),\Delta_t^{(n)}\Big\rangle\\&\quad+\frac{1}{2m}\sum_{j=1}^{m}\frac{1}{L_j}\Big\|U_j^\top\Big(\nabla\widetilde{F}_{\mathcal{S},-n}(\boldsymbol{x}^t)-\nabla\widetilde{F}_{\mathcal{S},-n}(\boldsymbol{x}^{(n),t})\Big)\Big\|_2^2\\&\leq-\frac{2L_{\min}-L_f}{2mL_fL_{\min}}\Big\|\nabla\widetilde{F}_{\mathcal{S},-n}(\boldsymbol{x}^t)-\nabla\widetilde{F}_{\mathcal{S},-n}(\boldsymbol{x}^{(n),t})\Big\|_2^2.\end{aligned} \tag{90}$$

For the cross term between the common empirical part and the single-sample perturbation, Cauchy–Schwarz and $L_j\geq L_{\min}$ imply

$$\begin{aligned}&\frac{1}{mN}\sum_{j=1}^{m}\frac{1}{L_j}\Big\langle U_j^\top\Big(\nabla\widetilde{F}_{\mathcal{S},-n}(\boldsymbol{x}^t)-\nabla\widetilde{F}_{\mathcal{S},-n}(\boldsymbol{x}^{(n),t})\Big),U_j^\top\Big(\nabla f(\boldsymbol{x}^t;\xi_n)-\nabla f(\boldsymbol{x}^{(n),t};\xi_n')\Big)\Big\rangle\\&\leq\frac{1}{mNL_{\min}}\Big\|\nabla\widetilde{F}_{\mathcal{S},-n}(\boldsymbol{x}^t)-\nabla\widetilde{F}_{\mathcal{S},-n}(\boldsymbol{x}^{(n),t})\Big\|_2\Big\|\nabla f(\boldsymbol{x}^t;\xi_n)-\nabla f(\boldsymbol{x}^{(n),t};\xi_n')\Big\|_2.\end{aligned} \tag{91}$$

Applying Young's inequality to absorb the mixed term into the negative drift gives

$$\begin{aligned}&\frac{1}{mN}\sum_{j=1}^{m}\frac{1}{L_j}\Big\langle U_j^\top\Big(\nabla\widetilde{F}_{\mathcal{S},-n}(\boldsymbol{x}^t)-\nabla\widetilde{F}_{\mathcal{S},-n}(\boldsymbol{x}^{(n),t})\Big),U_j^\top\Big(\nabla f(\boldsymbol{x}^t;\xi_n)-\nabla f(\boldsymbol{x}^{(n),t};\xi_n')\Big)\Big\rangle\\&\leq\frac{2L_{\min}-L_f}{2mL_fL_{\min}}\Big\|\nabla\widetilde{F}_{\mathcal{S},-n}(\boldsymbol{x}^t)-\nabla\widetilde{F}_{\mathcal{S},-n}(\boldsymbol{x}^{(n),t})\Big\|_2^2\\&\quad+\frac{L_f}{2mN^2L_{\min}(2L_{\min}-L_f)}\Big\|\nabla f(\boldsymbol{x}^t;\xi_n)-\nabla f(\boldsymbol{x}^{(n),t};\xi_n')\Big\|_2^2.\end{aligned} \tag{92}$$

The first term on the right-hand side of (92) cancels the negative common-part term in (90).

The remaining quadratic perturbation term in the expansion satisfies

$$\begin{aligned}&\frac{1}{2mN^2}\sum_{j=1}^{m}\frac{1}{L_j}\Big\|U_j^\top\Big(\nabla f(\boldsymbol{x}^t;\xi_n)-\nabla f(\boldsymbol{x}^{(n),t};\xi_n')\Big)\Big\|_2^2\\&\leq\frac{1}{2mN^2L_{\min}}\Big\|\nabla f(\boldsymbol{x}^t;\xi_n)-\nabla f(\boldsymbol{x}^{(n),t};\xi_n')\Big\|_2^2.\end{aligned} \tag{93}$$

The linear perturbation term is controlled by the weighted energy itself. Since $\mathcal{E}_t^{(n)}\geq\frac{L_{\min}}{2}\|\Delta_t^{(n)}\|_2^2$, we have

$$\begin{aligned}&\Big|-\frac{1}{mN}\Big\langle\nabla f(\boldsymbol{x}^t;\xi_n)-\nabla f(\boldsymbol{x}^{(n),t};\xi_n'),\Delta_t^{(n)}\Big\rangle\Big|\\&\leq\frac{\sqrt{2}}{mN\sqrt{L_{\min}}}\big(\mathcal{E}_t^{(n)}\big)^{1/2}\Big\|\nabla f(\boldsymbol{x}^t;\xi_n)-\nabla f(\boldsymbol{x}^{(n),t};\xi_n')\Big\|_2.\end{aligned} \tag{94}$$

Combining (88)–(94) and using $\frac{L_f}{2L_{\min}(2L_{\min}-L_f)} + \frac{1}{2L_{\min}} = \frac{1}{2L_{\min}-L_f}$, we obtain the one-step bound

$$\mathbb{E}_t[\mathcal{E}_{t+1}^{(n)}] \le \mathcal{E}_t^{(n)} + \frac{\sqrt{2}}{mN\sqrt{L_{\min}}}\big(\mathcal{E}_t^{(n)}\big)^{1/2}\left\|\nabla f(\boldsymbol{x}^t;\xi_n) - \nabla f(\boldsymbol{x}^{(n),t};\xi_n')\right\|_2 + \frac{1}{mN^2(2L_{\min}-L_f)}\left\|\nabla f(\boldsymbol{x}^t;\xi_n) - \nabla f(\boldsymbol{x}^{(n),t};\xi_n')\right\|_2^2. \tag{95}$$

Taking total expectation and averaging over n, Lemma F.5 gives

$$\left\|\nabla f(\boldsymbol{x}^t;\xi_n) - \nabla f(\boldsymbol{x}^{(n),t};\xi_n')\right\|_2^2 \le 4L_f\Big(f(\boldsymbol{x}^t;\xi_n) + f(\boldsymbol{x}^{(n),t};\xi_n')\Big). \tag{96}$$

By exchangeability of the samples and the common law of the algorithmic randomness in both runs,

$$\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[f(\boldsymbol{x}^t;\xi_n) + f(\boldsymbol{x}^{(n),t};\xi_n')\Big] = 2\,\mathbb{E}[F_{\mathcal{S}}(\boldsymbol{x}^t)].$$

Therefore, averaging (95) over $n = 1,\dots,N$ and applying Cauchy–Schwarz to the mixed term yield a recurrence for the coupled trajectories in the curvature-weighted distance:

$$\bar{\mathcal{E}}_{t+1} \le \bar{\mathcal{E}}_t + \frac{4\sqrt{L_f}}{mN\sqrt{L_{\min}}}\bar{\mathcal{E}}_t^{1/2}\big(\mathbb{E}[F_{\mathcal{S}}(\boldsymbol{x}^t)]\big)^{1/2} + \frac{8L_f}{mN^2(2L_{\min}-L_f)}\mathbb{E}[F_{\mathcal{S}}(\boldsymbol{x}^t)]. \tag{97}$$

Since $\left(\frac{4\sqrt{L_f}}{mN\sqrt{L_{\min}}}\right)^2 = \frac{16L_f}{m^2N^2L_{\min}} \le 4\cdot\frac{8L_f}{mN^2(2L_{\min}-L_f)}$, where we used $m \ge 1$ and $2L_{\min}-L_f \le 2L_{\min}$, the right-hand side of (97) is bounded by a perfect square:

$$\bar{\mathcal{E}}_{t+1} \le \left(\bar{\mathcal{E}}_t^{1/2} + \left[\frac{8L_f}{mN^2(2L_{\min}-L_f)}\mathbb{E}[F_{\mathcal{S}}(\boldsymbol{x}^t)]\right]^{1/2}\right)^2. \tag{98}$$

Taking square roots, summing over $t = 0,\dots,T-1$, using $\bar{\mathcal{E}}_0 = 0$, and applying Cauchy–Schwarz gives the accumulated empirical-risk contribution to the trajectory discrepancy:

$$\bar{\mathcal{E}}_T \le \frac{8L_fT}{mN^2(2L_{\min}-L_f)}\sum_{t=0}^{T-1}\mathbb{E}[F_{\mathcal{S}}(\boldsymbol{x}^t)]. \tag{99}$$

It remains to control the empirical-risk sum. Fix a deterministic comparator $\boldsymbol{x} \in \mathbb{R}^d$, independent of $\mathcal{S}$, $\mathcal{S}'$, and the algorithmic randomness, and define

$$D_{\mathcal{L}}(\boldsymbol{z},\boldsymbol{x}) := \frac{1}{2}\sum_{i=1}^{m}L_i\|U_i^\top(\boldsymbol{z}-\boldsymbol{x})\|_2^2.$$

