# ACCELERATING SINGLE-PASS SGD FOR GENERALIZED LINEAR PREDICTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

We study generalized linear prediction under a streaming setting, where each iteration uses only one fresh data point for a gradient-level update. While momentum is well-established in deterministic optimization, a fundamental open question is whether it can *accelerate* such single-pass non-quadratic stochastic optimization. We propose the *first* algorithm that successfully incorporates momentum via a novel data-dependent proximal method, achieving dual-momentum acceleration. Our derived excess risk bound decomposes into three components: an *improved* optimization error, a minimax *optimal* statistical error, and a *higher-order* model-misspecification error. The proof handles mis-specification via a fine-grained stationary analysis of inner updates, while localizing statistical error through a two-phase outer-loop analysis. As a result, we resolve the open problem posed by Jain et al. (2018a) and demonstrate that momentum acceleration is more effective than variance reduction for generalized linear prediction in the streaming setting.

## 1 INTRODUCTION

This paper considers Generalized Linear Prediction (GLP) in the form of:

$$\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) = \mathbb{E}_{(\mathbf{a},b)\sim\mathcal{D}}\, \ell(\mathbf{a}^\top \mathbf{x}, b), \tag{GLP}$$

which minimizes the expected convex loss $\ell$ over linear prediction of data $\mathbf{a}$ with $(\mathbf{a}, b)$ drawn from a distribution $\mathcal{D}$. This problem is ubiquitous in machine learning. It encapsulates generalized linear models (GLMs), where specific choices of the loss $\ell$ recover fundamental learning problems including linear and logistic regression. We study this problem in a large-scale streaming setting, where an algorithm is permitted only a gradient-level $\mathcal{O}(d)$ computation from a *fresh* data point per iteration. The central interest is to improve the data/runtime complexity required to approximately solve (GLP). We primarily focus on loss functions $\ell$ that are strongly convex and smooth for the first argument, and we denote $\alpha$ as the condition number of the loss (see Section 3 for the definition).

A direct comparison baseline of constrained streaming algorithms is those free of computational and memory constraints. Among them, a commonly used approach is empirical risk minimization:

$$\widehat{\mathbf{x}} \in \arg\min_{\mathbf{x}} \frac{1}{N}\sum_{i=1}^{N} \ell(\mathbf{a}_i^\top \mathbf{x}, b_i). \tag{1}$$

For (1), a non-asymptotic sample complexity of $\frac{\alpha\sigma_0^2}{\varepsilon}$ for the excess risk $F(\widehat{\mathbf{x}}) - F(\mathbf{x}^*)$ can be established via the celebrated localization Rademacher technique (Bartlett et al., 2005; Wainwright, 2019), where $\mathbf{x}^*$ denotes the minimizer of (GLP), $\sigma_0^2$ relates to the Lipschitzness of $\ell$. The sample complexity is at least $\frac{\alpha\sigma_*^2}{\varepsilon}$, where $\sigma_*^2 = \mathrm{tr}(\mathbf{H}^{-1}\mathbf{Q})$, $\mathbf{Q}$ is the stochastic gradient covariance at $\mathbf{x}^*$, and $\mathbf{H}$ is an upper bound of $\nabla^2 F$ (see Section 3.1 for definition). In Appendix G, we show that this sample complexity is optimal under only strong convexity and smoothness conditions on $\ell$.

Turning to the streaming setting, a natural question arises: how much additional data is required under the streaming constraint? It should be noted that standard Stochastic Gradient Descent (SGD) is inefficient for solving (GLP), as it fails to exploit the problem structure. However, for well-specified linear regression (where $\alpha = 1$), it is known that certain variants of SGD can indeed



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

improve the sample complexity to $\kappa \log(\frac{1}{\varepsilon}) + \frac{\sigma_*^2}{\varepsilon}$ (Jain et al., 2018b; Zou et al., 2023; Ge et al., 2019; Wu et al., 2022). Interestingly, the complexity decomposes into the optimal statistical term and an optimization term that scales as $\kappa$, which denotes the condition number of the data distribution (see Section 3).

In recent years, several works (Frostig et al., 2015; Li et al., 2022) have attempted to solve (GLP) in the general setting, obtaining complexity bounds that decompose into a statistical term and an optimization term imposed by the streaming constraint. The core idea of these algorithms is to use *variance-reduction* techniques that separate the gradient variance into the variance at the optimum and an optimization-controllable error component, thereby localizing the statistical error. However, a common limitation is that their optimization complexity scales linearly with the problem condition number $\alpha^2\kappa$. Since this complexity characterizes the time needed to reduce the initial error, the excess risk does not reach a non-trivial level (i.e., $o(1)$) unless the sample size exceeds $\alpha^2\kappa$.

In deterministic optimization, momentum techniques (Polyak, 1964; Nesterov, 1983) are well-known to accelerate convergence across different settings. Yet, how to effectively incorporate momentum into streaming optimization remains largely open. For general stochastic convex problems, there is a prevailing view that momentum may not offer efficient acceleration. For instance, Agarwal et al. (2012) establishes the optimality of SGD under general strongly convex objectives with bounded gradient variance. In contrast, for well-specified linear regression—as a structured estimation problem—Jain et al. (2018a) shows that the optimization term can indeed be accelerated while preserving the optimal statistical complexity. They introduce a statistical condition number of the data, denoted as $\tilde{\kappa}$ (detailed in Section 3), which is smaller than $\kappa$, and improve the optimization complexity to $\sqrt{\kappa\tilde{\kappa}}\log(\frac{1}{\varepsilon})$. Their work leaves two open problems: (i) extending the analysis to model mis-specification, and (ii) exploring more general estimation settings where momentum remains efficient. This paper addresses both by proposing the *first* algorithm that incorporates momentum for (GLP) without relying on fixed Hessian structure and model specification assumption.

### 1.1 Review: Previous Results

**Well-specified Linear Regression.** The central idea enabling SGD-type algorithms to attain optimal statistical complexity—along with an additional optimization term—is to model SGD as a stochastic process and to establish convergence of the variable (Dieuleveut & Bach, 2016; Jain et al., 2018b; Ge et al., 2019; Zou et al., 2023; Wu et al., 2022; Li et al., 2024). Jain et al. (2018a) improves the optimization term and attains the complexity of (ignoring logarithmic factors)

$$\sqrt{\kappa\tilde{\kappa}} + \frac{\sigma_*^2}{\varepsilon}$$

by employing momentum with a finer-grained analysis. This result, however, is restricted to well-specified linear regression, since the stationary state analysis critically relies on the quadratic form of the objective and the commutativity of the gradient noise covariance and the Hessian.

**Variance Reduction for Generalized Linear Prediction.** Frostig et al. (2015) and Li et al. (2022) study a slightly more general stochastic convex optimization problem that does not assume a linear structure in the individual functions. They require the objective $F$ to be $\mu_h$-strongly convex and each component to be $L_h$-smooth, and Frostig et al. (2015) further crucially impose a Hessian dominance condition $\nabla^2 F(\mathbf{x}^*) \preceq \alpha_h \nabla^2 F(\mathbf{x})$ tailored to generalized linear prediction. Adapting their results to our setting, Frostig et al. (2015) attains a complexity[1] of

$$\alpha^2\kappa + \frac{\alpha^2\sigma_*^2}{\varepsilon}, \tag{2}$$

where the term $\alpha^2\kappa$ corresponds to the optimization cost and $\frac{\alpha^2\sigma_*^2}{\varepsilon}$ is the statistical sample complexity. Li et al. (2022) obtains the sample complexity[2] of

$$\alpha^2\kappa \cdot \frac{\lambda_{\max}(\mathbf{\Sigma})}{\lambda_{\min}(\mathbf{\Sigma})} + \frac{\alpha\sigma_*^2}{\varepsilon} + \mathcal{O}\left(\frac{1}{\varepsilon^{2/3}}\right). \tag{3}$$

[1] Frostig et al. (2015) obtain a sample complexity of $\alpha_h\kappa_h + \frac{\alpha_h \operatorname{tr}((\nabla F(\mathbf{x}^*))^{-1}\mathbf{Q})}{\varepsilon}$ in Corollary 4, where $\kappa_h = L_h/\mu_h$. In the worst case, one has $\alpha_h = \alpha$, $\kappa_h = \alpha\kappa$ and $\operatorname{tr}((\nabla F(\mathbf{x}^*))^{-1}\mathbf{Q}) = \alpha\sigma_*^2$, which is (2).

[2] Li et al. (2022) obtain a sample complexity of $\kappa_h + \frac{\ell_\Xi^2}{\mu_h^2} + \frac{\operatorname{tr}((\nabla F(\mathbf{x}^*))^{-1}\mathbf{Q})}{\varepsilon} + \mathcal{O}(\varepsilon^{-2/3})$ in Corollary 3. In the worst case, one has $\kappa_h = \alpha\kappa$, $\ell_\Xi^2/\mu_h^2 = \alpha^2\kappa \cdot \frac{\lambda_{\max}(\mathbf{\Sigma})}{\lambda_{\min}(\mathbf{\Sigma})}$ and $\operatorname{tr}((\nabla F(\mathbf{x}^*))^{-1}\mathbf{Q}) = \alpha\sigma_*^2$, which is (3).

Their work requires an additional Hessian smoothness assumption to further reduce the statistical complexity. Moreover, the last term above hides problem-dependent factors, including the Hessian smoothness parameter. It remains open whether variance reduction is fundamentally necessary for (GLP) and how to improve the dependence on the optimization complexity.

### 1.2 Our Results and Implications

This paper proposes momentum acceleration for (GLP) in the streaming setting. Our algorithm's main idea lies in a data-dependent proximal scheme, with the proximal operator explicitly constructed from the expected data covariance $\mathbf{\Sigma}$. This makes double momentum acceleration applicable to both the inner and outer loops, and we obtain a complexity of

$$\left(\sqrt{\alpha\kappa\tilde{\kappa}} + \alpha^2\tilde{\kappa}\right) + \frac{\alpha\sigma_*^2}{\varepsilon} + \left(\frac{\alpha^2\tilde{\kappa}^2 \operatorname{tr}\mathbf{Q}}{L_\ell \mu \varepsilon}\right)^{1/3}.$$

The above excess risk bound is composed of three interpretable components: (i) an optimization term with *improved* dependence on the problem and data condition numbers, where the first term is from the doubly accelerated scheme, and the second term is the effect of noise; (ii) the estimation term matching the *optimal* statistical risk, and (iii) a *higher-order* mis-specification error term capturing the coupling effect of constrained computation with model mis-specification, for which we provide refined characterization. Our framework can be extended to non-strongly convex objectives and naturally accommodates the use of unlabeled data, mini-batching, and parallel implementations.

The analysis of our algorithm is composed of two parts: (i) In the inner loop, the sub-problem resembles linear regression with model mis-specification. We provide a fine-grained characterization of the effect of model mis-specification; (ii) The outer loop requires us to localize the statistical error to $\mathbf{x}^*$. We propose a two-phase analysis showing that variance reduction is not needed in GLP.

As a result, we resolve the open problems posed by Jain et al. (2018a). We also demonstrate that momentum acceleration is more efficient than variance reduction for streaming generalized linear prediction. This finding is fundamentally different from known results for non-convex optimization in the streaming setting. There, the goal is to find an approximate stationary point, and momentum fails to improve the worst-case rate over plain SGD even with Hessian smoothness assumptions (Fang et al., 2019; Jin et al., 2018). However, variance-reduction methods such as SVRG improve $\varepsilon^{-3.5}$ to $\varepsilon^{-10/3}$ (Allen-Zhu & Hazan, 2016; Reddi et al., 2016), with SPIDER (Fang et al., 2018) and SARAH (Nguyen et al., 2017) attaining the optimal $\varepsilon^{-3}$. We hope our work offers insights toward designing accelerated streaming methods for broader classes of convex and non-convex problems.

**Notations.** For a vector $\mathbf{x}$, let $\|\mathbf{x}\|$ denote 2-norm, and for a positive semi-definite (PSD) matrix $\mathbf{A}$, let $\|\mathbf{x}\|_{\mathbf{A}} = \sqrt{\mathbf{x}^\top \mathbf{A}\mathbf{x}}$. For PSD matrix $\mathbf{\Sigma}$, we denote the maximum and minimum eigenvalue of $\mathbf{\Sigma}$ by $\lambda_{\max}(\mathbf{\Sigma})$ and $\lambda_{\min}(\mathbf{\Sigma})$, respectively. Let $\kappa(\mathbf{\Sigma}) = \frac{\lambda_{\max}(\mathbf{\Sigma})}{\lambda_{\min}(\mathbf{\Sigma})}$ denote the condition number of $\mathbf{\Sigma}$. We use standard asymptotic notations, and use $\tilde{\mathcal{O}}(\cdot)$, $\tilde{\Omega}(\cdot)$, and $\tilde{\Theta}(\cdot)$ to hide logarithmic factors. We write $a_n \lesssim b_n$ (resp. $a_n \gtrsim b_n$) if $a_n = \mathcal{O}(b_n)$ (resp. $a_n = \Omega(b_n)$).