On the event $\{i_t = j\}$, the blockwise step satisfies $\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \frac{1}{L_j}U_jU_j^\top\nabla F_{\mathcal{S}}(\boldsymbol{x}^t)$. Expanding $D_{\mathcal{L}}(\boldsymbol{x}^{t+1},\boldsymbol{x})$ along this update and averaging over the uniformly sampled block gives

$$\mathbb{E}_t[D_{\mathcal{L}}(\boldsymbol{x}^{t+1},\boldsymbol{x})] = D_{\mathcal{L}}(\boldsymbol{x}^t,\boldsymbol{x}) - \frac{1}{m}\left\langle\nabla F_{\mathcal{S}}(\boldsymbol{x}^t),\boldsymbol{x}^t-\boldsymbol{x}\right\rangle + \frac{1}{2m}\sum_{j=1}^{m}\frac{1}{L_j}\|U_j^\top\nabla F_{\mathcal{S}}(\boldsymbol{x}^t)\|_2^2. \tag{100}$$

By Assumption 4.4, the same blockwise step also satisfies

$$\mathbb{E}_t[F_{\mathcal{S}}(\boldsymbol{x}^{t+1})] \le F_{\mathcal{S}}(\boldsymbol{x}^t) - \frac{1}{2m}\sum_{j=1}^{m}\frac{1}{L_j}\|U_j^\top\nabla F_{\mathcal{S}}(\boldsymbol{x}^t)\|_2^2. \tag{101}$$

Adding (100) and (101) cancels the block-gradient quadratic terms. By convexity of $F_{\mathcal{S}}$,

$$\frac{1}{m}\Big(F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x})\Big) \le D_{\mathcal{L}}(\boldsymbol{x}^t,\boldsymbol{x}) + F_{\mathcal{S}}(\boldsymbol{x}^t)$$

$$- \mathbb{E}_t\big[D_{\mathcal{L}}(\boldsymbol{x}^{t+1}, \boldsymbol{x}) + F_{\mathcal{S}}(\boldsymbol{x}^{t+1})\big]\,. \tag{102}$$

Taking total expectation and summing (102) from $t = 0$ to $T-1$, the right-hand side telescopes. Since $D_{\mathcal{L}} \geq 0$ and $F_{\mathcal{S}} \geq 0$ by Assumption 4.1, dropping the negative terminal contribution gives

$$\sum_{t=0}^{T-1} \mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x})\big] \leq m D_{\mathcal{L}}(\boldsymbol{x}^0, \boldsymbol{x}) + m\,\mathbb{E}[F_{\mathcal{S}}(\boldsymbol{x}^0)]. \tag{103}$$

Adding $T\,\mathbb{E}[F_{\mathcal{S}}(\boldsymbol{x})]$ to both sides, and using that $\boldsymbol{x}^0$ and $\boldsymbol{x}$ are deterministic and independent of $\mathcal{S}$, we have $\mathbb{E}[F_{\mathcal{S}}(\boldsymbol{x}^0)] = F(\boldsymbol{x}^0), \mathbb{E}[F_{\mathcal{S}}(\boldsymbol{x})] = F(\boldsymbol{x})$. Hence

$$\sum_{t=0}^{T-1} \mathbb{E}[F_{\mathcal{S}}(\boldsymbol{x}^t)] \leq m D_{\mathcal{L}}(\boldsymbol{x}^0, \boldsymbol{x}) + m F(\boldsymbol{x}^0) + T\,F(\boldsymbol{x}). \tag{104}$$

Substituting (104) into (99) yields

$$\bar{\mathcal{E}}_T \leq \frac{8 L_f T}{N^2(2L_{\min} - L_f)} \left[D_{\mathcal{L}}(\boldsymbol{x}^0, \boldsymbol{x}) + F(\boldsymbol{x}^0) + \frac{T}{m} F(\boldsymbol{x})\right]. \tag{105}$$

Finally, since $\mathcal{E}_T^{(n)} \geq \frac{L_{\min}}{2}\|\boldsymbol{x}^T - \boldsymbol{x}^{(n),T}\|_2^2$, we conclude that

$$\begin{aligned}\frac{1}{N}\sum_{n=1}^{N} \mathbb{E}\Big[\|\boldsymbol{x}^T - \boldsymbol{x}^{(n),T}\|_2^2\Big] &\leq \frac{2}{L_{\min}} \bar{\mathcal{E}}_T \\ &\leq \frac{16 L_f T}{N^2 L_{\min}(2L_{\min} - L_f)} \left[D_{\mathcal{L}}(\boldsymbol{x}^0, \boldsymbol{x}) + F(\boldsymbol{x}^0) + \frac{T}{m} F(\boldsymbol{x})\right]. \end{aligned} \tag{106}$$

Taking square roots proves the corollary. □

G.1.5 Proof of Theorem 4.11

Proof. Applying Lemma 3.3 to the output $\mathcal{A}(\mathcal{S}) = \boldsymbol{x}^T$, we obtain, for every $\gamma > 0$,

$$\epsilon_{\text{gen}} \leq \frac{1}{\gamma}\,\mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^T)\big] + \frac{L_f(1+\gamma)}{2N} \sum_{n=1}^{N} \mathbb{E}\Big[\|\boldsymbol{x}^T - \boldsymbol{x}^{(n),T}\|_2^2\Big]\,. \tag{107}$$

We apply Theorem 4.7 with the deterministic comparator $\boldsymbol{x} = \boldsymbol{x}^\star$. Since $\boldsymbol{x}^\star$ is independent of $\mathcal{S}$, $\mathcal{S}'$, and the algorithmic randomness,

$$\begin{aligned}\frac{1}{N}\sum_{n=1}^{N} \mathbb{E}\Big[\|\boldsymbol{x}^T - \boldsymbol{x}^{(n),T}\|_2^2\Big] \leq \frac{4 L_f T}{N^2} \Bigg[&\eta_0 \|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|_2^2 + 2\eta_0^2 F(\boldsymbol{x}^0) \\ &+ 2\alpha\Big(\sum_{t=0}^{T-1} \eta_t^2\Big) F(\boldsymbol{x}^\star)\Bigg]. \end{aligned} \tag{108}$$

Substituting (108) into (107) gives

$$\begin{aligned}\epsilon_{\text{gen}} \leq\ &\frac{1}{\gamma}\,\mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^T)\big] \\ &+ \frac{2L_f^2(1+\gamma)T}{N^2} \left[\eta_0 \|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|_2^2 + 2\eta_0^2 F(\boldsymbol{x}^0) + 2\alpha\Big(\sum_{t=0}^{T-1} \eta_t^2\Big) F(\boldsymbol{x}^\star)\right]. \end{aligned} \tag{109}$$

It remains to bound $\mathbb{E}[F_{\mathcal{S}}(\boldsymbol{x}^T)]$.

Since $\eta_t \leq \frac{1}{L_f} \leq \frac{1}{L_{\max}}$, Lemma F.3 implies that $\mathbb{E}[F_{\mathcal{S}}(\boldsymbol{x}^t)]$ is nonincreasing in t. Hence, for every $0 \leq t \leq T-1$,

$$\mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^T) - F_{\mathcal{S}}(\boldsymbol{x}^\star)\big] \leq \mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x}^\star)\big]\,.$$

Multiplying by η_t^2, summing over $t = 0, \dots, T-1$, and applying Lemma F.4(ii) with the deterministic comparator $\boldsymbol{x} = \boldsymbol{x}^\star$, we get

$$\mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^T) - F_{\mathcal{S}}(\boldsymbol{x}^\star)\big] \le \frac{\eta_0 \|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|_2^2 + 2\eta_0^2\, \mathbb{E}[F_{\mathcal{S}}(\boldsymbol{x}^0)]}{2\alpha \sum_{t=0}^{T-1} \eta_t^2}. \tag{110}$$

Since $\boldsymbol{x}^0$ and $\boldsymbol{x}^\star$ are deterministic and independent of $\mathcal{S}$, $\mathbb{E}[F_{\mathcal{S}}(\boldsymbol{x}^0)] = F(\boldsymbol{x}^0)$, $\mathbb{E}[F_{\mathcal{S}}(\boldsymbol{x}^\star)] = F(\boldsymbol{x}^\star)$. Therefore,

$$\mathbb{E}[F_{\mathcal{S}}(\boldsymbol{x}^T)] \le F(\boldsymbol{x}^\star) + \frac{\eta_0 \|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|_2^2 + 2\eta_0^2 F(\boldsymbol{x}^0)}{2\alpha \sum_{t=0}^{T-1} \eta_t^2}. \tag{111}$$

Substituting (111) into (109) proves the theorem. □

G.1.6 Proof of Theorem D.2

Proof. By the excess-risk decomposition (6), $\mathbb{E}[F(\boldsymbol{x}^T)] - F(\boldsymbol{x}^\star) \le \epsilon_{\text{gen}} + \epsilon_{\text{opt}}$. Theorem 4.11 gives, for every $\gamma > 0$,

$$\begin{aligned}\epsilon_{\text{gen}} \le \frac{1}{\gamma}\Bigg[F(\boldsymbol{x}^\star) + \frac{\eta_0 \|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|_2^2 + 2\eta_0^2 F(\boldsymbol{x}^0)}{2\alpha \sum_{t=0}^{T-1} \eta_t^2}\Bigg] \\ + \frac{2L_f^2(1+\gamma)T}{N^2}\Bigg[\eta_0 \|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|_2^2 + 2\eta_0^2 F(\boldsymbol{x}^0) + 2\alpha\Big(\sum_{t=0}^{T-1} \eta_t^2\Big) F(\boldsymbol{x}^\star)\Bigg].\end{aligned} \tag{112}$$

On the other hand, Theorem C.1 yields

$$\begin{aligned}\epsilon_{\text{opt}} &= \mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^T) - F_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}}^\star)\big] \\ &\le \frac{(1 + \eta_0 L_f)\mathbb{E}\big[\|\boldsymbol{x}^0 - \boldsymbol{x}_{\mathcal{S}}^\star\|_2^2\big]}{2\alpha \sum_{t=0}^{T-1} \eta_t}.\end{aligned} \tag{113}$$

Combining (112) and (113) with (6) gives Eq. (17). □

G.1.7 Proof of Corollary 4.13 (general case)

Proof. We use the pre-optimization bound obtained in the proof of Theorem D.2. Set $\eta_t \equiv 1/L_f$. Then $\sum_{t=0}^{T-1} \eta_t = \frac{T}{L_f}$, $\sum_{t=0}^{T-1} \eta_t^2 = \frac{T}{L_f^2}$, $1 + \eta_0 L_f = 2$. For every $\gamma > 0$, substitution gives

$$\begin{aligned}\mathbb{E}[F(\boldsymbol{x}^T)] - F(\boldsymbol{x}^\star) \le \frac{L_f\, \mathbb{E}[\|\boldsymbol{x}^0 - \boldsymbol{x}_{\mathcal{S}}^\star\|_2^2]}{\alpha T} + \frac{1}{\gamma}\Bigg[F(\boldsymbol{x}^\star) + \frac{L_f \|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|_2^2 + 2F(\boldsymbol{x}^0)}{2\alpha T}\Bigg] \\ + \frac{2L_f(1+\gamma)T}{N^2}\Bigg[\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|_2^2 + \frac{2}{L_f} F(\boldsymbol{x}^0) + \frac{2\alpha T}{L_f} F(\boldsymbol{x}^\star)\Bigg].\end{aligned} \tag{114}$$

The stated assumption gives $L_f \mathbb{E}\|\boldsymbol{x}^0 - \boldsymbol{x}_{\mathcal{S}}^\star\|_2^2 \le C_0$, $L_f \|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|_2^2 + 2F(\boldsymbol{x}^0) \le 2C_0$, and $F(\boldsymbol{x}^\star) \le C_0$. Thus the right-hand side of (114) is at most

$$C_0\left[\frac{1}{\alpha T} + \frac{1}{\gamma} + \frac{1}{\gamma \alpha T} + 4(1+\gamma)\left(\frac{T}{N^2} + \frac{\alpha T^2}{N^2}\right)\right]. \tag{115}$$

No smoothness-dependent factor is absorbed into C_0 here. With $a = \alpha^{-1/4} N^{-1/2}$, take $T = \lceil \alpha^{-3/4}\sqrt{N} \rceil$ and $\gamma = N/(\sqrt{\alpha}\, T)$. Because $\alpha^{-3/4}\sqrt{N} \ge 1$, we have $\alpha^{-3/4}\sqrt{N} \le T \le 2\alpha^{-3/4}\sqrt{N}$, and therefore

$$\frac{1}{\alpha T} \le a, \qquad \frac{1}{\gamma} \le 2a, \qquad \frac{1}{\gamma \alpha T} = \frac{\gamma T}{N^2} = a^2,$$
$$\frac{T}{N^2} \le 2a^3, \qquad \frac{\alpha T^2}{N^2} \le 4a^2, \qquad \frac{\gamma \alpha T^2}{N^2} = \frac{1}{\gamma} \le 2a.$$

Substituting these bounds into (115) gives

$$\mathbb{E}[F(\boldsymbol{x}^T)] - F(\boldsymbol{x}^\star) \le C_0\big(11a + 21a^2 + 8a^3\big). \tag{116}$$

The term T/N^2 is of order a^3 and cannot be absorbed into a^2 uniformly in a. For $a \le 1$, however, (116) gives $\mathbb{E}[F(\boldsymbol{x}^T)] - F(\boldsymbol{x}^\star) \le 40C_0 a$.

For $a > 1$ we use a complementary distance estimate. Nonnegativity and smoothness give $\mathbb{E}_\xi\|\nabla f(\boldsymbol{x}^\star;\xi)\|_2^2 \le 2L_f F(\boldsymbol{x}^\star) < \infty$; together with the uniform gradient Lipschitz bound, this allows differentiation under the expectation, so F is L_f-smooth and $\nabla F(\boldsymbol{x}^\star) = 0$. Conditional on $\mathcal{S}$ and the past draws, $\mathbb{E}_t[\Pi_t] = \alpha I_d$ and $\Pi_t^2 = \Pi_t$, so the update with step $1/L_f$ satisfies

$$\begin{aligned}\mathbb{E}_t\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^\star\|_2^2 &= \|\boldsymbol{x}^t - \boldsymbol{x}^\star\|_2^2 - \frac{2\alpha}{L_f}\langle \nabla F_{\mathcal{S}}(\boldsymbol{x}^t), \boldsymbol{x}^t - \boldsymbol{x}^\star\rangle + \frac{\alpha}{L_f^2}\|\nabla F_{\mathcal{S}}(\boldsymbol{x}^t)\|_2^2 \\ &\le \|\boldsymbol{x}^t - \boldsymbol{x}^\star\|_2^2 + \frac{2\alpha}{L_f}F_{\mathcal{S}}(\boldsymbol{x}^\star). \end{aligned}\tag{117}$$

The inequality uses convexity, $\langle \nabla F_{\mathcal{S}}(\boldsymbol{x}^t), \boldsymbol{x}^t - \boldsymbol{x}^\star\rangle \ge F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x}^\star)$, and self-bounding, $\|\nabla F_{\mathcal{S}}(\boldsymbol{x}^t)\|_2^2 \le 2L_f F_{\mathcal{S}}(\boldsymbol{x}^t)$ (Lemma F.5, applied to the nonnegative, L_f-smooth empirical risk).

Iterating, averaging over $\mathcal{S}$, and using that $\boldsymbol{x}^\star$ is deterministic with $\mathbb{E}_{\mathcal{S}}F_{\mathcal{S}}(\boldsymbol{x}^\star) = F(\boldsymbol{x}^\star)$, we get $\mathbb{E}\|\boldsymbol{x}^T - \boldsymbol{x}^\star\|_2^2 \le \|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|_2^2 + 2\alpha T F(\boldsymbol{x}^\star)/L_f$; in particular the iterates have finite second moments and integrable population risks. Population smoothness at $\boldsymbol{x}^\star$ then gives

$$\mathbb{E}[F(\boldsymbol{x}^T)] - F(\boldsymbol{x}^\star) \le \frac{L_f}{2}\mathbb{E}\|\boldsymbol{x}^T - \boldsymbol{x}^\star\|_2^2 \le \frac{L_f}{2}\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|_2^2 + \alpha T F(\boldsymbol{x}^\star) \le C_0\Big(\frac{1}{2} + \alpha T\Big). \tag{118}$$

For $a > 1$, $\alpha T \le 2\alpha^{1/4}\sqrt{N} = 2/a < 2$, so the excess risk is at most $\frac{5}{2}C_0$. Combining the two cases proves

$$\mathbb{E}[F(\boldsymbol{x}^T)] - F(\boldsymbol{x}^\star) \le 40C_0\min\{a, 1\} \le 40C_0(a + a^2) = 40C_0\left[\frac{1}{\alpha^{1/4}\sqrt{N}} + \frac{1}{\sqrt{\alpha}\,N}\right] \tag{119}$$

for every integer $N \ge 1$ and every $\alpha \in (0, 1]$. This establishes both the uniform bound and the two-term form. The condition $N \gtrsim \alpha^{-1/2}$ only makes the second term of the two-term form dominated by the first; it is not needed for (119). A vanishing guarantee requires $a \to 0$, equivalently $N\sqrt{\alpha} \to \infty$, with C_0 bounded. □