## 2 Related Work

**Stochastic Approximation.** Stochastic Approximation methods date back to the seminal work of Robbins & Monro (1951). A line of work has developed non-asymptotic rate for SGD in various settings, and analyzed the effect of averaging schemes and step size schedules (Moulines & Bach, 2011; Bach & Moulines, 2013; Défossez & Bach, 2015; Dieuleveut & Bach, 2016; Ge et al., 2019; Zhang et al., 2025; Sheshukova et al., 2025). For linear regression setting, refined non-asymptotic analyses show that iterate-averaged SGD attains the optimal statistical rate $\sigma^2 d/n$ (Bach & Moulines, 2013; Jain et al., 2018b; Dieuleveut et al., 2017; Jain et al., 2018a). However, such optimal rates do not generally extend beyond quadratic objectives. Recent works have developed algorithms to achieve such an optimal statistical rate (Frostig et al., 2015; Li et al., 2022). Specifically, Streaming SVRG (Frostig et al., 2015) adopts the idea of variance reduction to the streaming setting and achieves the statistical performance of ERM. ROOT-SGD (Li et al., 2022) achieves a nonasymptotic rate, and its asymptotic rate matches the Cramér-Rao lower bound. These analyses rely on third-order continuous conditions such as self-concordance or Hessian smoothness, while we only require second-order conditions.

**Momentum Acceleration and Variance Reduction.** Momentum methods originate from the heavy-ball method (Polyak, 1964) and Nesterov's accelerated gradient descent (Nesterov, 1983), which achieve accelerated and optimal convergence rates. In the streaming setting, the effect of momentum is more subtle, as it does not improve the worst-case statistical rate compared to SGD in general (Agarwal et al., 2012). However, for linear regression, several works establish accelerated convergence of the optimization error and optimal statistical error by exploiting the quadratic structure (Jain et al., 2018a; Dieuleveut et al., 2017; Pan et al., 2024; Li et al., 2024; Zhang et al., 2024; Liu et al., 2025). However, their analyses rely heavily on matrix calculations, which are non-trivial to extend to general objectives. For non-convex problems, Fang et al. (2019) proved the complexity for SGD to find a second-order stationary point is $\mathcal{O}(\varepsilon^{-3.5})$, matching that of momentum-based methods (Jin et al., 2018). Variance reduction techniques accelerate stochastic optimization by exploiting problem structure to reduce stochastic noise. For convex problems, in the finite-sum setting, SDCA (Shalev-Shwartz & Zhang, 2013) considers the same problem as (GLP). A line of work has developed various algorithms with momentum acceleration (Shalev-Shwartz & Zhang, 2014; Lin et al., 2018; Allen-Zhu, 2018). In the streaming setting, SVRG (Frostig et al., 2015) and ROOT-SGD (Li et al., 2022) improve the statistical term but offer no acceleration to the optimization term.

## 3 Problem Setup

We restate Generalized Linear Prediction as follows: $\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) \overset{\text{def}}{=} \mathbb{E}_{\mathbf{a},b\sim\mathcal{D}}\, \ell(\mathbf{a}^\top\mathbf{x}, b)$. We make the following assumptions on the loss $\ell$ and the data distribution $\mathcal{D}$.

**Assumption 1** (Condition Number of Loss Function)**.** *Let $\ell''(\cdot,\cdot)$ denote the second derivative with respect to its first argument. Assume there exists $L_\ell, \mu_\ell > 0$ such that $\mu_\ell \leqslant \ell''(\mathbf{a}^\top\mathbf{x}, b) \leqslant L_\ell$.*

Let $\alpha \overset{\text{def}}{=} L_\ell/\mu_\ell$ be the condition number of the loss function $\ell$. Assumption 1 implies that $\ell$ is $L_\ell$-smooth and $\mu_\ell$-strongly convex, which is widely used in optimization literature (Nesterov, 2018). This assumption is also commonly adopted in the analysis of generalized linear models, such through tools like local Rademacher complexity (Bartlett et al., 2005; Wainwright, 2019).

**Assumption 2** (Second Moment of Data)**.** *We assume that the second-order moment matrix $\mathbf{\Sigma} \overset{\text{def}}{=} \mathbb{E}\mathbf{a}\mathbf{a}^\top$ is finite and positive definite, with minimum eigenvalue $\mu \overset{\text{def}}{=} \lambda_{\min}(\mathbf{\Sigma}) > 0$.*

Assumption 2 is a mild regularity condition. Combined with Assumption 1, it implies that the objective $F$ is $L_\ell\lambda_{\max}(\mathbf{\Sigma})$-smooth and $\mu_\ell\mu$-strongly convex. In addition, we will extend our results to the non-strongly convex case in Section 8.

**Assumption 3** (Fourth Moment of Data)**.** *We assume the following conditions of the fourth moment:*

$$\mathbb{E}_{\mathbf{a}\sim\mathcal{D}}\left(\|\mathbf{a}\|^2\mathbf{a}\mathbf{a}^\top\right) \preceq R^2\mathbf{\Sigma}, \quad \mathbb{E}_{\mathbf{a},b\sim\mathcal{D}}\left(\|\mathbf{a}\|^2_{\mathbf{\Sigma}^{-1}}\mathbf{a}\mathbf{a}^\top\right) \preceq \tilde{\kappa}\mathbf{\Sigma}.$$

Assumption 3 is commonly used in the analysis of SGD for linear regression (Jain et al., 2018a; Zou et al., 2023; Wu et al., 2022; Li et al., 2024). The quantity $R$ can be viewed as the effective radius of the data $\mathbf{a} \sim \mathcal{D}$. Specifically, if $\|\mathbf{a}\| \leqslant L$ almost surely, then $R = L$. The quantity $\tilde{\kappa}$ can be understood similarly. We refer to $\kappa = R^2/\mu$ as the condition number of $\mathcal{D}$ and $\tilde{\kappa}$ as the statistical condition number following Défossez & Bach (2015); Jain et al. (2018a). Note $\tilde{\kappa} \leqslant \kappa$ always holds since $\mathbb{E}(\|\mathbf{a}\|^2_{\mathbf{\Sigma}^{-1}}\mathbf{a}\mathbf{a}^\top) \preceq \frac{1}{\mu}\mathbb{E}_{\mathbf{a}\sim\mathcal{D}}(\|\mathbf{a}\|^2\mathbf{a}\mathbf{a}^\top) \preceq \kappa\mathbf{\Sigma}$ (Jain et al., 2018a).

In the linear regression setting, Jain et al. (2018a) improve the optimization complexity from $\kappa$ to $\sqrt{\kappa\tilde{\kappa}}$, where $\tilde{\kappa} \leqslant \kappa$. The improvement is significant when the data is poor-conditioned, i.e. $\tilde{\kappa} \ll \kappa$. The statistical condition $\tilde{\kappa}$ is necessary since $\tilde{\kappa}$ governs the concentration of the empirical covariance matrix to $\mathbf{\Sigma}$ and in some sense, it captures the statistical complexity under the noiseless setting (Jain et al., 2018a). We refer the reader to Jain et al. (2018a) for more discussion on the necessity of $\tilde{\kappa}$.

*Remark* 1 (Effect of Acceleration). In the Gaussian design $\mathbf{a} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, one may take $\kappa = 3\,\mathrm{tr}\,\mathbf{\Sigma}/\mu$ and $\tilde{\kappa} = 3d$. So $\kappa \geqslant \tilde{\kappa}$. More generally, if $\mathbf{\Sigma}^{-1/2}\mathbf{a}$ is $\sigma_a^2$-sub-Gaussian, then the condition holds with $\kappa = 16\sigma_a^4\,\mathrm{tr}\,\mathbf{\Sigma}/\mu$ and $\tilde{\kappa} = 16\sigma_a^4 d$ (Zou et al., 2023). In the worst case, $\mathrm{tr}\,\mathbf{\Sigma}/\mu = \kappa(\mathbf{\Sigma})\tilde{\kappa}$.

### 3.1 Summary of Problem-Dependent Quantities

We collect the quantities and present the goal for the overall complexity. We denote $\mathbf{H} = L_\ell\mathbf{\Sigma}$, which is an upper bound of the objective Hessian $\nabla^2 F$. Let $\mathbf{x}^*$ be the minimizer of $F$, we define the second moment matrix of the gradient noise at $\mathbf{x}^*$ as $\mathbf{Q} = \mathbb{E}_{\mathbf{a},b\sim\mathcal{D}}\left(\ell'(\mathbf{a}^\top\mathbf{x}^*, b)\right)^2\mathbf{a}\mathbf{a}^\top$.

**Algorithm 1** **S**tochastic **A**ccelerated **D**ata-Dependent **A**lgorithm (SADA)

**Input:** Initialization $\tilde{\mathbf{x}}_0$, regularization parameters $\{h_k\}_{k=1}^K$, step sizes $\eta$, $\gamma$, and momentum parameters $\{\beta_k\}_{k=1}^K$, $\theta$, $\tilde{\mathbf{x}}_{-1} = \tilde{\mathbf{x}}_0$

**for** $k = 1, 2, \ldots, K$ **do**

  $\tilde{\mathbf{y}}_{k-1} \leftarrow \tilde{\mathbf{x}}_{k-1} + \beta_k(\tilde{\mathbf{x}}_{k-1} - \tilde{\mathbf{x}}_{k-2})$ ▷ Extrapolation

  $\mathbf{x}_0 \leftarrow \tilde{\mathbf{y}}_{k-1}$, $\mathbf{z}_0 \leftarrow \tilde{\mathbf{y}}_{k-1}$

  **for** $t = 1, 2, \ldots, T$ **do** ▷ Inner loop for solving subproblem (4)

    Sample fresh data $(\mathbf{a}_t, b_t) \sim \mathcal{D}$

    $\mathbf{y}_{t-1} \leftarrow \frac{1}{1+\theta}\mathbf{x}_{t-1} + \frac{\theta}{1+\theta}\mathbf{z}_{t-1}$

    $\widehat{\mathbf{g}}_t \leftarrow h_k \ell'(\mathbf{a}_t^\top \tilde{\mathbf{y}}_{k-1}, b_t)\mathbf{a}_t + \left[\mathbf{a}_t^\top(\mathbf{y}_{t-1} - \tilde{\mathbf{y}}_{k-1})\right]\mathbf{a}_t$ ▷ Compute $\widehat{\mathbf{g}}_t$ by streaming data

    $\mathbf{x}_t \leftarrow \mathbf{y}_{t-1} - \eta\widehat{\mathbf{g}}_t$

    $\mathbf{z}_t \leftarrow \theta\mathbf{y}_{t-1} + (1-\theta)\mathbf{z}_{t-1} - \gamma\widehat{\mathbf{g}}_t$

  **end for**

  $\tilde{\mathbf{x}}_k \leftarrow \frac{2}{T}\sum_{t=T/2+1}^{T}\mathbf{x}_t$ ▷ Tail-averaging scheme

**end for**

**return** $\tilde{\mathbf{x}}_K$

Our goals are as follows:

- For optimization complexity, the goal is to *improve* the dependence on the loss condition number $\alpha$ and data condition number $\kappa$, $\tilde{\kappa}$. In particular, we are interested in improving the dependence on $\kappa$. As shown in Remark 1, $\kappa$ can be times larger than $\tilde{\kappa}$.
- For the statistical complexity, our objective is to *preserve* the optimal rate $\alpha\,\mathrm{tr}(\mathbf{H}^{-1}\mathbf{Q})/\varepsilon$. For well-specified linear regression model $b = \mathbf{a}^\top\mathbf{x}^* + \varepsilon_{\mathrm{noise}}$, this expression reduces to $\frac{\sigma^2 d}{\varepsilon}$, where $\sigma^2$ is the variance of the noise $\varepsilon_{\mathrm{noise}}$.
- For additional complexity, it may emerge from the coupling between the optimization process and model mis-specification. We expect this complexity to exhibit a lower-order dependence on $\varepsilon^{-1}$ and $\kappa$ (i.e., to appear as a higher-order term in the excess risk).

## 4 Stochastic Accelerated Data-Dependent Algorithm

We propose the **Stochastic Accelerated Data-Dependent Algorithm (SADA)** to solve (GLP), as summarized in Algorithm 1. SADA combines momentum and data-dependent proximal methods to achieve acceleration in the streaming setting. At a high level, the outer loop iteratively constructs data-dependent proximal subproblems. The proximal term is induced by the data covariance and cannot be accessed explicitly. The inner loop solves the subproblem by using streaming data to approximate the proximal term, and returns the tail-average of the last half iterations.

### 4.1 Inner Loop: Accelerated Solver with Tail-Averaging

The inner loop adopts momentum to accelerate the convergence of optimization error and a tail-averaging scheme to reduce variance. At the $k$-th outer iteration, let $\tilde{\mathbf{y}}_{k-1}$ denote the extrapolated variable. The inner loop solves the following data-dependent proximal subproblem:

$$\min_{\mathbf{x}\in\mathbb{R}^d} \mathbb{E}_{\mathbf{a},b\sim\mathcal{D}} h_k\langle \ell'(\mathbf{a}^\top\tilde{\mathbf{y}}_{k-1}, b)\mathbf{a}, \mathbf{x} - \tilde{\mathbf{y}}_{k-1}\rangle + \frac{1}{2}\|\mathbf{x} - \tilde{\mathbf{y}}_{k-1}\|_{\mathbf{\Sigma}}^2, \tag{4}$$

where $\ell'(\cdot,\cdot)$ denotes the first-order partial derivative of $\ell$ with respect to its first argument.

Since population covariance $\mathbf{\Sigma}$ is inaccessible, the inner loop uses fresh samples $(\mathbf{a}_t, b_t) \sim \mathcal{D}$ to approximate $\mathbf{\Sigma}$ by $\mathbf{a}_t\mathbf{a}_t^\top$. This resembles linear regression with data pair $(\mathbf{a}, -h_k\ell'(\mathbf{a}^\top\tilde{\mathbf{y}}_{k-1}, b))$. However, model misspecification arises, whose analysis is non-trivial. Previous work has focused only on the well-specified setting. To address this gap, we propose a Layer-Peeled Decomposition to characterize its stationary distribution. Details are shown in Section 6.1.

### 4.2 Outer Loop: Data-Dependent Proximal Method with Acceleration

The outer loop iteratively constructs the proximal subproblem (4) and uses momentum to accelerate convergence. Since acceleration often amplifies the error, we need to carefully control the inner loop optimization accuracy. SADA employs a two-phase step size to accelerate convergence while controlling stochastic noise: a large constant step size in the first phase rapidly reduces the optimization error, followed by a gradually decaying step size in the second phase to control stochastic noise.

## 5 Convergence Result and Algorithmic Complexity

**Hyperparameter Choice.** We choose inner loop hyperparameters as follows:

$$\eta \leqslant \frac{1}{16R^2}, \quad \gamma = \frac{1}{4}\sqrt{\frac{\eta}{\tilde{\kappa}\mu}}, \quad \theta = \frac{1}{4}\sqrt{\frac{\mu\eta}{\tilde{\kappa}}}, \quad T \geqslant \tilde{\Omega}\left(\sqrt{\frac{\tilde{\kappa}}{\mu\eta}}\right). \tag{5}$$

For the outer loop, we set step size $h_k = \frac{2\alpha\tilde{\theta}_k^2}{L_\ell}$, and momentum parameters $\beta_k = \frac{1-\tilde{\theta}_k}{1+\tilde{\theta}_k}$, where

$$\tilde{\theta}_{\max} \asymp \min\left\{\frac{1}{\sqrt{\alpha}}, \frac{T}{\alpha^2\tilde{\kappa}}\right\}, \quad \tilde{\theta}_k = \begin{cases} \tilde{\theta}_{\max}, & k \leqslant K/2, \\ \frac{4}{4/\tilde{\theta}_{\max} + k - K/2}, & k > K/2. \end{cases} \tag{6}$$

*Remark* 2 (Interpretation of Hyperparameters). In the above setting, the regularization parameter $h_k$ can be viewed as the step size of the outer loop. $\eta$ is required to be smaller than $1/R^2$ to guarantee convergence, and $\gamma$ and $\theta$ are acceleration hyperparameters, which are proposed in stochastic coordinate descent (Nesterov, 2012) and linear regression (Jain et al., 2018a; Li et al., 2024). The inner loop requires $T \geqslant \tilde{\Omega}(\sqrt{\tilde{\kappa}/(\mu\eta)})$ for sufficient convergence. In the outer loop, the step size $h_k$ is bounded by $1/L_\ell$ and an additional term $\frac{T^2}{L_\ell\alpha^3\tilde{\kappa}^2}$ imposed by additional noise. In the first phase ($k \leqslant K/2$), we choose the maximum step size to accelerate the convergence; in the second phase ($k > K/2$), we carefully decay the step size to control noise.

*Remark* 3. For the Gaussian and sub-Gaussian setting in Remark 1, the inner loop hyperparameters are $\eta \lesssim \frac{1}{\operatorname{tr}\mathbf{\Sigma}}$, $\gamma \asymp \sqrt{\frac{\eta}{\mu d}}$, $\theta \asymp \sqrt{\frac{\mu\eta}{d}}$ and $T \gtrsim \tilde{\Omega}\left(\frac{1}{\mu\gamma}\right)$. For the worst case where $\frac{\operatorname{tr}\mathbf{\Sigma}}{\mu} = \kappa(\mathbf{\Sigma})d$, step size $\gamma$ is larger than $\eta$ by a factor of $\sqrt{\kappa(\mathbf{\Sigma})}$. Since $\eta$ is also the step size of plain SGD (Jain et al., 2018b; Zou et al., 2023), this suggests an acceleration by a factor of $\sqrt{\kappa(\mathbf{\Sigma})}$ in the convergence rate.

**Theorem 5.1.** *Suppose Assumptions 1, 2 and 3 hold. Let Algorithm 1 start from $\tilde{\mathbf{x}}_0$, and choose the hyperparameters as specified above in* (5) *and* (6)*. Let $\tilde{\mathbf{x}}_K$ denote the output of Algorithm 1 after $K$ outer iterations, each consisting of $T \geqslant \tilde{\Omega}(\sqrt{\tilde{\kappa}/(\mu\eta)})$ inner updates. Then we have*

$$\mathbb{E}F(\tilde{\mathbf{x}}_K) - F(\mathbf{x}^*) \lesssim \exp\left(-\frac{c_0 K}{\sqrt{\alpha} + \alpha^2\tilde{\kappa}/T}\right)\left(F(\tilde{\mathbf{x}}_0) - F(\mathbf{x}^*)\right) + \frac{\alpha\operatorname{tr}(\mathbf{H}^{-1}\mathbf{Q})}{n} + \frac{\alpha\eta\tilde{\kappa}\operatorname{tr}\mathbf{Q}}{L_\ell n},$$

*where $n = KT$ is the sample size and $c_0$ is a universal constant.*

**Corollary 5.1** (Sample Complexity)**.** *Under the setting of Theorem 5.1, to obtain a solution $\tilde{\mathbf{x}}_K$ with excess risk at most $\varepsilon$, the required sample size $n = KT$ is*

$$\tilde{\mathcal{O}}\Bigg(\underbrace{\left(\sqrt{\alpha\kappa\tilde{\kappa}} + \alpha^2\tilde{\kappa}\right)}_{\textit{Optimization term}} + \underbrace{\frac{\alpha\operatorname{tr}(\mathbf{H}^{-1}\mathbf{Q})}{\varepsilon}}_{\textit{Statistical term}} + \underbrace{\left(\frac{\alpha^2\tilde{\kappa}^2\operatorname{tr}\mathbf{Q}}{L_\ell\mu\varepsilon}\right)^{1/3}}_{\textit{Mis-specification term}}\Bigg).$$

*We choose the hyperparameter as $\tilde{\Theta}\left(\frac{\alpha\tilde{\kappa}}{\mu n^2}\right) \leqslant \eta \leqslant \tilde{\Theta}\left(\min\left\{\frac{1}{R^2}, \frac{n\varepsilon L_\ell}{\alpha\tilde{\kappa}\operatorname{tr}\mathbf{Q}}\right\}\right)$, $T = \tilde{\Theta}(\sqrt{\tilde{\kappa}/(\mu\eta)})$, $K = n/T$ and other hyperparameters as specified in* (5) *and* (6)*.*

*Remark* 4. The sample bound ensures an admissible $\eta$ (Appendix E.3). While the choice of $\eta$ depends on the pre-known sample size $n$, it can be extended to an infinite-horizon setting (where the algorithm may be stopped at any time) using the doubling trick (Hazan & Kale, 2014), incurring only logarithmic overhead (see Appendix J for more discussion). Moreover, the choice of hyperparameters requires pre-knowing some problem-dependent parameters, which is common in optimization methods (Nesterov, 2018). Such dependence may be removed by restarting (O'donoghue & Candes, 2015; Necoara et al., 2019) or line-search procedures (Bubeck et al., 2015). We leave further investigation for future work.

**Optimization Term.** The optimization complexity consists of two terms. The first term, $\sqrt{\alpha\kappa\tilde{\kappa}}$, corresponds to an accelerated decay of the initial error. This rate reflects a double acceleration effect, induced by momentum mechanisms at both the outer and inner loops of Algorithm 1. The second term $\alpha^2\tilde{\kappa}$ arises from the additional noise caused by localization with magnitude proportional to the distance $\|\tilde{\mathbf{y}}_k - \mathbf{x}^*\|_{\mathbf{\Sigma}}^2$. We improve the complexity from previous $\alpha^2\kappa$ in VR methods to $\alpha^2\tilde{\kappa}$.

As a representative example, consider the Gaussian case in Remark 1 and assume outer condition number $\alpha \asymp 1$ (e.g., linear regression or logistic regression with normalized features and constant level optimal solution). In this case, Algorithm 1 achieves an optimization complexity of order $\mathcal{O}((d\operatorname{tr}\mathbf{\Sigma}/\mu)^{1/2})$, improving over the unaccelerated rate $\mathcal{O}(\operatorname{tr}\mathbf{\Sigma}/\mu)$ (note that $d \leqslant \operatorname{tr}\mathbf{\Sigma}/\mu$). In the worst case, $\operatorname{tr}\mathbf{\Sigma}/\mu = \kappa(\mathbf{\Sigma})d$, so our algorithm improves the optimization convergence rate by a factor of $\sqrt{\kappa(\mathbf{\Sigma})}$. The improvement is significant when $\mathbf{\Sigma}$ is poorly conditioned.

**Statistical Term.** The statistical term corresponds to the noise covariance $\mathbf{Q}$ at the minimizer $\mathbf{x}^*$ suppressed along the eigen-direction of the lower bound of the objective Hessian, which is $\mathbf{H}/\alpha$. Specifically, the noise $\mathbf{Q}$ is filtered through the local curvature: directions with larger curvature contract the noise, whereas flatter directions amplify it. This statistical term cannot be improved without additional assumptions on the third-order smoothness of $\ell$. We construct a problem class in Appendix G whose minimax risk is lower bounded by $\Omega(\alpha\operatorname{tr}(\mathbf{H}^{-1}\mathbf{Q})/n)$, thereby establishing the worst-case optimality of our bound.

**Mis-specification Term.** The mis-specification term is a coupling effect of the approximation error $\mathbf{a}\mathbf{a}^\top \neq \mathbf{\Sigma}$ (also known as fourth moment effect (Jain et al., 2018a; Li et al., 2024)) and the possibility that the noise covariance $\mathbf{Q} \neq c\mathbf{H}$ for any $c \geqslant 0$. Note that the complexity does not depend on $\kappa$, and on $\alpha^{2/3}$, $\mu^{-1/3}$, and $\varepsilon^{-1/3}$. For high accuracy with $\varepsilon$, a small step size is required. Theorem 5.1 implies that the misspecification error vanishes asymptotically. For well-specified models with $\mathbf{Q} = \sigma^2\mathbf{\Sigma}$, the mis-specification term can be absorbed into the statistical term with some additional analysis. This holds directly for Gaussian and sub-Gaussian distributions in Remark 1 by applying $\kappa \asymp \operatorname{tr}\mathbf{\Sigma}/\mu$ and $\tilde{\kappa} \asymp d$ to the bound in Theorem 5.1, so the mis-specification term $\frac{\alpha\eta\tilde{\kappa}\operatorname{tr}\mathbf{Q}}{L_\ell n} = \frac{\sigma^2\alpha\eta d\operatorname{tr}\mathbf{\Sigma}}{n} \leqslant \frac{\sigma^2\alpha d}{n} = \frac{\alpha\operatorname{tr}(\mathbf{H}^{-1}\mathbf{Q})}{n}$, where we use $\eta\operatorname{tr}\mathbf{\Sigma} \leqslant 1$.

**Answers to Open Problem.** Our result answers the open problem by Jain et al. (2018a) in the General Linear Prediction setting with minimum assumptions. Concretely, we show that SADA achieves the asymptotic optimal convergence rate in the following sense:

$$\limsup_{n\to\infty} \frac{\mathbb{E}F(\mathbf{x}_n^{\mathrm{SADA}}) - F(\mathbf{x}^*)}{\operatorname{tr}(\mathbf{H}^{-1}\mathbf{Q})/n} = \mathcal{O}(\alpha),$$

where, as discussed above, the term $\mathcal{O}(\alpha)$ is optimal in our setting.