G.1.8 Proof of Corollary 4.13 (low-noise case)

Proof. Under the additional low-noise assumption $F(\boldsymbol{x}^\star) \le C_0/N$, the bound (114) becomes

$$\begin{aligned}\mathbb{E}[F(\boldsymbol{x}^T)] - F(\boldsymbol{x}^\star) \le\ & O\left(\frac{C_0}{\alpha T}\right) + O\left(\frac{1}{\gamma}\left[\frac{C_0}{N} + \frac{C_0}{\alpha T}\right]\right) \\ & + O\left((1+\gamma)\left[\frac{C_0 T}{N^2} + \frac{C_0\alpha T^2}{N^3}\right]\right). \end{aligned}\tag{120}$$

Choose $\gamma = 1$ and $T = \lceil N/\sqrt{\alpha}\rceil$. For every $N \ge 1$ and $\alpha \in (0, 1]$, $N/\sqrt{\alpha} \le T \le 2N/\sqrt{\alpha}$, so

$$\frac{1}{\alpha T} \le \frac{1}{N\sqrt{\alpha}}, \qquad \frac{T}{N^2} \le \frac{2}{N\sqrt{\alpha}}, \qquad \frac{\alpha T^2}{N^3} \le \frac{4}{N}.$$

Since $\alpha \le 1$, the terms of order $1/N$ are dominated by $1/(N\sqrt{\alpha})$, and no relation between N and α is needed. Substituting these relations into (120) yields

$$\mathbb{E}[F(\boldsymbol{x}^T)] - F(\boldsymbol{x}^\star) = O\left(\frac{C_0}{N\sqrt{\alpha}}\right).$$

□

G.1.9 Proof of Theorem D.1

Proof. Applying Lemma 3.3 to the output $\mathcal{A}(\mathcal{S}) = \boldsymbol{x}^T$, we obtain, for every $\gamma > 0$,

$$\epsilon_{\text{gen}} \leq \frac{1}{\gamma}\,\mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^T)\big] + \frac{L_f(1+\gamma)}{2N}\sum_{n=1}^{N}\mathbb{E}\Big[\|\boldsymbol{x}^T - \boldsymbol{x}^{(n),T}\|_2^2\Big]\,. \tag{121}$$

We apply Corollary 4.9 with the deterministic comparator $\boldsymbol{x} = \boldsymbol{x}^\star$. Since $\boldsymbol{x}^\star$ is independent of $\mathcal{S}, \mathcal{S}'$, and the algorithmic randomness,

$$\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\|\boldsymbol{x}^T - \boldsymbol{x}^{(n),T}\|_2^2\Big] \leq \frac{16L_fT}{N^2L_{\min}(2L_{\min} - L_f)}\left[D_{\mathcal{L}}(\boldsymbol{x}^0, \boldsymbol{x}^\star) + F(\boldsymbol{x}^0) + \frac{T}{m}F(\boldsymbol{x}^\star)\right]. \tag{122}$$

Substituting (122) into (121) yields

$$\begin{aligned}\epsilon_{\text{gen}} \leq{}& \frac{1}{\gamma}\,\mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^T)\big] \\ &+ \frac{8L_f^2(1+\gamma)T}{N^2L_{\min}(2L_{\min} - L_f)}\left[D_{\mathcal{L}}(\boldsymbol{x}^0, \boldsymbol{x}^\star) + F(\boldsymbol{x}^0) + \frac{T}{m}F(\boldsymbol{x}^\star)\right].\end{aligned} \tag{123}$$

It remains to control $\mathbb{E}[F_{\mathcal{S}}(\boldsymbol{x}^T)]$. For any deterministic comparator $\boldsymbol{x} \in \mathbb{R}^d$, independent of $\mathcal{S}$ and the algorithmic randomness, the block update with step size $1/L_{i_t}$ satisfies, by the same weighted-potential argument used for BCD with blockwise step sizes,

$$\frac{1}{m}\big(F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x})\big) \leq D_{\mathcal{L}}(\boldsymbol{x}^t, \boldsymbol{x}) + F_{\mathcal{S}}(\boldsymbol{x}^t) - \mathbb{E}_t\big[D_{\mathcal{L}}(\boldsymbol{x}^{t+1}, \boldsymbol{x}) + F_{\mathcal{S}}(\boldsymbol{x}^{t+1})\big]\,. \tag{124}$$

Indeed, expanding $D_{\mathcal{L}}$ along the blockwise step and using Assumption 4.4, the block-gradient quadratic terms cancel; convexity gives

$$F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x}) \leq \big\langle \nabla F_{\mathcal{S}}(\boldsymbol{x}^t), \boldsymbol{x}^t - \boldsymbol{x}\big\rangle\,.$$

Taking total expectation in (124), summing over $t = 0, \ldots, T-1$, and using $D_{\mathcal{L}} \geq 0$ and $F_{\mathcal{S}} \geq 0$ to drop the negative terminal contribution gives

$$\sum_{t=0}^{T-1}\mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x})\big] \leq m\,D_{\mathcal{L}}(\boldsymbol{x}^0, \boldsymbol{x}) + m\,\mathbb{E}[F_{\mathcal{S}}(\boldsymbol{x}^0)]. \tag{125}$$

Since $\boldsymbol{x}^0$ is deterministic, $\mathbb{E}[F_{\mathcal{S}}(\boldsymbol{x}^0)] = F(\boldsymbol{x}^0)$. Moreover, each blockwise step is a pathwise descent step for $F_{\mathcal{S}}$, and hence $F_{\mathcal{S}}(\boldsymbol{x}^{t+1}) \leq F_{\mathcal{S}}(\boldsymbol{x}^t)$ for every realization. Therefore,

$$T\,\mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^T) - F_{\mathcal{S}}(\boldsymbol{x})\big] \leq \sum_{t=0}^{T-1}\mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x})\big]\,. \tag{126}$$

Combining (125) and (126), and then choosing $\boldsymbol{x} = \boldsymbol{x}^\star$, gives

$$\mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^T)\big] \leq F(\boldsymbol{x}^\star) + \frac{m}{T}\Big(D_{\mathcal{L}}(\boldsymbol{x}^0, \boldsymbol{x}^\star) + F(\boldsymbol{x}^0)\Big), \tag{127}$$

since $\boldsymbol{x}^\star$ is deterministic and $\mathbb{E}[F_{\mathcal{S}}(\boldsymbol{x}^\star)] = F(\boldsymbol{x}^\star)$. Substituting (127) into (123) gives (16). □

G.1.10 Proof of Theorem D.3

Proof. By the excess-risk decomposition (6), $\mathbb{E}[F(\boldsymbol{x}^T)] - F(\boldsymbol{x}^\star) \leq \epsilon_{\text{gen}} + \epsilon_{\text{opt}}$. The optimization term is controlled by Corollary C.3:

$$\begin{aligned}\epsilon_{\text{opt}} &= \mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^T) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\big] \\ &\leq \frac{m}{T}\left(\mathbb{E}\big[D_{\mathcal{L}}(\boldsymbol{x}^0, \boldsymbol{x}^\star_{\mathcal{S}})\big] + \frac{L_f}{2}\mathbb{E}\big[\|\boldsymbol{x}^0 - \boldsymbol{x}^\star_{\mathcal{S}}\|_2^2\big]\right).\end{aligned} \tag{128}$$

The generalization term is bounded by Theorem D.1; that is, for every $\gamma > 0$,

$$\epsilon_{\text{gen}} \le \frac{1}{\gamma}\left[F(\boldsymbol{x}^\star) + \frac{m}{T}\Big(D_{\mathcal{L}}(\boldsymbol{x}^0, \boldsymbol{x}^\star) + F(\boldsymbol{x}^0)\Big)\right] + \frac{8L_f^2(1+\gamma)T}{N^2 L_{\min}(2L_{\min} - L_f)}\left[D_{\mathcal{L}}(\boldsymbol{x}^0, \boldsymbol{x}^\star) + F(\boldsymbol{x}^0) + \frac{T}{m}F(\boldsymbol{x}^\star)\right]. \tag{129}$$

Combining (128) and (129) with (6) gives Eq. (18). □

G.1.11 Proof of Corollary D.4

Proof. We keep the curvature factor $K = L_f^2/[L_{\min}(2L_{\min} - L_f)]$ explicit rather than absorbing it into C_0. Equation (18) and the stated bound on C_0 give, for every $\gamma > 0$,

$$\mathbb{E}\big[F(\boldsymbol{x}^T)\big] - F(\boldsymbol{x}^\star) \le C_0\left[\frac{m}{T} + \frac{1}{\gamma} + \frac{m}{\gamma T} + 8K(1+\gamma)\left(\frac{T}{N^2} + \frac{T^2}{mN^2}\right)\right]. \tag{130}$$

Take $T = \lceil m^{3/4}\sqrt{N}\rceil$ and $\gamma = N\sqrt{m}/T$. Since $m^{3/4}\sqrt{N} \le T \le 2m^{3/4}\sqrt{N}$, substitution gives

$$\mathbb{E}\big[F(\boldsymbol{x}^T)\big] - F(\boldsymbol{x}^\star) \le C_0\left[(3+16K)\frac{m^{1/4}}{\sqrt{N}} + (1+40K)\frac{\sqrt{m}}{N} + 16K\frac{m^{3/4}}{N^{3/2}}\right]. \tag{131}$$

If $m \le C_m N^2$ for a fixed $C_m > 0$, then $m^{1/4}/\sqrt{N} \le C_m^{1/4}$, so the second and third terms are at most $C_m^{1/4}$ and $C_m^{1/2}$ times the first order, and

$$\mathbb{E}[F(\boldsymbol{x}^T)] - F(\boldsymbol{x}^\star) \le \frac{C_0 m^{1/4}}{\sqrt{N}}\left[3 + 16K + (1+40K)C_m^{1/4} + 16KC_m^{1/2}\right] = O\left(\frac{C_0(1+K)m^{1/4}}{\sqrt{N}}\right),$$

where the $O(\cdot)$ constant depends only on C_m. The strict inequality $L_f < 2L_{\min}$ makes K finite for a fixed problem but does not control K uniformly over a family of problems. □

G.1.12 Proof of Corollary D.5

Proof. Under the additional assumption $F(\boldsymbol{x}^\star) \le C_0/N$, the γ-dependent bound in Eq. (18) gives, with K as in the previous proof,

$$\mathbb{E}\big[F(\boldsymbol{x}^T)\big] - F(\boldsymbol{x}^\star) \le C_0\left[\frac{m}{T} + \frac{1}{\gamma N} + \frac{m}{\gamma T} + 8K(1+\gamma)\left(\frac{T}{N^2} + \frac{T^2}{mN^3}\right)\right]. \tag{132}$$

Take $\gamma = 1$ and $T = \lceil N\sqrt{m}\rceil$. For every $N, m \ge 1$, $N\sqrt{m} \le T \le 2N\sqrt{m}$, hence $m/T \le \sqrt{m}/N$, $T/N^2 \le 2\sqrt{m}/N$, and $T^2/(mN^3) \le 4/N$. Substitution yields

$$\mathbb{E}\big[F(\boldsymbol{x}^T)\big] - F(\boldsymbol{x}^\star) \le C_0\left[(2+32K)\frac{\sqrt{m}}{N} + \frac{1+64K}{N}\right] \le (3+96K)\frac{C_0\sqrt{m}}{N}.$$

This proof uses only $m \ge 1$, without the general-convex condition $m = O(N^2)$. □

G.2 Proofs for the nonconvex case

G.2.1 Proof of Theorem C.5

Proof. Fix a dataset $\mathcal{S}$. By Lemma F.8, the iterates remain in the local PŁ region, and for each $t \ge 0$,

$$\mathbb{E}_t\big[F_{\mathcal{S}}(\boldsymbol{x}^{t+1}) - F_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}}^\star)\big] \le \left[1 - 2\alpha\beta_{\mathcal{S}}\eta_t\left(1 - \frac{L_{\max}\eta_t}{2}\right)\right]_+ \big(F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x}_{\mathcal{S}}^\star)\big). \tag{133}$$

Iterating this conditional recursion and using the tower property gives

$$\begin{aligned}\mathbb{E}_{\mathcal{A}|\mathcal{S}}\left[F_{\mathcal{S}}(\boldsymbol{x}^T) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\right] &\le \prod_{t=0}^{T-1}\left[1 - 2\alpha\beta_{\mathcal{S}}\eta_t\left(1 - \frac{L_{\max}\eta_t}{2}\right)\right]_+ \left(F_{\mathcal{S}}(\boldsymbol{x}^0) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\right) \\ &\le \exp\left(-2\alpha\beta_{\mathcal{S}}\sum_{t=0}^{T-1}\eta_t\left(1 - \frac{L_{\max}\eta_t}{2}\right)\right)\left(F_{\mathcal{S}}(\boldsymbol{x}^0) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\right) \\ &= e^{-\mathfrak{C}_{\mathcal{S}}(T)}\left(F_{\mathcal{S}}(\boldsymbol{x}^0) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\right), \end{aligned} \tag{134}$$

where the second inequality uses $[1-u]_+ \le e^{-u}$ for $u \ge 0$. Taking expectation over $\mathcal{S}$ proves the residual-form bound in (13).

By Assumption 4.2 and Remark 4.3, $F_{\mathcal{S}}$ is L_f-smooth. At the unconstrained minimizer $\boldsymbol{x}^\star_{\mathcal{S}}$, we have $\nabla F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}}) = 0$. Smoothness at $(\boldsymbol{x}^\star_{\mathcal{S}}, \boldsymbol{x}^0)$ bounds the initial residual by the squared initial distance, giving the distance form of the optimization bound:

$$\begin{aligned} F_{\mathcal{S}}(\boldsymbol{x}^0) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}}) &\le \left\langle \nabla F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}}), \boldsymbol{x}^0 - \boldsymbol{x}^\star_{\mathcal{S}}\right\rangle + \frac{L_f}{2}\|\boldsymbol{x}^0 - \boldsymbol{x}^\star_{\mathcal{S}}\|_2^2 \\ &= \frac{L_f}{2}\|\boldsymbol{x}^0 - \boldsymbol{x}^\star_{\mathcal{S}}\|_2^2. \end{aligned} \tag{135}$$

Substituting (135) into (134) and then taking expectation over $\mathcal{S}$ proves the second inequality in (13). □

G.2.2 Proof of Corollary C.8

Proof. Fix a dataset $\mathcal{S}$. For the BCD specialization, $P_i = U_i$, $\rho_i = 1/m$, and the sampled block i_t is updated with step size $1/L_{i_t}$. If $i_t = j$, then $\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \frac{1}{L_j}U_j U_j^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}^t)$. Applying Assumption 4.4 at $\boldsymbol{x}^t$ in the direction $-L_j^{-1}U_j^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}^t)$, we obtain

$$F_{\mathcal{S}}(\boldsymbol{x}^{t+1}) \le F_{\mathcal{S}}(\boldsymbol{x}^t) - \frac{1}{2L_j}\left\|U_j^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}^t)\right\|_2^2. \tag{136}$$

Thus $F_{\mathcal{S}}(\boldsymbol{x}^t)$ is pathwise nonincreasing, and hence $\boldsymbol{x}^t \in \mathcal{U}_0$ for all $t \ge 0$. Taking conditional expectation over the uniform block choice and subtracting $F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})$ gives

$$\begin{aligned}\mathbb{E}_t\left[F_{\mathcal{S}}(\boldsymbol{x}^{t+1}) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\right] &\le F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}}) \\ &\quad - \frac{1}{2m}\sum_{j=1}^m \frac{1}{L_j}\left\|U_j^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}^t)\right\|_2^2. \end{aligned} \tag{137}$$

Since $\boldsymbol{x}^t \in \mathcal{U}_0$, Definition C.7 implies $\sum_{j=1}^m \frac{1}{L_j}\left\|U_j^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}^t)\right\|_2^2 \ge 2\beta_{\mathcal{S}}^{\mathcal{L}}\left(F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\right)$. Substituting this bound into (137) yields

$$\mathbb{E}_t\left[F_{\mathcal{S}}(\boldsymbol{x}^{t+1}) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\right] \le \left(1 - \frac{\beta_{\mathcal{S}}^{\mathcal{L}}}{m}\right)\left(F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\right). \tag{138}$$

Because $F_{\mathcal{S}}(\boldsymbol{x}^{t+1}) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}}) \ge 0$ almost surely, the same argument gives the positive-part recursion

$$\mathbb{E}_t\left[F_{\mathcal{S}}(\boldsymbol{x}^{t+1}) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\right] \le \left[1 - \frac{\beta_{\mathcal{S}}^{\mathcal{L}}}{m}\right]_+ \left(F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\right). \tag{139}$$

Iterating (139) via the tower property and using $[1-u]_+ \le e^{-u}$ for $u \ge 0$, we get

$$\mathbb{E}_{\mathcal{A}|\mathcal{S}}\left[F_{\mathcal{S}}(\boldsymbol{x}^T) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\right] \le \exp\left(-\frac{\beta_{\mathcal{S}}^{\mathcal{L}}}{m}T\right)\left(F_{\mathcal{S}}(\boldsymbol{x}^0) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\right). \tag{140}$$

Taking expectation over $\mathcal{S}$ proves the first inequality in (15).