## 6 Proof Sketch

### 6.1 Part I: Analysis of Inner Loop

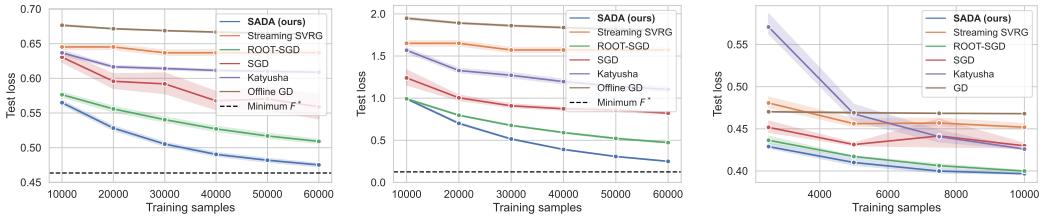
The subproblem (4) resembles linear regression with data pair $(\mathbf{a}, -h_k\ell'(\mathbf{a}^\top\tilde{\mathbf{y}}_{k-1}, b))$. We denote the minimizer of (4) by $\tilde{\mathbf{x}}_k^*$. Applying the standard bias-variance decomposition (Dieuleveut & Bach, 2016; Jain et al., 2018a; Zou et al., 2023; Wu et al., 2022; Li et al., 2024), we can decompose the dynamics of $\boldsymbol{\eta}_t \overset{\text{def}}{=} \begin{pmatrix}\mathbf{x}_t - \tilde{\mathbf{x}}_k^* \\ \mathbf{y}_t - \tilde{\mathbf{x}}_k^*\end{pmatrix} \in \mathbb{R}^{2d}$ into the sum of bias term $\boldsymbol{\eta}_t^{\mathrm{bias}}$ and variance term $\boldsymbol{\eta}_t^{\mathrm{var}}$:

$$\begin{cases}\boldsymbol{\eta}_t^{\mathrm{bias}} = \widehat{\mathbf{A}}_t\boldsymbol{\eta}_{t-1}^{\mathrm{bias}}, & \boldsymbol{\eta}_0^{\mathrm{bias}} = \boldsymbol{\eta}_0, \\ \boldsymbol{\eta}_t^{\mathrm{var}} = \widehat{\mathbf{A}}_t\boldsymbol{\eta}_{t-1}^{\mathrm{var}} + \boldsymbol{\zeta}_t, & \boldsymbol{\eta}_0^{\mathrm{var}} = \mathbf{0}, \end{cases} \quad \text{where} \quad \widehat{\mathbf{A}}_t = \begin{pmatrix}\mathbf{O} & \mathbf{I} - \eta\mathbf{a}_t\mathbf{a}_t^\top \\ -\frac{1-\theta}{1+\theta}\mathbf{I} & \frac{2}{1+\theta}\mathbf{I} - \frac{\eta+\theta\gamma}{1+\theta}\mathbf{a}_t\mathbf{a}_t^\top\end{pmatrix}.$$

The inner loop output $\tilde{\mathbf{x}}_k$ can be decomposed as $\tilde{\mathbf{x}}_k = \tilde{\mathbf{x}}_k^* + h_k\mathbf{r}_k + h_k\mathbf{v}_k$, where $\mathbf{r}_k \in \mathbb{R}^d$ is the bias term, and $\mathbf{v}_k \in \mathbb{R}^d$ is the noise term with zero mean. The overall result in the inner loop is shown in Lemma D.1. The bound of $\mathbf{r}_k$ follows from Jain et al. (2018a) using standard acceleration technique shown in Appendix D.3. The main focus is to derive a bound of $\|\mathbf{v}_k\|_{\mathbf{H}}^2$, which is non-trivial.

**Bound of $\mathbf{v}_k$.** We need to analyze the covariance $\mathbf{C}_t = \mathbb{E}\left(\boldsymbol{\eta}_t^{\mathrm{var}}(\boldsymbol{\eta}_t^{\mathrm{var}})^\top\right)$, which satisfies

$$\mathbf{C}_t = \mathcal{B} \circ \mathbf{C}_{t-1} + \begin{pmatrix}\eta^2\mathbf{R} & \eta q\mathbf{R} \\ \eta q\mathbf{R} & q^2\mathbf{R}\end{pmatrix}, \quad \mathbf{C}_0 = \mathbf{O}.$$




(a) Simulation, Logistic Regression (b) Simulation, Linear Regression (c) CIFAR-10, Logistic Regression

Figure 1: Comparison of SADA with other algorithms. The x-axis represents the number of training samples used by each method. SADA achieves lower test loss than the other methods. Shaded regions indicate 95% confidence intervals computed over independent runs.

where $\mathbf{R} \in \mathbb{R}^{d\times d}$ is the noise covariance and $\mathcal{B} : \mathbf{C} \mapsto \mathbb{E}\widehat{\mathbf{A}}_t\mathbf{C}\widehat{\mathbf{A}}_t^\top$ is a linear matrix operator acts on $\mathbb{R}^{2d\times 2d}$ and $\circ$ represents the operation of a linear matrix operator on a matrix. Note that in prior work on well-specified linear regression, where $\mathbf{R} = \sigma^2\mathbf{\Sigma}$ commutes with $\mathbf{\Sigma}$, the stationary property of $\mathbf{C}_t$ is relatively simple to analyze. In contrast, our setting necessitates understanding the behavior of $\mathbf{C}_t$ for a general $\mathbf{R}$. We propose the Layer-Peeled Decomposition, which decomposes $\mathbf{C}_t$ into layers of dynamics $\tilde{\mathbf{C}}_t$ and $\tilde{\mathbf{C}}_t^{(\ell)}$. Specifically, we introduce $\mathbf{C}_t = \tilde{\mathbf{C}}_t + \sum_{\ell=0}^{T}\tilde{\mathbf{C}}_t^{(\ell)}$, where

$$\begin{aligned}
&\tilde{\mathbf{C}}_t = \mathbf{A}\tilde{\mathbf{C}}_{t-1}\mathbf{A} + \mathbf{R}, && \tilde{\mathbf{C}}_0 = \mathbf{O}, && \tilde{\mathbf{C}}_t^{(0)} = \tilde{\mathbf{C}}_t,\\
&\tilde{\mathbf{C}}_t^{(\ell)} = \mathbf{A}\tilde{\mathbf{C}}_{t-1}^{(\ell)}\mathbf{A} + (\mathcal{B} - \mathbf{A}\otimes\mathbf{A})\circ\tilde{\mathbf{C}}_t^{(\ell-1)}, && \tilde{\mathbf{C}}_0^{(\ell)} = \mathbf{O}, && \text{for } \ell = 1, 2, \dots
\end{aligned}$$

and $\mathbf{A} = \mathbb{E}\widehat{\mathbf{A}}_t$. Intuitively, $\tilde{\mathbf{C}}_t$ can be viewed as the covariance dynamics by replacing $\mathbf{a}\mathbf{a}^\top$ by $\mathbf{\Sigma}$, and $\tilde{\mathbf{C}}_t^{(\ell)}$ collects these approximation errors. In order to study $\tilde{\mathbf{C}}_t$ and $\tilde{\mathbf{C}}_t^{(\ell)}$, we introduce a core auxiliary dynamics $\tilde{\mathbf{\Pi}}_t(\mathbf{M}) \in \mathbb{R}^{2d\times 2d}$ for $\mathbf{M} \in \mathbb{R}^{d\times d}$ as follows:

$$\tilde{\mathbf{\Pi}}_t(\mathbf{M}) = \mathbf{A}\tilde{\mathbf{\Pi}}_{t-1}(\mathbf{M})\mathbf{A} + \begin{pmatrix}\eta^2\mathbf{M} & \eta q\mathbf{M}\\ \eta q\mathbf{M} & q^2\mathbf{M}\end{pmatrix}, \quad \tilde{\mathbf{\Pi}}_0(\mathbf{M}) = \mathbf{O}. \tag{7}$$

Because we have $\tilde{\mathbf{C}}_t = \tilde{\mathbf{\Pi}}_t(\mathbf{R})$, and $\tilde{\mathbf{C}}_{t-1}^{(\ell)}$ can be bounded by $\tilde{\mathbf{\Pi}}_t(\mathbf{R}^{(\ell)})$ for some $\mathbf{R}^{(\ell)}$ which relates to the stationary state of $\mathbf{C}_t^{(\ell-1)}$. As a result, we can recursively bound $\tilde{\mathbf{C}}_t^{(\ell)}$ and its limit.

The detailed proof for bounding the variance is provided in Appendix D.2. In Definition D.2, we introduce a linear mapping $L_{\text{var}}$ to calculate the loss of a general covariance matrix sequence. In Appendix D.2.2, we study the dynamics of (7). We obtain a sharp upper bound of $L_{\text{var}}(\{\tilde{\mathbf{\Pi}}_t(\mathbf{M})\}_{t\in\mathbb{N}})$ in Lemma D.13 and stationary covariance $\tilde{\mathbf{\Pi}}_\infty(\mathbf{M}) = \lim_{t\to\infty}\tilde{\mathbf{\Pi}}_t(\mathbf{M})$ in Lemma D.16. In Appendix D.2.3, we study the noise covariance matrix $\mathbf{R}$, and decompose $\operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{R})$ and $\operatorname{tr}(\mathbf{R})$. Finally, in Appendix D.2.5, we introduce $\tilde{\mathbf{C}}_t$ and $\tilde{\mathbf{C}}_t^{(\ell)}$. We first show that the variance can be further decomposed as follows: $\mathbb{E}\|\mathbf{v}_k\|_{\mathbf{\Sigma}}^2 = L_{\text{var}}(\{\tilde{\mathbf{C}}_t\}_{t\in\mathbb{N}}) + \sum_{\ell=1}^{T}L_{\text{var}}(\{\tilde{\mathbf{C}}_t^{(\ell)}\}_{t\in\mathbb{N}})$ in Lemma D.22. Then, by plugging in Lemma D.27, we have the final bound of variance.

### 6.2 Part II: Analysis of Outer Loop

The bounds of the outer loop mainly exploit the structure of the solution $\tilde{\mathbf{x}}_k$ returned by the inner loop. Our proof technique follows the analysis of AGD, but with a careful control of the noise term.

**Lemma 6.1.** *Suppose we have* $h_k \lesssim \min\left\{\frac{1}{L_\ell}, \frac{T^2}{L_\ell\alpha^3\tilde{\kappa}^2}\right\}$ *and set* $\theta_k = \sqrt{\frac{h_k}{2\alpha}}$. *Let the energy* $L_k = \mathbb{E}F(\tilde{\mathbf{x}}_k) - F(\mathbf{x}^*) + \frac{2\theta_k^2}{3L_\ell h_k}\mathbb{E}_{k-1}\|\tilde{\mathbf{z}}_k - \mathbf{x}^*\|_{\mathbf{H}}^2$, *then we have*

$$L_k \leqslant \left(1 - \frac{\theta_k}{2}\right)L_{k-1} + \frac{3h_k\sigma_0^2}{2}, \quad \text{where } \sigma_0^2 \lesssim \frac{L_\ell\operatorname{tr}(\mathbf{H}^{-1}\mathbf{Q})}{T} + \frac{\eta\tilde{\kappa}\operatorname{tr}\mathbf{Q}}{T}.$$

The two-phase step size schedule (6) uses the largest step to contract initial error for Phase I ($t \leqslant K/2$), then $h_{k+K/2} \asymp 1/(\mu_\ell(a+k)^2)$ to control accumulated noise for Phase II ($t > K/2$).

## 7 EXPERIMENTS

We conduct numerical simulation and experiments on CIFAR-10 and MNIST to evaluate the empirical performance of SADA and confirm the theoretical insights. Due to space limit, we defer the results on MNIST to Appendix I.1.1 and detailed implementation to Appendix I.2.

**Numerical Simulation.** We consider logistic regression (Figure 1(a)) and linear regression (Figure 1(b)) with $d = 20$ and covariance $\mathbf{\Sigma} = \mathrm{diag}\{k^{-4}\}_{k=1}^{d}$. We set $\mathbf{x}^*$ such that $x_k^* = 2/\sqrt{dk^{-4}}$, where $\lambda_k = k^{-4}$, and draw $\mathbf{a} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. For logistic regression, label $b = \pm 1$ follows $\mathbb{P}(b = 1 \mid \mathbf{a}) = 1/(1 + \exp(-\mathbf{a}^\top \mathbf{x}^*))$. For linear regression, $b = \mathbf{a}^\top \mathbf{x}^* + 0.5\varepsilon$, where $\varepsilon \sim \mathcal{N}(0, 1)$. For each sample size, we select the learning rate via grid search. As shown in Figure 1(a)–(b), SADA consistently attains lower test loss than other methods across a wide range of sample sizes.

**Real Dataset.** We further evaluate the methods on binary logistic regression using CIFAR-10 (Krizhevsky et al., 2009) on the first two classes (Figure 1(c)). For each sample size, we select the hyperparameters via grid search. The results in Figure 1(c) show that SADA achieves competitive or lower test loss compared with the other methods across different sample sizes.

## 8 EXTENSIONS AND DISCUSSIONS

In this section, we present extensions of SADA. Due to space limitation, we defer the complexity of SADA with mini-batching and its parallel implementation to Appendix J.

**Non-strongly Convex Case.** We apply the reduction technique to extend our results to the non-strongly convex case. Suppose $\|\mathbf{x}^*\| \leqslant D$ and $\|\mathbf{x}^*\|_{\mathbf{H}} \leqslant M$, then we optimize the surrogate function:

$$G(\mathbf{x}) = \mathbb{E}_{\mathbf{a},b\sim\mathcal{D}}\left(\ell(\mathbf{a}^\top \mathbf{x}, b) + \underbrace{\frac{\varepsilon}{2M^2}\left(\mathbf{a}^\top \mathbf{x}\right)^2}_{\text{Term (A)}}\right) + \underbrace{\frac{\varepsilon}{2D^2}\|\mathbf{x}\|^2}_{\text{Term (B)}},$$

where Term (A) is needed when $\ell$ is non-strongly convex ($\mu_\ell \leqslant \varepsilon$), and Term (B) is required when $\mathbf{\Sigma}$ is poor-conditioned ($\mu \leqslant \varepsilon$). The algorithm SADA-NonSC (Algorithm 2) is deferred to Appendix H.

**Corollary 8.1** (Sample Complexity)**.** *Under the setting of Theorem 5.1. The required sample size is*

$$\left(\sqrt{\frac{L_\ell R^2 \tilde{\kappa}}{\mu'_\ell \mu'}} + \frac{L_\ell^2 \tilde{\kappa}}{(\mu'_\ell)^2}\right) + \frac{L_\ell \,\mathrm{tr}((\mathbf{H} + \varepsilon \mathbf{I}/D^2)^{-1}\mathbf{Q})}{\mu'_\ell \varepsilon} + \left(\frac{L_\ell \tilde{\kappa}^2 \,\mathrm{tr}\,\mathbf{Q}}{(\mu'_\ell)^2 \mu' \varepsilon}\right)^{1/3},$$

*where* $\mu'_\ell = \max\{\mu_\ell, \varepsilon\}$ *and* $\mu' = \max\{\mu, \varepsilon\}$.