It remains to convert the initial residual into a distance term. By Assumption 4.2 and Remark 4.3, $F_{\mathcal{S}}$ is L_f-smooth. Since $\boldsymbol{x}^\star_{\mathcal{S}}$ is an unconstrained minimizer of $F_{\mathcal{S}}$, we have $\nabla F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}}) = 0$. Therefore,

$$F_{\mathcal{S}}(\boldsymbol{x}^0) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}}) \le \left\langle \nabla F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}}), \boldsymbol{x}^0 - \boldsymbol{x}^\star_{\mathcal{S}}\right\rangle + \frac{L_f}{2}\|\boldsymbol{x}^0 - \boldsymbol{x}^\star_{\mathcal{S}}\|_2^2$$

$$= \frac{L_f}{2}\|\boldsymbol{x}^0 - \boldsymbol{x}^\star_{\mathcal{S}}\|_2^2. \tag{141}$$

Substituting (141) into (140) and then taking expectation over $\mathcal{S}$ proves the second inequality in (15). □

G.2.3 Proof of Lemma 4.17

Proof. Write $\boldsymbol{x} = \mathcal{A}(\mathcal{S})$ and $F^\star_{\mathcal{S}} = F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})$, with the same minimum-value notation for neighboring datasets. We use the pointwise ratio on $\mathcal{S}^{(1)}$ at $\boldsymbol{x}$ throughout. All expectations below include the algorithm's independent random seed. If the neighboring ratio is zero with positive probability, the lemma's right-hand side is defined as $+\infty$ and the bound is immediate. Otherwise the ratio lies in $(0, +\infty]$ almost surely. Symmetry expresses the generalization gap as

$$\epsilon_{\text{gen}} = N\,\mathbb{E}[F_{\mathcal{S}^{(1)}}(\boldsymbol{x}) - F_{\mathcal{S}}(\boldsymbol{x})].$$

Since $\mathcal{S}$ and $\mathcal{S}^{(1)}$ have the same distribution, their expected minimum empirical risks agree. Subtracting the minima and dropping the nonnegative training residual therefore yields

$$\begin{aligned}\epsilon_{\text{gen}} &\le N\,\mathbb{E}\left[F_{\mathcal{S}^{(1)}}(\boldsymbol{x}) - F^\star_{\mathcal{S}^{(1)}}\right] \\ &= \frac{N}{2}\,\mathbb{E}\left[\frac{\|\nabla F_{\mathcal{S}^{(1)}}(\boldsymbol{x})\|_2^2}{\beta^{\text{pt}}_{\mathcal{S}^{(1)}}(\boldsymbol{x})}\right].\end{aligned}$$

For positive ratios, the equality follows from their definition. At a neighboring minimizer, both the residual and inverse ratio are zero. The replacement identity

$$\nabla F_{\mathcal{S}^{(1)}}(\boldsymbol{x}) = \nabla F_{\mathcal{S}}(\boldsymbol{x}) + \frac{\nabla f(\boldsymbol{x};\xi_1') - \nabla f(\boldsymbol{x};\xi_1)}{N}$$

and $\|u+v\|_2^2 \le 2\|u\|_2^2 + 2\|v\|_2^2$ imply

$$\begin{aligned}\epsilon_{\text{gen}} \le\ &\mathbb{E}\left[\frac{N\|\nabla F_{\mathcal{S}}(\boldsymbol{x})\|_2^2}{\beta^{\text{pt}}_{\mathcal{S}^{(1)}}(\boldsymbol{x})}\right] \\ &+ \mathbb{E}\left[\frac{2\|\nabla f(\boldsymbol{x};\xi_1)\|_2^2 + 2\|\nabla f(\boldsymbol{x};\xi_1')\|_2^2}{N\beta^{\text{pt}}_{\mathcal{S}^{(1)}}(\boldsymbol{x})}\right].\end{aligned}$$

Finally, $F_{\mathcal{S}} - F^\star_{\mathcal{S}}$ is nonnegative and L_f-smooth, so its self-bounding inequality gives $\|\nabla F_{\mathcal{S}}(\boldsymbol{x})\|_2^2 \le 2L_f(F_{\mathcal{S}}(\boldsymbol{x}) - F^\star_{\mathcal{S}})$. Substitution proves the lemma. This argument also applies after fixing any seed that makes the randomized algorithm symmetric, and then averaging over that seed. Well-defined expectations are required; subsequent finite-moment assumptions ensure the bound is informative. □

G.2.4 Proof of Theorem D.7

Proof. By the excess-risk decomposition and the optimality of $\boldsymbol{x}^\star_{\mathcal{S}}$ for $F_{\mathcal{S}}$, we have

$$\mathbb{E}[F(\boldsymbol{x}^T)] - F(\boldsymbol{x}^\star) \le \epsilon_{\text{gen}} + \mathbb{E}\left[F_{\mathcal{S}}(\boldsymbol{x}^T) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\right]. \tag{142}$$

Theorem C.5 gives

$$\mathbb{E}\left[F_{\mathcal{S}}(\boldsymbol{x}^T) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\right] \le \mathbb{E}\left[e^{-\mathfrak{C}_{\mathcal{S}}(T)}\left(F_{\mathcal{S}}(\boldsymbol{x}^0) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\right)\right]. \tag{143}$$

Moreover, Lemma 4.17 yields

$$\begin{aligned}\epsilon_{\text{gen}} \le\ &\mathbb{E}\left[\frac{2\|\nabla f(\boldsymbol{x}^T;\xi_1)\|_2^2 + 2\|\nabla f(\boldsymbol{x}^T;\xi_1')\|_2^2}{N\,\beta^{\text{pt}}_{\mathcal{S}^{(1)}}(\boldsymbol{x}^T)}\right] \\ &+ \mathbb{E}\left[\frac{2L_f N\left(F_{\mathcal{S}}(\boldsymbol{x}^T) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\right)}{\beta^{\text{pt}}_{\mathcal{S}^{(1)}}(\boldsymbol{x}^T)}\right].\end{aligned} \tag{144}$$

Combining (142), (143), and (144) proves the claim. □

G.2.5 Proof of Corollary 4.18

Proof. Let $T = T(\mathcal{S}) \geq 1$ be chosen measurably and permutation-invariantly. Since the base algorithm is symmetric and the horizon depends on the data only through a permutation-invariant rule, the stopped algorithm remains symmetric. Moreover, conditional on $\mathcal{S}$, the horizon $T(\mathcal{S})$ is fixed and is independent of the subsequent subspace randomness. Thus the conditional arguments underlying Theorem D.7 apply to the stopped output $\boldsymbol{x}^T = \boldsymbol{x}^{T(\mathcal{S})}$.

Applying Hölder's inequality to the sample-gradient term in Theorem D.7 gives

$$\begin{aligned}
&\mathbb{E}\left[\frac{2\|\nabla f(\boldsymbol{x}^T;\xi_1)\|_2^2 + 2\|\nabla f(\boldsymbol{x}^T;\xi_1')\|_2^2}{N\,\beta^{\mathrm{pt}}_{\mathcal{S}^{(1)}}(\boldsymbol{x}^T)}\right] \\
&\quad\leq \frac{2}{N}\left(\mathbb{E}\Big[\big(\|\nabla f(\boldsymbol{x}^T;\xi_1)\|_2^2 + \|\nabla f(\boldsymbol{x}^T;\xi_1')\|_2^2\big)^p\Big]\right)^{1/p}\left(\mathbb{E}\Big[\big(\beta^{\mathrm{pt}}_{\mathcal{S}^{(1)}}(\boldsymbol{x}^T)\big)^{-q}\Big]\right)^{1/q} \\
&\quad= \frac{2}{N}\,\mathfrak{G}_p(T)\,\mathfrak{I}_q(T)^{1/q}. \qquad (145)
\end{aligned}$$

The proof of Lemma F.8 gives pathwise monotonicity of the empirical residual, and its one-step recursion gives the weighted residual contraction with $c_{\mathcal{S},t} = 2\alpha\beta_{\mathcal{S}}\eta_t\left(1 - \frac{L_{\max}\eta_t}{2}\right)$. Since Lemma F.10 is proved conditionally on $\mathcal{S}$, the same tower-property argument applies to the $\mathcal{S}$-measurable horizon $T(\mathcal{S})$.

Hence

$$\mathbb{E}\left[\frac{F_{\mathcal{S}}(\boldsymbol{x}^T) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})}{\beta^{\mathrm{pt}}_{\mathcal{S}^{(1)}}(\boldsymbol{x}^T)}\right] \leq \left(\mathbb{E}\Big[e^{-\mathfrak{C}_{\mathcal{S}}(T)}R_0^p\Big]\right)^{1/p}\mathfrak{I}_q(T)^{1/q}. \qquad (146)$$

Substituting (145) and (146) into Theorem D.7 yields

$$\begin{aligned}
\mathbb{E}[F(\boldsymbol{x}^T)] - F(\boldsymbol{x}^\star) \leq\ &\mathbb{E}\Big[e^{-\mathfrak{C}_{\mathcal{S}}(T)}R_0\Big] + \frac{2}{N}\,\mathfrak{G}_p(T)\,\mathfrak{I}_q(T)^{1/q} \\
&+ 2L_f N\left(\mathbb{E}\Big[e^{-\mathfrak{C}_{\mathcal{S}}(T)}R_0^p\Big]\right)^{1/p}\mathfrak{I}_q(T)^{1/q}. \qquad (147)
\end{aligned}$$

The stopping rule gives the following bounds on the two residual terms:

$$e^{-\mathfrak{C}_{\mathcal{S}}(T)}R_0 \leq \frac{1}{N^2L_f}, \qquad e^{-\mathfrak{C}_{\mathcal{S}}(T)}R_0^p \leq \frac{1}{(N^2L_f)^p}. \qquad (148)$$

If $N^2L_fR_0 \leq 1$, then $\log_+(N^2L_fR_0) = 0$, and (148) follows from $e^{-\mathfrak{C}_{\mathcal{S}}(T)} \leq 1$ and $R_0 \leq (N^2L_f)^{-1}$. If $N^2L_fR_0 > 1$, then the stopping rule gives $e^{-\mathfrak{C}_{\mathcal{S}}(T)} \leq (N^2L_fR_0)^{-p}$. Therefore $e^{-\mathfrak{C}_{\mathcal{S}}(T)}R_0^p \leq (N^2L_f)^{-p}$, and also, since $R_0 > (N^2L_f)^{-1}$ and $p > 1$,

$$e^{-\mathfrak{C}_{\mathcal{S}}(T)}R_0 \leq (N^2L_f)^{-p}R_0^{1-p} \leq (N^2L_f)^{-p}(N^2L_f)^{p-1} = \frac{1}{N^2L_f}.$$

This proves (148).

Taking expectations in (148) and substituting the resulting bounds into (147), we obtain

$$\mathbb{E}[F(\boldsymbol{x}^T)] - F(\boldsymbol{x}^\star) \leq \frac{1}{N^2L_f} + \frac{2}{N}\,\mathfrak{G}_p(T)\,\mathfrak{I}_q(T)^{1/q} + \frac{2}{N}\,\mathfrak{I}_q(T)^{1/q},$$

which proves (12).

Finally, if $\eta_t \equiv 1/L_{\max}$, then

$$\mathfrak{C}_{\mathcal{S}}(T) = 2\alpha\beta_{\mathcal{S}}\sum_{t=0}^{T-1}\frac{1}{L_{\max}}\left(1 - \frac{1}{2}\right) = \frac{\alpha\beta_{\mathcal{S}}}{L_{\max}}T.$$

Substituting this identity into the stopping condition gives the constant-step specialization. □

G.2.6 Proof of Corollary D.9

Proof. By the excess-risk decomposition and the optimality of $\boldsymbol{x}^\star_{\mathcal{S}}$ for $F_{\mathcal{S}}$, we have

$$\mathbb{E}[F(\boldsymbol{x}^T)] - F(\boldsymbol{x}^\star) \le \epsilon_{\text{gen}} + \mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^T) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\big]\,. \tag{149}$$

Corollary C.8 gives

$$\mathbb{E}\big[F_{\mathcal{S}}(\boldsymbol{x}^T) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\big] \le \mathbb{E}\Big[e^{-\beta^{\mathcal{L}}_{\mathcal{S}} T/m}\left(F_{\mathcal{S}}(\boldsymbol{x}^0) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\right)\Big]\,. \tag{150}$$

Moreover, Lemma 4.17 yields

$$\begin{aligned}\epsilon_{\text{gen}} \le \mathbb{E}\left[\frac{2\|\nabla f(\boldsymbol{x}^T;\xi_1)\|_2^2 + 2\|\nabla f(\boldsymbol{x}^T;\xi_1')\|_2^2}{N\,\beta^{\text{pt}}_{\mathcal{S}^{(1)}}(\boldsymbol{x}^T)}\right] \\ + \mathbb{E}\left[\frac{2L_f N\left(F_{\mathcal{S}}(\boldsymbol{x}^T) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\right)}{\beta^{\text{pt}}_{\mathcal{S}^{(1)}}(\boldsymbol{x}^T)}\right].\end{aligned} \tag{151}$$

Combining (149), (150), and (151) proves (20). □

G.2.7 Proof of Corollary D.10

Proof. Since $T = T(\mathcal{S})$ is measurable and permutation-invariant, the stopped BCD algorithm with blockwise step sizes remains symmetric. Conditional on $\mathcal{S}$, the horizon $T(\mathcal{S})$ is fixed and independent of the subsequent block-sampling randomness. Thus the conditional arguments used in Corollaries C.8 and D.9 apply to the stopped output $\boldsymbol{x}^T = \boldsymbol{x}^{T(\mathcal{S})}$.

Applying Hölder's inequality to the sample-gradient term in (20) gives

$$\begin{aligned}&\mathbb{E}\left[\frac{2\|\nabla f(\boldsymbol{x}^T;\xi_1)\|_2^2 + 2\|\nabla f(\boldsymbol{x}^T;\xi_1')\|_2^2}{N\,\beta^{\text{pt}}_{\mathcal{S}^{(1)}}(\boldsymbol{x}^T)}\right] \\ &\quad\le \frac{2}{N}\left(\mathbb{E}\Big[\big(\|\nabla f(\boldsymbol{x}^T;\xi_1)\|_2^2 + \|\nabla f(\boldsymbol{x}^T;\xi_1')\|_2^2\big)^p\Big]\right)^{1/p}\left(\mathbb{E}\Big[\big(\beta^{\text{pt}}_{\mathcal{S}^{(1)}}(\boldsymbol{x}^T)\big)^{-q}\Big]\right)^{1/q} \\ &\quad= \frac{2}{N}\mathfrak{G}_p(T)\mathfrak{I}_q(T)^{1/q}.\end{aligned} \tag{152}$$

By the proof of Corollary C.8, the empirical residual is nonincreasing along every realized trajectory and satisfies the conditional contraction below.

$$\mathbb{E}_t\big[F_{\mathcal{S}}(\boldsymbol{x}^{t+1}) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\big] \le \left[1 - \frac{\beta^{\mathcal{L}}_{\mathcal{S}}}{m}\right]_+ \big(F_{\mathcal{S}}(\boldsymbol{x}^t) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})\big)\,. \tag{153}$$

Since Lemma F.10 is proved conditionally on $\mathcal{S}$, the same tower-property argument applies to the $\mathcal{S}$-measurable horizon $T(\mathcal{S})$. Applying it with $c_{\mathcal{S},t} = \frac{\beta^{\mathcal{L}}_{\mathcal{S}}}{m}, t = 0,\dots,T-1$, yields

$$\mathbb{E}\left[\frac{F_{\mathcal{S}}(\boldsymbol{x}^T) - F_{\mathcal{S}}(\boldsymbol{x}^\star_{\mathcal{S}})}{\beta^{\text{pt}}_{\mathcal{S}^{(1)}}(\boldsymbol{x}^T)}\right] \le \left(\mathbb{E}\Big[e^{-\beta^{\mathcal{L}}_{\mathcal{S}} T/m} R_0^p\Big]\right)^{1/p}\mathfrak{I}_q(T)^{1/q}. \tag{154}$$

Substituting (152) and (154) into (20) gives

$$\begin{aligned}\mathbb{E}[F(\boldsymbol{x}^T)] - F(\boldsymbol{x}^\star) \le \mathbb{E}\Big[e^{-\beta^{\mathcal{L}}_{\mathcal{S}} T/m} R_0\Big] + \frac{2}{N}\mathfrak{G}_p(T)\mathfrak{I}_q(T)^{1/q} \\ + 2L_f N\left(\mathbb{E}\Big[e^{-\beta^{\mathcal{L}}_{\mathcal{S}} T/m} R_0^p\Big]\right)^{1/p}\mathfrak{I}_q(T)^{1/q}.\end{aligned} \tag{155}$$

The stopping rule gives the following bounds on the two residual terms:

$$e^{-\beta^{\mathcal{L}}_{\mathcal{S}} T/m} R_0 \le \frac{1}{N^2 L_f}, \qquad e^{-\beta^{\mathcal{L}}_{\mathcal{S}} T/m} R_0^p \le \frac{1}{(N^2 L_f)^p}. \tag{156}$$

If $N^2 L_f R_0 \le 1$, then $R_0 \le (N^2 L_f)^{-1}$, and (156) follows from $e^{-\beta_S^{\mathcal{L}} T/m} \le 1$. If $N^2 L_f R_0 > 1$, then (21) gives $e^{-\beta_S^{\mathcal{L}} T/m} \le (N^2 L_f R_0)^{-p}$. Therefore $e^{-\beta_S^{\mathcal{L}} T/m} R_0^p \le \frac{1}{(N^2 L_f)^p}$, and, since $p > 1$ and $R_0 > (N^2 L_f)^{-1}$,

$$e^{-\beta_S^{\mathcal{L}} T/m} R_0 \le (N^2 L_f)^{-p} R_0^{1-p} \le (N^2 L_f)^{-p} (N^2 L_f)^{p-1} = \frac{1}{N^2 L_f}.$$

This proves (156).