**Use of Unlabeled Data.** The inner loop covariance estimate $\mathbf{a}\mathbf{a}^\top$ needs no label. Hence, we can use $m$ unlabeled samples to improve the estimate (see Algorithm 3 in Appendix H). We can improve $R^2$, $\kappa$ and $\tilde{\kappa}$ to $R_m^2 = \frac{R^2 + m\lambda_{\max}(\mathbf{\Sigma})}{m+1}$, $\kappa_m = \frac{R_m^2}{\mu}$ and $\tilde{\kappa}_m = \frac{\tilde{\kappa}+m}{m+1}$. The sample complexity is

$$n \geqslant \tilde{\Theta}\left(\left(\sqrt{\alpha \kappa_m \tilde{\kappa}_m} + \alpha^2 \tilde{\kappa}\right) + \frac{\alpha \,\mathrm{tr}(\mathbf{H}^{-1}\mathbf{Q})}{\varepsilon} + \left(\frac{\alpha^2 \tilde{\kappa}_m^2 \,\mathrm{tr}\,\mathbf{Q}}{L_\ell \mu \varepsilon}\right)^{1/3}\right).$$

The total unlabeled sample size is at least $\tilde{\Omega}(m\sqrt{\alpha\tilde{\kappa}_m/\mu\eta})$. We set $\eta = \tilde{\Theta}\left(\min\left\{\frac{1}{R_m^2}, \frac{n\varepsilon L_\ell}{\alpha \tilde{\kappa}_m \,\mathrm{tr}\,\mathbf{Q}}\right\}\right)$, $K = \tilde{\Theta}(\sqrt{\alpha})$ and $T = n/K$. The redundancy in the labeled-data complexity can be attributed to optimization and mis-specification errors, and is quantified directly by $R_m^2$, $\kappa_m$, and $\tilde{\kappa}_m$.

**High-order Smoothness Assumption.** Our analysis, like the standard gradient-based optimization and local Rademacher complexity framework, assumes no third-order smoothness of $\ell$. Under this minimal regularity, the obtained statistical rate is minimax optimal. If Hessian smoothness is given, one may expect the sharper rate $\mathrm{tr}((\nabla^2 F(\mathbf{x}^*))^{-1}\mathbf{Q})/n$, matching the asymptotic efficiency of ERM and Cramér-Rao lower bound. A direct approach would be to first run SADA until the iterate enters a neighborhood of $\mathbf{x}^*$, e.g., $\|\tilde{\mathbf{x}} - \mathbf{x}^*\| \leqslant \frac{1}{\mu_\ell \mu}$, then replace $\mathbf{a}\mathbf{a}^\top$ by $\ell''(\mathbf{a}^\top \tilde{\mathbf{x}}, b)\mathbf{a}\mathbf{a}^\top$ and use a carefully chosen step size. However, the analysis is much more involved. Moreover, higher-order methods, such as cubic regularization (Nesterov & Polyak, 2006), could be more appropriate as it achieves faster rate for the outer loop when the Hessian is smooth. We leave this as future work.

### AI USE STATEMENT

In this work, we used generative AI tools to aid or polish writing. We have not used generative AI tools for retrieval and discovery, for research ideation or execution, to draft sections of the paper, for proving mathematical claims and for generating synthetic datasets are not applicable to this work. We have reviewed all AI-assisted work. We checked LLM-generated text and verified the correctness. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

### REPRODUCIBILITY STATEMENT

Our theoretical results are supported by complete proofs sketched in Section 6 and fully provided in Appendix B-H. The proofs in Appendix are established under more general assumptions stated in Appendix B and the results under Assumptions 1, 2 and 3 of the main paper follow as special cases. The details are in Appendix B. For the empirical evaluation, we provided details in Appendix I. These details are intended to facilitate the reproduction of all experimental results reported in the main paper.

## REFERENCES

Alekh Agarwal, Peter L. Bartlett, Pradeep Ravikumar, and Martin J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012.

Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *Journal of Machine Learning Research*, 18(221):1–51, 2018.

Zeyuan Allen-Zhu and Elad Hazan. Variance reduction for faster non-convex optimization. In *International Conference on Machine Learning*, 2016.

Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Neural Information Processing Systems*, 2013.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

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

Alexandre Défossez and Francis Bach. Averaged least-mean-squares: Bias-variance trade-offs and optimal sampling distributions. In *Artificial Intelligence and Statistics*, 2015.

Aymeric Dieuleveut and Francis Bach. Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics*, 44(4):1363–1399, 2016.

Aymeric Dieuleveut, Nicolas Flammarion, and Francis Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *Journal of Machine Learning Research*, 18(101):1–51, 2017.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Neural Information Processing Systems*, 2018.

Cong Fang, Zhouchen Lin, and Tong Zhang. Sharp analysis for nonconvex SGD escaping from saddle points. In *Conference on Learning Theory*, 2019.

Roy Frostig, Rong Ge, Sham M Kakade, and Aaron Sidford. Competing with the empirical risk minimizer in a single pass. In *Conference on Learning Theory*, 2015.

Rong Ge, Sham M Kakade, Rahul Kidambi, and Praneeth Netrapalli. The step decay schedule: A near optimal, geometrically decaying learning rate procedure for least squares. In *Neural Information Processing Systems*, 2019.

Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: Optimal algorithms for stochastic strongly-convex optimization. *The Journal of Machine Learning Research*, 15(1): 2489–2512, 2014.

Prateek Jain, Sham M Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Accelerating stochastic gradient descent for least squares regression. In *Conference on Learning Theory*, 2018a.

Prateek Jain, Sham M Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Parallelizing stochastic gradient descent for least squares regression: Mini-batching, averaging, and model misspecification. *Journal of Machine Learning Research*, 18(223):1–42, 2018b.

Chi Jin, Praneeth Netrapalli, and Michael I Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. In *Conference on Learning Theory*, 2018.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-100 and cifar-10 (canadian institute for advanced research), 2009. URL `http://www.cs.toronto.edu/~kriz/cifar.html`. MIT License.

Chris Junchi Li, Wenlong Mou, Martin Wainwright, and Michael Jordan. Root-SGD: Sharp nonasymptotics and asymptotic efficiency in a single algorithm. In *Conference on Learning Theory*, 2022.

Xuheng Li, Yihe Deng, Jingfeng Wu, Dongruo Zhou, and Quanquan Gu. Risk bounds of accelerated SGD for overparameterized linear regression. In *International Conference on Learning Representations*, 2024.

Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. Catalyst acceleration for first-order convex optimization: From theory to practice. *Journal of Machine Learning Research*, 18(212):1–54, 2018.

Yuanshi Liu, Haihan Zhang, Qian Chen, and Cong Fang. Optimal algorithms in linear regression under covariate shift: On the importance of precondition. *arXiv preprint arXiv:2502.09047*, 2025.

Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Neural Information Processing Systems*, 2011.

Jaouad Mourtada. Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *The Annals of Statistics*, 50(4):2157–2178, 2022.

Ion Necoara, Yu Nesterov, and Francois Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical programming*, 175(1):69–107, 2019.

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

Yurii Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Proceedings of the USSR Academy of Sciences*, 269:543–547, 1983.

Yurii Nesterov. *Lectures on Convex Optimization*, volume 137. Springer, 2018.

Yurii Nesterov and Boris T Polyak. Cubic regularization of Newton method and its global performance. *Mathematical programming*, 108(1):177–205, 2006.

Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, 2017.

Brendan O'donoghue and Emmanuel Candes. Adaptive restart for accelerated gradient schemes. *Foundations of computational mathematics*, 15(3):715–732, 2015.

Rui Pan, Yuxing Liu, Xiaoyu Wang, and Tong Zhang. Accelerated convergence of stochastic heavy ball method under anisotropic gradient noise. In *International Conference on Learning Representations*, 2024.

Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Poczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *International Conference on Machine Learning*, 2016.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, pp. 400–407, 1951.

Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss. *The Journal of Machine Learning Research*, 14(1):567–599, 2013.

Shai Shalev-Shwartz and Tong Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. In *International Conference on Machine Learning*, 2014.

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

Martin J Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge university press, 2019.

Jingfeng Wu, Difan Zou, Vladimir Braverman, Quanquan Gu, and Sham Kakade. Last iterate risk bounds of SGD with decaying stepsize for overparameterized linear regression. In *International Conference on Machine Learning*, 2022.

Haihan Zhang, Yuanshi Liu, Qianwen Chen, and Cong Fang. The optimality of (accelerated) SGD for high-dimensional quadratic optimization. *arXiv preprint arXiv:2409.09745*, 2024.

Haihan Zhang, Weicheng Lin, Yuanshi Liu, and Cong Fang. Learning curves of stochastic gradient descent in kernel regression. In *Forty-Second International Conference on Machine Learning*, 2025.

Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham M Kakade. Benign overfitting of constant-stepsize SGD for linear regression. *Journal of Machine Learning Research*, 24 (326):1–58, 2023.

# CONTENTS


**1 Introduction** **1**

1.1 Review: Previous Results . . . . . 2

1.2 Our Results and Implications . . . . . 3

**2 Related Work** **3**

**3 Problem Setup** **4**

3.1 Summary of Problem-Dependent Quantities . . . . . 4

**4 Stochastic Accelerated Data-Dependent Algorithm** **5**

4.1 Inner Loop: Accelerated Solver with Tail-Averaging . . . . . 5

4.2 Outer Loop: Data-Dependent Proximal Method with Acceleration . . . . . 6

**5 Convergence Result and Algorithmic Complexity** **6**

**6 Proof Sketch** **7**

6.1 Part I: Analysis of Inner Loop . . . . . 7

6.2 Part II: Analysis of Outer Loop . . . . . 8

**7 Experiments** **9**

**8 Extensions and Discussions** **9**

**A Organization** **14**

**B Generalized Assumptions** **15**

**C Hyperparameter Choice** **16**

**D Part I: Analysis of Inner Loop** **16**

D.1 Preliminary . . . . . 17

D.1.1 Bias-variance Decomposition . . . . . 17

D.1.2 Covariance Dynamics . . . . . 18

D.1.3 Properties of Momentum Matrix . . . . . 18

D.2 Variance Upper Bound . . . . . 20

D.2.1 Layer-peeled Decomposition . . . . . 21

D.2.2 Core Auxiliary Dynamics . . . . . 22

D.2.3 Decomposition of Noise Covariance . . . . . 26

D.2.4 Bound of Layer 0 Dynamics $\tilde{\mathbf{C}}_t$ . . . . . 27

D.2.5 Bound of Layer $\ell$ Dynamics $\tilde{\mathbf{C}}_t^{(\ell)}$ . . . . . 27

D.2.6 Proof of Lemma D.6 . . . . . 29

D.3 Bias Upper Bound . . . . . 29
D.4 Proof of Lemma D.1 . . . . . 31
D.5 Auxiliary Lemmas . . . . . 32

**E Part II: Analysis of Outer Loop** **32**
E.1 Excess Risk Bound of Outer Loop . . . . . 32
E.2 Proof of Theorem 5.1 . . . . . 35
E.3 Proof of Corollary 5.1 . . . . . 36

**F Excess Risk of ERM** **36**
F.1 Proof of Theorem F.1 . . . . . 37

**G Worst-case Optimality of Statistical Term** **38**
G.1 Proof of Theorem G.1 . . . . . 38

**H Extensions of SADA** **39**
H.1 SADA for Non-strongly Convex Objectives . . . . . 39
H.2 SADA with Unlabeled Data . . . . . 40
H.3 SADA with Mini-batching and Parallelization . . . . . 41

**I Additional Experiments and Experimental Setting** **41**
I.1 Additonal Experiments . . . . . 42
I.1.1 MNIST . . . . . 42
I.1.2 Ablation Study . . . . . 42
I.1.3 Hyperparameter Sensitivity Analysis . . . . . 42
I.2 Experimental Setting . . . . . 42
I.2.1 Numerical Simulation . . . . . 42
I.2.2 Real Dataset . . . . . 43
I.2.3 Algorithmic Details and Search Ranges . . . . . 43

**J Additional Discussion** **44**
J.1 Doubling Trick . . . . . 44
J.2 Hyperparameter Choice . . . . . 44

# A ORGANIZATION

The appendix is organized as follows:

- In Appendix B, we present two general assumptions (Assumptions 4 and 5) to characterize the stochastic gradient noise of objective $F$. Under Assumptions 1, 2 and 3 in the main text, Assumptions 4 and 5 hold with $B = \alpha R^2$ and $L = \alpha\tilde{\kappa}$ (see Remark 5). The results in the Appendix are stated with dependence on $B$ and $L$, and immediately transferred to the results in the main text by choosing $B = \alpha R^2$ and $L = \alpha\tilde{\kappa}$.

- In Appendix C, we detail the hyperparameter choice with explicit dependence on the constant and logarithmic factors.
- In Appendix D, we present the analysis of inner loop of Algorithm 1.
- In Appendix E, we present the analysis of outer loop of Algorithm 1.
- In Appendix F, we present a proof of the non-asymptotic excess risk upper bound of ERM using the localization technique (Bartlett et al., 2005; Wainwright, 2019).
- In Appendix G, we show that the statistical term in Theorem 5.1 is minimax optimal by constructing a hard problem instance. For this problem, the excess risk must be lower bounded by $\Omega(\alpha \operatorname{tr}(\mathbf{H}^{-1}\mathbf{Q})/n)$.
- In Appendix H, we present algorithms that work for the non-strongly convex case (Algorithm 2) and can use unlabeled data (Algorithm 3).

## B Generalized Assumptions

This section presents two general assumptions that provide a fine characterization of the gradient noise, and then we show that they hold under Assumptions 1, 2, and 3.

**Assumption 4** (Bound of 2-norm of Gradient Noise)**.** *We assume there exists $B > 0$ such that for all $\mathbf{x}$, the following holds:*

$$\mathbb{E}_{\mathbf{a},b\sim\mathcal{D}}\left(\ell'(\mathbf{a}^\top\mathbf{x},b)-\ell'(\mathbf{a}^\top\mathbf{x}^*,b)\right)^2\|\mathbf{a}\|^2 \leqslant 2L_\ell B(F(\mathbf{x})-F(\mathbf{x}^*)).$$

**Assumption 5** (Bound of $\mathbf{\Sigma}^{-1}$-norm of Gradient Noise)**.** *We assume there exists $L > 0$ such that for all $\mathbf{x}$, the following holds:*

$$\mathbb{E}_{\mathbf{a},b\sim\mathcal{D}}\left(\ell'(\mathbf{a}^\top\mathbf{x},b)-\ell'(\mathbf{a}^\top\mathbf{x}^*,b)\right)^2\|\mathbf{a}\|^2_{\mathbf{\Sigma}^{-1}} \leqslant 2L_\ell L(F(\mathbf{x})-F(\mathbf{x}^*)).$$

*Remark* 5. Under Assumption 1, 2 and 3, the above assumptions hold with $B = \alpha R^2$ and $L = \alpha\tilde{\kappa}$, as shown in Lemma B.1 and Lemma B.2. In addition, if a sharper bound is available, then the convergence rate in Theorem 5.1 can be improved. See Theorem E.1 for the possibly sharper convergence rate that explicitly depends on $B$ and $L$.

We verify Assumptions 4 and 5 under Assumptions 1, 2 and 3.

**Lemma B.1** (Verification of Assumption 4)**.** *Suppose Assumptions 1, 2 and 3 hold. Then for any $\mathbf{x} \in \mathbb{R}^d$, we have*

$$\mathbb{E}_{\mathbf{a},b\sim\mathcal{D}}\left(\ell'(\mathbf{a}^\top\mathbf{x},b)-\ell'(\mathbf{a}^\top\mathbf{x}^*,b)\right)^2\|\mathbf{a}\|^2 \leqslant 2L_\ell\alpha R^2(F(\mathbf{x})-F(\mathbf{x}^*)).$$

*Proof.* By Assumption 1, we have $0 \leqslant \ell''(\cdot,\cdot) \leqslant L_\ell$. Thus, $|\ell'(\mathbf{a}^\top\mathbf{x},b)-\ell'(\mathbf{a}^\top\mathbf{x}^*,b)| \leqslant L_\ell|\mathbf{a}^\top(\mathbf{x}-\mathbf{x}^*)|$. By Assumption 3, the left-hand side can be bounded as follows:

$$\begin{aligned}\mathbb{E}_{\mathbf{a},b\sim\mathcal{D}}\left(\ell'(\mathbf{a}^\top\mathbf{x},b)-\ell'(\mathbf{a}^\top\mathbf{x}^*,b)\right)^2\|\mathbf{a}\|^2 &\leqslant L_\ell^2\mathbb{E}_{\mathbf{a},b\sim\mathcal{D}}\left(\mathbf{a}^\top(\mathbf{x}-\mathbf{x}^*)\right)^2\|\mathbf{a}\|^2\\ &\leqslant L_\ell^2R^2\|\mathbf{x}-\mathbf{x}^*\|^2_{\mathbf{\Sigma}} \leqslant 2L_\ell\alpha R^2(F(\mathbf{x})-F(\mathbf{x}^*)).\end{aligned}$$

This completes the proof. □

**Lemma B.2** (Verification of Assumption 5)**.** *Suppose Assumptions 1, 2 and 3 hold. Then for any $\mathbf{x} \in \mathbb{R}^d$, we have*

$$\mathbb{E}_{\mathbf{a},b\sim\mathcal{D}}\left(\ell'(\mathbf{a}^\top\mathbf{x},b)-\ell'(\mathbf{a}^\top\mathbf{x}^*,b)\right)^2\|\mathbf{a}\|^2_{\mathbf{\Sigma}^{-1}} \leqslant 2L_\ell\alpha\tilde{\kappa}(F(\mathbf{x})-F(\mathbf{x}^*)).$$

*Proof.* By Assumptions 1 and 3, the left-hand side can be bounded as follows:

$$\begin{aligned}\mathbb{E}_{\mathbf{a},b\sim\mathcal{D}}\left(\ell'(\mathbf{a}^\top\mathbf{x},b)-\ell'(\mathbf{a}^\top\mathbf{x}^*,b)\right)^2\|\mathbf{a}\|^2_{\mathbf{\Sigma}^{-1}} &\leqslant L_\ell^2\mathbb{E}_{\mathbf{a},b\sim\mathcal{D}}\left(\mathbf{a}^\top(\mathbf{x}-\mathbf{x}^*)\right)^2\|\mathbf{a}\|^2_{\mathbf{\Sigma}^{-1}}\\ &\leqslant L_\ell^2\tilde{\kappa}\|\mathbf{x}-\mathbf{x}^*\|^2_{\mathbf{\Sigma}} \leqslant 2L_\ell\alpha\tilde{\kappa}(F(\mathbf{x})-F(\mathbf{x}^*)).\end{aligned}$$

This completes the proof. □

# C Hyperparameter Choice

In this section, we provide hyperparameter setup that depends on the constants in Assumptions 4 and 5. We also provide explicit constants and logarithmic factors for hyperparameters $T$ and $\tilde{\theta}_{\max}$. We choose the hyperparameters $\eta$ and $T$ such that

$$\eta \leqslant \frac{1}{16R^2}, \quad T \geqslant \sqrt{\frac{\tilde{\kappa}}{\mu\eta}} \log \kappa(\mathbf{\Sigma}) \log \frac{4}{\tilde{\theta}_K^2}.$$

Then choose the hyperparameters of the inner loop as follows:

$$\gamma = \frac{1}{4}\sqrt{\frac{\eta}{\tilde{\kappa}\mu}}, \quad \theta = \frac{1}{4}\sqrt{\frac{\mu\eta}{\tilde{\kappa}}}.$$

For the outer loop, let $L_{\text{eff}} = 160(6L + \tilde{\kappa}(7 + 16\eta B))$, where $B$ and $L$ are defined in Assumptions 4 and 5 (see also Remark 5). We define

$$\tilde{\theta}_{\max} = \min\left\{\sqrt{\frac{1}{12\alpha}}, \frac{T}{12\sqrt{2}\alpha L_{\text{eff}}}\right\}, \quad \tilde{\theta}_k = \begin{cases} \tilde{\theta}_{\max}, & k \leqslant K/2, \\ \dfrac{4}{4/\tilde{\theta}_{\max} + k - K/2}, & k > K/2. \end{cases}$$

Then we set the hyperparameters of the outer loop as follows:

$$h_k = \frac{2\alpha\tilde{\theta}_k^2}{L_\ell}, \quad \beta_k = \frac{1-\tilde{\theta}_k}{1+\tilde{\theta}_k}.$$

# D Part I: Analysis of Inner Loop

**Organization.** In this section, we study the inner loop corresponding to the $k$-th outer iteration of Algorithm 1.

- In Appendix D.1, we introduce preliminary tools for analyzing linear regression. In Appendix D.1.1, we introduce the bias-variance decomposition technique (Dieuleveut & Bach, 2016; Jain et al., 2018a; Zou et al., 2023) to decompose the dynamics into a bias term, which characterizes the optimization error, and a variance term, which characterizes the stochastic noise. In Appendix D.1.2, we derive the dynamics of the bias iterate covariance $\mathbf{B}_t$ and the variance iterate covariance $\mathbf{C}_t$. In Appendix D.1.3, we summarize the results concerning the momentum matrix $\mathbf{A}$, which governs the contraction of the dynamics $\boldsymbol{\eta}_t^{\text{bias}}$ and $\boldsymbol{\eta}_t^{\text{var}}$.
- In Appendix D.2, we present the complete proof for bounding $\mathbf{C}_t$, which is sketched in Section 6.1. In Appendix D.2.2, we propose the core auxiliary dynamics $\tilde{\mathbf{\Pi}}_t(\cdot)$. Then, we derive an upper bound of $L_{\text{var}}(\{\tilde{\mathbf{\Pi}}_t(\mathbf{M})\}_{t\in\mathbb{N}})$ and analyze the stationary covariance $\tilde{\mathbf{\Pi}}_\infty(\mathbf{M})$. By a sharp bound of $\tilde{\mathbf{\Pi}}_t(\cdot)$, we are able to derive the bound of the variance term. In Appendix D.2.4, we apply the tools in Appendix D.2.2 to prove bounds on the layer 0 dynamics $\tilde{\mathbf{C}}_t$. In Appendix D.2.5, we apply the tools in Appendix D.2.2 to prove bounds on the dynamics $\tilde{\mathbf{C}}_t^{(\ell)}$ for $\ell \geqslant 1$.
- In Appendix D.3, we present the proof for bounding the bias iterate $\boldsymbol{\eta}_t^{\text{bias}}$. The proof mainly follows Jain et al. (2018a).
- In Appendix D.4, we present the proof of Lemma D.1, which is the full version of Lemma D.1. This lemma is the basis for the analysis of the outer loop.

To simplify the notation, we denote $\tilde{\mathbf{x}} \overset{\text{def}}{=} \tilde{\mathbf{x}}_{k-1}$, $\tilde{\mathbf{y}} \overset{\text{def}}{=} \tilde{\mathbf{y}}_{k-1}$ and $\tilde{\mathbf{x}}_+ \overset{\text{def}}{=} \tilde{\mathbf{x}}_k$. Let $\tilde{\mathbf{x}}_+^*$ denote the minimizer of the $k$-th outer iteration objective

$$\min_{\mathbf{x}\in\mathbb{R}^d} \mathbb{E}_{\mathbf{a},b\sim\mathcal{D}}\langle \ell'(\mathbf{a}^\top\tilde{\mathbf{y}}_{k-1}, b)\mathbf{a}, \mathbf{x} - \tilde{\mathbf{y}}_{k-1}\rangle + \frac{1}{2h_k}\|\mathbf{x} - \tilde{\mathbf{y}}_{k-1}\|_{\mathbf{\Sigma}}^2.$$

The goal of this section is to derive the following decomposition:

$$\tilde{\mathbf{x}}_+ = \tilde{\mathbf{x}}_+^* + h_k\mathbf{r}_k + h_k\mathbf{v}_k,$$

where $\mathbf{r}_k$ and $\mathbf{v}_k$ are random variables which satisfy the following lemma.

**Lemma D.1.** *Suppose Assumptions 1, 2, 3, 4 and 5 hold. Then*

$$\mathbb{E}\|\mathbf{r}\|_{\mathbf{H}}^2 \leqslant \frac{L_\ell h_k}{8\alpha}\|\nabla F(\tilde{\mathbf{y}}_{k-1})\|_{\mathbf{H}^{-1}}^2, \quad \mathbb{E}\mathbf{v} = \mathbf{0},$$

$$\mathbb{E}\|\mathbf{v}\|_{\mathbf{H}}^2 \leqslant \frac{320L_\ell\left(3L_\ell \operatorname{tr}(\mathbf{H}^{-1}\mathbf{Q}) + 8\eta\tilde{\kappa}\operatorname{tr}\mathbf{Q}\right)}{T} + \frac{160L_\ell^2(6L + \tilde{\kappa}(7+16\eta B))(F(\tilde{\mathbf{y}}) - F(\mathbf{x}^*))}{T},$$

*where the expectation is taken with respect to the samples drawn in the $k$-th outer iteration.*

Without loss of generality, we assume $h_k = 1$ in the following of this section. From now on, we choose a basis that diagonalizes $\mathbf{\Sigma}$. Therefore, $\mathbf{\Sigma}$ and $\mathbf{H}$ are diagonal matrices.

## D.1 Preliminary

### D.1.1 Bias-variance Decomposition

We first provide a concise matrix form of the inner loop update.