Taking expectations in (156) and substituting the resulting bounds into (155), we obtain

$$\mathbb{E}[F(\boldsymbol{x}^T)] - F(\boldsymbol{x}^\star) \le \frac{1}{N^2 L_f} + \frac{2}{N} \mathfrak{G}_p(T) \mathfrak{I}_q(T)^{1/q} + \frac{2}{N} \mathfrak{I}_q(T)^{1/q},$$

which is (22). □

H Summary of notation

Table 6 groups the main symbols by their role in the analysis.

Table 6: Main notation.

Symbol	Meaning
Learning problem and risk	
$\boldsymbol{x}$	Parameter vector in $\mathbb{R}^d$.
d	Ambient dimension.
ξ	Data observation.
$\mathcal{D}$	Data distribution.
$\mathcal{S}$	Training sample of N i.i.d. observations from $\mathcal{D}$.
$\mathcal{S}'$	Independent copy of $\mathcal{S}$.
$\mathcal{S}^{(n)}$	Neighbor replacing ξ_n with an independent copy ξ_n'.
N	Training sample size.
$f(\boldsymbol{x}; \xi)$	Per-sample loss.
$F(\boldsymbol{x})$, $F_{\mathcal{S}}(\boldsymbol{x})$	Population and empirical risks, respectively.
$\boldsymbol{x}^\star$, $\boldsymbol{x}_{\mathcal{S}}^\star$	Population and empirical risk minimizers, respectively.
$\mathcal{A}(\mathcal{S})$	Randomized algorithm output.
ϵ_{gen}, ϵ_{opt}	Expected generalization and optimization errors, respectively.
$\epsilon_{\mathcal{A}}$	On-average loss-stability parameter.
γ	Balancing parameter in convex generalization bounds.
C_0	Risk and initialization bound; family comparisons assume uniformity.
Subspace updates and curvature	
m	Number of subspace types.
i_t	Subspace type sampled at step t.
ρ_i	Sampling probability of type i.
P_i	Possibly random orthonormal basis of subspace type i.
r_i	Dimension of subspace type i.
Π_t	Sampled orthogonal projector $P_{i_t} P_{i_t}^\top$.
α	Coverage $\sum_i \rho_i r_i / d$ under joint isotropy.
$\bar{r}$	Mean subspace dimension $m^{-1} \sum_i r_i$.
$\boldsymbol{x}^t$	Iterate after t updates.
T	Update horizon; the algorithm returns $\boldsymbol{x}^T$.
η_t	Step size at iteration t.
$\boldsymbol{g}_{i_t}^t$	Subspace gradient $P_{i_t}^\top \nabla F_{\mathcal{S}}(\boldsymbol{x}^t)$.
L_f	Per-sample gradient-Lipschitz constant.

Table 6: Main notation (continued).

Symbol	Meaning
L_i	Empirical-risk smoothness along subspace type i.
$L_{\min}, L_{\max}$	Minimum and maximum of the subspace constants L_i.
U_i	Coordinate-selection matrix for block i.
$\mathcal{L}$	Block smoothness constants $\{L_i\}_{i=1}^m$.
$D_{\mathcal{L}}(\boldsymbol{y}, \boldsymbol{z})$	Block-weighted squared distance (Section 4.2).
K	Curvature factor $L_f^2/[L_{\min}(2L_{\min} - L_f)]$, for $L_f < 2L_{\min}$.
κ	Feature scale of the last five blocks in Figure 8.
Nonconvex analysis and output moments	
$\mathcal{U}_0$	Initial empirical-risk sublevel set.
$\beta_{\mathcal{S}}$	Local PŁ constant on $\mathcal{U}_0$.
$\beta_{\mathcal{S}}^{\mathcal{L}}$	Blockwise local PŁ constant (Definition C.7).
$\beta_{\mathcal{S}}^{\mathrm{pt}}(\boldsymbol{x})$	Pointwise PŁ ratio; $+\infty$ at empirical minimizers.
$\beta_{\mathcal{S}^{(1)}}^{\mathrm{pt}}(\boldsymbol{x}^T)$	Neighboring pointwise PŁ ratio at the trained output.
R_0	Initial empirical-risk residual.
$\mathfrak{C}_{\mathcal{S}}(T)$	RSD contraction clock (Theorem C.5).
p, q	Hölder conjugates: $p > 1$ and $q = p/(p-1)$.
$\mathfrak{G}_p(T)$	L^p norm of the sum of two squared sample-gradient norms (Eq. (10)).
$\mathfrak{I}_q(T)$	Raw qth inverse-curvature moment (Eq. (10)).
$\log_+(u)$	Truncated logarithm $\log(u \vee 1)$, for $u \geq 0$.
Coupling and conditional expectation	
$\boldsymbol{x}^{(n),t}$	Neighboring-data iterate with the same subspace draws.
$\Delta_t^{(n)}$	Coupled iterate difference $\boldsymbol{x}^t - \boldsymbol{x}^{(n),t}$.
$\bar{\Delta}_t^2$	Expected squared discrepancy, averaged over replacements.
$\mathcal{E}_t^{(n)}$	Block-weighted discrepancy $D_{\mathcal{L}}(\boldsymbol{x}^t, \boldsymbol{x}^{(n),t})$.
$\bar{\mathcal{E}}_t$	Expected block-weighted discrepancy, averaged over replacements.
$\mathbb{E}_t[\cdot]$	Expectation over the next subspace draw, given data and history.
$\mathbb{E}_{\mathcal{A}}[\cdot \mid \mathcal{S}]$	Algorithmic expectation conditional on $\mathcal{S}$.

Broader impact

Coverage-aware guarantees can help assess the statistical cost of resource-constrained training. Longer training at lower coverage can offset per-step savings and increase total resource use. Deployment requires separate evaluation of predictive reliability and resource use.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Sections 1–4 state stability and excess-risk bounds for randomized subspace descent under the stated sampling, smoothness, and moment conditions. Section 5 distinguishes qualitative comparisons, saved run-level diagnostics, and author-reported aggregate summaries.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 discusses fixed jointly isotropic sampling, smoothness, and the local PŁ, moment, and horizon conditions. Appendix A identifies the empirical reporting gaps.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The theorem statements specify or reference the RSD sampling and regularity assumptions, with convexity or local PŁ and moment conditions where required. The appendices provide the proofs, including the complementary distance argument for the uniform convex bound and explicit curvature constants for BCD with blockwise step sizes.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [No]

   Justification: Appendix A.3 specifies the controlled tanh protocol and its saved-record analysis. Other experiments lack some run-level records, evaluation settings, or uncertainty metadata, as detailed in Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: A complete public reproduction package for all experiments is not available. The controlled tanh study has saved measurements, a transcription of the executed code, and a paired-analysis script; other studies retain only incomplete configurations or aggregate results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [No]

   Justification: Appendix A reports the recovered settings. GPT-2 evaluation details, original-figure configurations, and some diagnostic implementation records remain incomplete.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

Justification: The tanh study reports marginal and paired outer-run bootstrap intervals with explicit resampling procedures. Least-squares intervals have incomplete metadata, while the original figures and GPT-2 results lack recoverable uncertainty definitions; therefore uncertainty reporting is not complete across all experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Original hardware, memory use, runtime, and total training cost are not recorded consistently across experiments. GPT-2 relative step times and normalized oracle cost do not replace these measurements.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is theoretical and experimental research on optimization and generalization in machine learning, using public datasets without recruiting human participants. To the best of the authors' knowledge, the research conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The broader-impact statement discusses the value of coverage-aware analysis and the resource costs of longer training horizons.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk models, scraped datasets, pre-trained language models, image generators, or other assets that would require special safeguards for responsible release.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: Dataset and model sources are cited, but original-run asset versions and license records are incomplete. The supplementary provenance notes distinguish the recovered material from the remaining author confirmations.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `https://paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: The supplementary package documents saved diagnostic outputs, transcribed code, and reanalysis procedures. A release license has not yet been specified by the authors, so documentation is not complete.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, user studies, or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, user studies, or research with human subjects, so IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs assisted writing, mathematical consistency checks, explanatory derivations, and code and analysis for the controlled neural-network diagnostic. The authors remain responsible for verifying these contributions; the saved diagnostic outputs were also reproduced during revision.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.