**Lemma D.2.** *The update of the inner loop has the following form:*

$$\begin{pmatrix}\mathbf{x}_t - \tilde{\mathbf{x}}_+^* \\ \mathbf{y}_t - \tilde{\mathbf{x}}_+^*\end{pmatrix} = \begin{pmatrix}\mathbf{O} & \mathbf{I} - \eta\mathbf{a}_t\mathbf{a}_t^\top \\ -c\mathbf{I} & (1+c)\mathbf{I} - q\mathbf{a}_t\mathbf{a}_t^\top\end{pmatrix}\begin{pmatrix}\mathbf{x}_{t-1} - \tilde{\mathbf{x}}_+^* \\ \mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*\end{pmatrix} + \begin{pmatrix}\eta\varepsilon_t\mathbf{a}_t \\ q\varepsilon_t\mathbf{a}_t\end{pmatrix}. \tag{8}$$

*where*

$$c = \frac{1-\theta}{1+\theta}, \quad q = \frac{\eta + \theta\gamma}{1+\theta}, \quad \varepsilon_t = -\ell'(\mathbf{a}_t^\top\tilde{\mathbf{y}}, b_t) + \mathbf{a}_t^\top(\tilde{\mathbf{y}} - \tilde{\mathbf{x}}_+^*). \tag{9}$$

*Proof.* Recall that the update rule is

$$\begin{aligned}\mathbf{x}_t &= \mathbf{y}_{t-1} - \eta\mathbf{g}_t, \\ \mathbf{z}_t &= \theta\mathbf{y}_{t-1} + (1-\theta)\mathbf{z}_{t-1} - \gamma\mathbf{g}_t, \\ \mathbf{y}_t &= \frac{1}{1+\theta}\mathbf{x}_t + \frac{\theta}{1+\theta}\mathbf{z}_t,\end{aligned}$$

where the gradient $\mathbf{g}_t$ is

$$\mathbf{g}_t = \ell'(\mathbf{a}_t^\top\tilde{\mathbf{y}}, b_t)\mathbf{a}_t + \mathbf{a}_t\mathbf{a}_t^\top(\mathbf{y}_{t-1} - \tilde{\mathbf{y}}),$$

By algebraic transformations to eliminate $\mathbf{z}_t$ and $\mathbf{z}_{t-1}$, we obtain the following matrix form:

$$\begin{pmatrix}\mathbf{x}_t \\ \mathbf{y}_t\end{pmatrix} = \begin{pmatrix}\mathbf{O} & \mathbf{I} \\ -c\mathbf{I} & (1+c)\mathbf{I}\end{pmatrix}\begin{pmatrix}\mathbf{x}_{t-1} \\ \mathbf{y}_{t-1}\end{pmatrix} - \begin{pmatrix}\eta\mathbf{g}_t \\ q\mathbf{g}_t\end{pmatrix}, \tag{10}$$

where $c$ and $q$ are define in (9). Note that by the definition of $\varepsilon_t$ in (9), we have

$$\mathbf{g}_t = \mathbf{a}_t\mathbf{a}_t^\top(\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*) - \varepsilon_t\mathbf{a}_t. \tag{11}$$

We complete the proof by substituting (11) into (10). □

Define

$$\boldsymbol{\eta}_t = \begin{pmatrix}\mathbf{x}_t - \tilde{\mathbf{x}}_+^* \\ \mathbf{y}_t - \tilde{\mathbf{x}}_+^*\end{pmatrix}, \quad \widehat{\mathbf{A}}_t = \begin{pmatrix}\mathbf{O} & \mathbf{I} - \eta\mathbf{a}_t\mathbf{a}_t^\top \\ -c\mathbf{I} & (1+c)\mathbf{I} - q\mathbf{a}_t\mathbf{a}_t^\top\end{pmatrix}, \quad \boldsymbol{\zeta}_t = \begin{pmatrix}\eta\varepsilon_t\mathbf{a}_t \\ q\varepsilon_t\mathbf{a}_t\end{pmatrix}.$$

Then the dynamic of $\boldsymbol{\eta}_t$ have the following form,

$$\boldsymbol{\eta}_t = \widehat{\mathbf{A}}_t\boldsymbol{\eta}_{t-1} + \boldsymbol{\zeta}_t, \quad \boldsymbol{\eta}_0 = \begin{pmatrix}\tilde{\mathbf{y}} - \tilde{\mathbf{x}}_+^* \\ \tilde{\mathbf{y}} - \tilde{\mathbf{x}}_+^*\end{pmatrix}.$$

Following the bias-variance decomposition technique (Dieuleveut & Bach, 2016; Jain et al., 2018a; Zou et al., 2023), we decompose $\boldsymbol{\eta}_t$ into the bias dynamic $\boldsymbol{\eta}_t^{\text{bias}}$ and variance dynamic $\boldsymbol{\eta}_t^{\text{var}}$ as follows:

$$\begin{aligned}\boldsymbol{\eta}_t^{\text{bias}} &= \widehat{\mathbf{A}}_t\boldsymbol{\eta}_{t-1}^{\text{bias}}, & \boldsymbol{\eta}_0^{\text{bias}} &= \begin{pmatrix}\tilde{\mathbf{y}} - \tilde{\mathbf{x}}_+^* \\ \tilde{\mathbf{y}} - \tilde{\mathbf{x}}_+^*\end{pmatrix}, \\ \boldsymbol{\eta}_t^{\text{var}} &= \widehat{\mathbf{A}}_t\boldsymbol{\eta}_{t-1}^{\text{var}} + \boldsymbol{\zeta}_t, & \boldsymbol{\eta}_0^{\text{var}} &= \mathbf{0}.\end{aligned} \tag{12}$$

Note that $\boldsymbol{\eta}_t = \boldsymbol{\eta}_t^{\text{bias}} + \boldsymbol{\eta}_t^{\text{var}}$ since the dynamics are linear. With the bias-variance decomposition given above, we construct $\mathbf{r}$ and $\mathbf{v}$ as follows.

**Definition D.1.** Let

$$\mathbf{r} = \frac{2}{h_k T}\left(\sum_{t=T/2+1}^{T} \boldsymbol{\eta}_t^{\text{bias}}\right)_1, \quad \mathbf{v} = \frac{2}{h_k T}\left(\sum_{t=T/2+1}^{T} \boldsymbol{\eta}_t^{\text{var}}\right)_1,$$

where for $\boldsymbol{\eta} \in \mathbb{R}^{2d}$, $\boldsymbol{\eta}_1 \in \mathbb{R}^d$ denotes the first $d$ elements of $\boldsymbol{\eta}$.

### D.1.2 Covariance Dynamics

Define

$$\mathbf{R} = \mathbb{E}_{\mathbf{a},b\sim\mathcal{D}}\left(\varepsilon^2 \mathbf{a}\mathbf{a}^\top\right).$$

where $\varepsilon$ is defined in (9). We introduce the following linear operators on $\mathbb{R}^{d\times d}$,

$$\mathcal{B} = \mathbb{E}(\widehat{\mathbf{A}}_t \otimes \widehat{\mathbf{A}}_t), \quad \tilde{\mathcal{B}} = \mathbf{A} \otimes \mathbf{A},$$

where $\mathbf{A} \otimes \mathbf{B}$ defines a linear matrix operator for matrix $\mathbf{A}$ and $\mathbf{B}$ that satisfies $(\mathbf{A} \otimes \mathbf{B}) \circ \mathbf{C} = \mathbf{A}\mathbf{C}\mathbf{B}^\top$, and $\circ$ denotes the operation of a linear matrix operator on a matrix, and. Let

$$\mathcal{M} = \mathbb{E}(\mathbf{a} \otimes \mathbf{a} \otimes \mathbf{a} \otimes \mathbf{a}), \quad \tilde{\mathcal{M}} = \boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma}. \tag{13}$$

Then for $\mathbf{M} = \begin{pmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{pmatrix} \in \mathbb{R}^{2d\times 2d}$,

$$(\mathcal{B} - \tilde{\mathcal{B}}) \circ \mathbf{M} = \begin{pmatrix} \eta^2(\mathcal{M} - \tilde{\mathcal{M}}) \circ \mathbf{M}_{22} & \eta q(\mathcal{M} - \tilde{\mathcal{M}}) \circ \mathbf{M}_{22} \\ \eta q(\mathcal{M} - \tilde{\mathcal{M}}) \circ \mathbf{M}_{22} & q^2(\mathcal{M} - \tilde{\mathcal{M}}) \circ \mathbf{M}_{22} \end{pmatrix}.$$

Let

$$\mathbf{C}_t = \mathbb{E}\left(\boldsymbol{\eta}_t^{\text{var}}(\boldsymbol{\eta}_t^{\text{var}})^\top\right),$$

Then we can write the update of $\mathbf{C}_t$ as follows

$$\mathbf{C}_t = \mathcal{B} \circ \mathbf{C}_{t-1} + \begin{pmatrix} \eta^2\mathbf{R} & \eta q\mathbf{R} \\ \eta q\mathbf{R} & q^2\mathbf{R} \end{pmatrix}, \quad \mathbf{C}_0 = \mathbf{O}.$$

### D.1.3 Properties of Momentum Matrix

We note that $\mathbf{A}$ is a block-diagonal matrix, and the $i$-th block[3] is

$$\mathbf{A}_i = \begin{pmatrix} 0 & 1-\eta\lambda_i \\ c & 1+c-q\lambda_i \end{pmatrix},$$

where $\lambda_i$ is the $i$-th eigenvalue of $\boldsymbol{\Sigma}$. We define

$$\mathbf{A}(\lambda) = \begin{pmatrix} 0 & 1-\eta\lambda \\ c & 1+c-q\lambda \end{pmatrix}.$$

**Lemma D.3.** *We have*

$$(\mathbf{I} - \mathbf{A}(\lambda))^{-1}\begin{pmatrix} \eta \\ q \end{pmatrix} = \frac{1}{\lambda}\begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

*Proof.* The inverse of $\mathbf{I} - \mathbf{A}(\lambda)$ is given by

$$(\mathbf{I} - \mathbf{A}(\lambda))^{-1} = \frac{1}{(q-c\eta)\lambda}\begin{pmatrix} q\lambda - c & 1-\eta\lambda \\ -c & 1 \end{pmatrix}.$$

Therefore,

$$(\mathbf{I} - \mathbf{A}(\lambda))^{-1}\begin{pmatrix} \eta \\ q \end{pmatrix} = \frac{1}{\lambda(q-c\eta)}\begin{pmatrix} \eta(q\lambda - c) + q(1-\eta\lambda) \\ c\eta + q \end{pmatrix} = \frac{1}{\lambda}\begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

This completes the proof. □

[3] To be more precise, $\mathbf{A}_i$ is the $2\times 2$ submatrix of $\mathbf{A}$ formed by the entries $(i,i)$, $(d+i,i)$, $(i,d+i)$ and $(d+i,d+i)$ of $\mathbf{A}$.

**Spectral properties of momentum matrix.** The eigenvalues of $\mathbf{A}(\lambda)$ are

$$x_1(\lambda) = \frac{1+c-q\lambda}{2} - \frac{\sqrt{(1+c-q\lambda)^2 - 4c(1-\eta\lambda)}}{2},$$
$$x_1(\lambda) = \frac{1+c-q\lambda}{2} + \frac{\sqrt{(1+c-q\lambda)^2 - 4c(1-\eta\lambda)}}{2}.$$

Let $(1+c-q\lambda)^2 - 4c(1-\eta\lambda) < 0$, we obtain the region of $\lambda$ such that $x_1(\lambda), x_2(\lambda) \in \mathbb{C}$:

$$\underbrace{\frac{(1-c)^2}{\left(\sqrt{q-c\eta} + \sqrt{c(q-\eta)}\right)^2}}_{\lambda_L} < \lambda < \underbrace{\frac{(1-c)^2}{\left(\sqrt{q-c\eta} - \sqrt{c(q-\eta)}\right)^2}}_{\lambda_U}.$$

We adopt Lemma E.2 from Li et al. (2024), which bounds $x_1(\lambda)$ and $x_2(\lambda)$.

**Lemma D.4.** *Let $\lambda \geqslant 0$ and $\eta\lambda < 1$.*

- *If $\lambda \leqslant \lambda_L$, then $x_1(\lambda)$ and $x_2(\lambda)$ are real, and $x_1(\lambda) \leqslant x_2(\lambda) \leqslant 1 - (q - c\delta)\lambda/(1-c)$.*
- *If $\lambda_L < \lambda < \lambda_U$, then $x_1(\lambda)$ and $x_2(\lambda)$ are complex, and $\|x_1(\lambda)\| = \|x_2(\lambda)\| = \sqrt{c\,(1-\delta\lambda)}$.*
- *If $\lambda \geqslant \lambda_U$, then $x_1(\lambda)$ and $x_2(\lambda)$ are real, and $x_1(\lambda) \leqslant x_2(\lambda) \leqslant c\delta/q$.*

**Bound of power of momentum matrix.** The $k$-th power of $\mathbf{A}(\lambda)$ is

$$\mathbf{A}^k(\lambda) = \begin{pmatrix} -c(1-\eta\lambda)a_{k-1}(\lambda) & (1-\eta\lambda)a_k(\lambda) \\ -ca_k(\lambda) & a_{k+1}(\lambda) \end{pmatrix}, \quad \text{where} \quad a_k(\lambda) = \frac{x_2^k(\lambda) - x_1^k(\lambda)}{x_2(\lambda) - x_1(\lambda)},$$

and $a_k(\lambda) \in \mathbb{R}$.

The following lemma is modified from Lemma 20 in (Liu et al., 2025), which bounds $\mathbf{A}^k(\lambda)$. We present the proof for completeness.

**Lemma D.5.** *Suppose $\eta\lambda < 1$ and $\eta \leqslant q$, then we have*

$$\left\| \left( \mathbf{A}^k(\lambda) \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right)_1 \right\| \leqslant 2.$$

*Proof.* For notational simplicity, we write $a_k = a_k(\lambda)$ and $x_{1,2} = x_{1,2}(\lambda)$. We have

$$\left( \mathbf{A}^k(\lambda) \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right)_1 = (1-\eta\lambda)(a_k - ca_{k-1}).$$

The result holds for $k = 0$, so we need to show that $\|a_k - ca_{k-1}\| \leqslant 2$ for $k \geqslant 1$.

1. If $\lambda \leqslant \lambda_L$, by Lemma D.4, and $\delta \leqslant q$, we have $a_k \geqslant 0$, and

$$x_1 \leqslant x_2 \leqslant 1 - \frac{q - c\delta}{1-c}\lambda \leqslant 1 - \delta\lambda.$$

By Vieta's formula, $x_1 x_2 = c(1-\delta\lambda)$, so $c \leqslant x_1 \leqslant x_2$. Thus, we bound $a_k - ca_{k-1}$ as follows:

$$a_k - ca_{k-1} \geqslant a_k - x_1 a_{k-1} = x_2^{k-1} > 0,$$
$$a_k - ca_{k-1} \leqslant a_k - x_1 x_2 a_{k-1} = x_2^{k-1} + (1-x_2)\sum_{i=1}^{k-1} x_1^i x_2^{k-i-1}$$
$$\leqslant x_2^{k-1}\left[1 + (k-1)(1-x_2)\right] \overset{a}{\leqslant} 1,$$

where $\overset{a}{\leqslant}$ uses Lemma D.32 and $k \geqslant 1$.

2. If $\lambda_L < \lambda < \lambda_U$, let $x_{1,2} = r(\cos\theta \pm \mathrm{i}\sin\theta)$, where $0 \leqslant \theta \leqslant \pi/2$. then we have $r = \sqrt{c(1-\delta\lambda)} \leqslant 1$ and $2r\cos\theta = x_1 + x_2 = 1 + c - q\lambda \geqslant 0$. Therefore,

$$\begin{aligned} a_k - ca_{k-1} &= \frac{r^{k-1}\sin(k\theta)}{\sin\theta} - \frac{r^{k-2}\sin((k-1)\theta)}{\sin\theta} \\ &= r^{k-2}\left(r\cos((k-1)\theta) + \frac{r-c}{\sin\theta}\sin((k-1)\theta) - r\tan\frac{\theta}{2}\sin((k-1)\theta)\right), \end{aligned}$$

Apply triangular inequality, and $\|\sin k\theta\| \leqslant 1$, $\|\cos k\theta\| \leqslant 1$, $\|\tan(\theta/2)\| \leqslant 1$

$$\|a_k - ca_{k-1}\| \overset{a}{\leqslant} r^{k-2}\left(r + (k-1)(1-r)\right) + r^{k-1} = r^{k-2}\left(1 + (k-2)(1-r)\right) + r^{k-1} \overset{b}{\leqslant} 2,$$

where $\overset{a}{\leqslant}$ holds since $r^2 \leqslant c \leqslant 1$ implies $\|r - c\| \leqslant 1 - r$ and $\|\sin(k\theta)/\sin\theta\| \leqslant k$, $\overset{b}{\leqslant}$ follows from Lemma D.32, $k \geqslant 1$ and $0 \leqslant r \leqslant 1$.

3. If $\lambda > \lambda_U$, we have $a_k \geqslant 0$, and $x_1 \leqslant x_2 \leqslant c\delta/q \leqslant c$ by Lemma D.4, and $\delta \leqslant q$. Then we have the following bounds:

$$\begin{aligned} a_k - ca_{k-1} \geqslant & a_k - a_{k-1} = \sum_{i=0}^{k-1} x_1^i x_2^{k-i-1} - \sum_{i=0}^{k-2} x_1^i x_2^{k-i-2} \\ = & x_1^{k-1} - (1-x_2)\sum_{i=0}^{k-2} x_1^i x_2^{k-i-2} \\ \geqslant & x_1^{k-1} - (k-1)(1-x_2)x_2^{k-2} \\ \geqslant & -x_2^{k-2}\left(1 + (k-2)x_2^{k-1}\right) \overset{a}{\geqslant} -1, \\ a_k - ca_{k-1} \leqslant & a_k - x_2 a_{k-1} = x_1^{k-1} \leqslant 1, \end{aligned}$$

where $\overset{a}{\geqslant}$ if from Lemma D.32 and $k \geqslant 1$.

We complete the proof by combining the above cases. □

### D.2 Variance Upper Bound

**Organization.** In the following, we introduce $L_{\mathrm{var}}(\{\mathbf{X}_t\}_{t\in\mathbb{N}})$, which acts like a loss function for covariance matrices $\{\mathbf{X}_t\}_{t\in\mathbb{N}}$. In Appendix D.2.1, we state the layer-peeled decomposition technique. We introduce dynamics $\tilde{\mathbf{C}}_t^{(\ell)}$ and establish the relation $\mathbf{C}_t = \tilde{\mathbf{C}}_t + \sum_{\ell=1}^{\infty}\tilde{\mathbf{C}}_t^{(\ell)}$ in Lemma D.10. In Appendix D.2.3, we decompose the noise matrix $\mathbf{R}$ defined in (25), which is the key step to establish the statistical term that depends on the noise covariance matrix at the minimizer $\mathbf{x}^*$ (i.e., $\mathbf{Q}$). In Appendix D.2.2, we construct the core auxiliary iteration $\tilde{\mathbf{\Pi}}_t(\mathbf{M})$ in (21), where the covariance matrix is a general PSD matrix $\mathbf{M}$. In Appendix D.2.4, we set $\mathbf{M} = \mathbf{R}$ to obtain the bounds of $\tilde{\mathbf{C}}_t$. In Appendix D.2.5, we bound $\tilde{\mathbf{C}}_t^{(\ell)}$ by recursively obtaining bounds of $\tilde{\mathbf{C}}_t^{(\ell)}$. Finally, by combining the bounds in Appendix D.2.4 and Appendix D.2.5, we obtain the variance bound by proving Lemma D.6 in Appendix D.2.6.

The goal of this section is to prove the following lemma.

**Lemma D.6** (Variance Upper Bound)**.** *Suppose Assumptions 1, 2, 3, 4 and 5 hold. Let* $\mathbf{v}$ *defined in Definition D.1 Then* $\mathbb{E}\mathbf{v} = \mathbf{0}$ *and*

$$\mathbb{E}\|\mathbf{v}\|_{\mathbf{\Sigma}}^2 \leqslant \frac{320\left(3\operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{Q}) + 8\eta\tilde{\kappa}L_\ell\operatorname{tr}\mathbf{Q}\right)}{T} + \frac{160L_\ell(6L + \tilde{\kappa}(7 + 16\eta B))(F(\tilde{\mathbf{y}}) - F(\mathbf{x}^*))}{T},$$

*where the expectation is taken with respect to the samples drawn in the* $k$*-th outer iteration.*

The proof is deferred to Appendix D.2.6.

To simplify notation, we introduce the following mapping $L_{\mathrm{var}}$ on a sequence of PSD matrices.

**Definition D.2.** For a PSD matrices sequence $\{\mathbf{X}_t \in \mathbb{R}^{2d\times 2d}\}_{t\in\mathbb{N}}$, we define

$$L_{\mathrm{var}}(\{\mathbf{X}_t\}_{t\in\mathbb{N}}) \stackrel{\mathrm{def}}{=} \frac{4}{T^2}\left\langle \begin{pmatrix}\mathbf{\Sigma} & \mathbf{O}\\ \mathbf{O} & \mathbf{O}\end{pmatrix}, \sum_{s,t=T/2+1}^{T} \mathbf{A}^{\max\{s-t,0\}}\mathbf{X}_{\min\{s,t\}}\left(\mathbf{A}^{\max\{t-s,0\}}\right)^\top \right\rangle. \quad (14)$$

The mapping $L_{\mathrm{var}}$ is linear and monotonic in the following sense.

**Lemma D.7.** *For PSD matrices sequence $\{\mathbf{X}_t\}_{t\in\mathbb{N}}$ and $\{\mathbf{T}_t\}_{t\in\mathbb{N}}$, we have*

$$L_{\mathrm{var}}(\{\alpha\mathbf{X}_t + \beta\mathbf{Y}_t\}_{t\in\mathbb{N}}) = \alpha L_{\mathrm{var}}(\{\mathbf{X}_t\}_{t\in\mathbb{N}}) + \beta L_{\mathrm{var}}(\{\mathbf{Y}_t\}_{t\in\mathbb{N}}). \quad (15)$$

**Lemma D.8.** *For PSD matrices sequence $\{\mathbf{X}_t\}_{t\in\mathbb{N}}$ and $\{\mathbf{Y}_t\}_{t\in\mathbb{N}}$ such that $\mathbf{X}_t \preceq \mathbf{Y}_t$ for all $t\in\mathbb{N}$, then we have*

$$L_{\mathrm{var}}(\{\mathbf{X}_t\}_{t\in\mathbb{N}}) \leqslant L_{\mathrm{var}}(\{\mathbf{Y}_t\}_{t\in\mathbb{N}}). \quad (16)$$

The proof follows directly from the definition of $L_{\mathrm{var}}$. We can represent $\mathbb{E}\|\mathbf{v}\|_{\mathbf{\Sigma}}^2$ by $L_{\mathrm{var}}$ as follows.

**Lemma D.9.** *We have*

$$\mathbb{E}\|\mathbf{v}\|_{\mathbf{\Sigma}}^2 = L_{\mathrm{var}}(\{\mathbf{C}_t\}_{t\in\mathbb{N}}).$$

*Proof.* By the definition of $\mathbf{v}$, we have

$$\begin{aligned}
\mathbb{E}\|\mathbf{v}\|_{\mathbf{\Sigma}}^2 &= \left\langle \begin{pmatrix}\mathbf{\Sigma} & \mathbf{O}\\ \mathbf{O} & \mathbf{O}\end{pmatrix}, \mathbb{E}\left[\left(\frac{2}{T}\sum_{t=T/2+1}^{T}\boldsymbol{\eta}_t^{\mathrm{var}}\right)\left(\frac{2}{T}\sum_{t=T/2+1}^{T}\boldsymbol{\eta}_t^{\mathrm{var}}\right)^\top\right]\right\rangle\\
&= \left\langle \begin{pmatrix}\mathbf{\Sigma} & \mathbf{O}\\ \mathbf{O} & \mathbf{O}\end{pmatrix}, \frac{4}{T^2}\sum_{s,t=T/2+1}^{T}\mathbb{E}(\boldsymbol{\eta}_s^{\mathrm{var}}\boldsymbol{\eta}_t^{\mathrm{var}})\right\rangle,\\
&\stackrel{a}{=} \left\langle \begin{pmatrix}\mathbf{\Sigma} & \mathbf{O}\\ \mathbf{O} & \mathbf{O}\end{pmatrix}, \frac{4}{T^2}\sum_{s,t=T/2+1}^{T}\mathbf{A}^{\max\{s-t,0\}}\mathbf{C}_{\min\{s,t\}}\left(\mathbf{A}^{\max\{t-s,0\}}\right)^\top\right\rangle,
\end{aligned}$$

where $\stackrel{a}{=}$ uses

$$\mathbb{E}\boldsymbol{\eta}_s^{\mathrm{var}}(\boldsymbol{\eta}_t^{\mathrm{var}})^\top = \mathbb{E}\left(\mathbb{E}\left(\boldsymbol{\eta}_s^{\mathrm{var}}|\mathcal{F}_t\right)(\boldsymbol{\eta}_t^{\mathrm{var}})^\top\right) = \mathbf{A}^{s-t}\mathbb{E}\boldsymbol{\eta}_t^{\mathrm{var}}(\boldsymbol{\eta}_t^{\mathrm{var}})^\top = \mathbf{A}^{s-t}\mathbf{C}_t, \quad s\geqslant t$$

and collecting the terms. □

The quantity $L_{\mathrm{var}}(\{\mathbf{X}_t\}_{t\in\mathbb{N}})$ can be interpreted as the loss incurred by the tail-averaged sequence with covariance $\mathbf{X}_t$. Specifically, we construct the dynamics $\boldsymbol{\psi}_t \in \mathbb{R}^{2d\times 2d}$ as follows:

$$\boldsymbol{\psi}_t = \mathbf{A}\boldsymbol{\psi}_{t-1} + \begin{pmatrix}\eta\boldsymbol{\xi}_t\\ q\boldsymbol{\xi}_t\end{pmatrix}, \quad \boldsymbol{\psi}_0 = \mathbf{0},$$

where $\boldsymbol{\xi}_t \sim$ i.i.d. $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\boldsymbol{\xi}})$ and $\mathbb{E}\boldsymbol{\psi}_t\boldsymbol{\psi}_t^\top = \mathbf{X}_t$, then we have (the proof is identical to the proof of Lemma D.9)

$$L_{\mathrm{var}}(\{\mathbf{X}_t\}_{t\in\mathbb{N}}) = \left\langle \begin{pmatrix}\mathbf{\Sigma} & \mathbf{O}\\ \mathbf{O} & \mathbf{O}\end{pmatrix}, \mathbb{E}\left[\left(\frac{2}{T}\sum_{t=T/2+1}^{T}\boldsymbol{\psi}_t\right)\left(\frac{2}{T}\sum_{t=T/2+1}^{T}\boldsymbol{\psi}_t\right)^\top\right]\right\rangle.$$

However, $L_{\mathrm{var}}(\{\mathbf{X}_t\}_{t\in\mathbb{N}})$ can be calculated for any sequence $\{\mathbf{X}_t\}_{t\in\mathbb{N}}$ and does not necessarily require $\mathbf{X}_t$ to be the covariance of some dynamics $\boldsymbol{\psi}_t$.

### D.2.1 Layer-peeled Decomposition

The goal of this section is to bound $\mathbb{E}\|\mathbf{v}\|_{\mathbf{\Sigma}}^2 = L_{\mathrm{var}}(\{\mathbf{C}_t\}_{t\in\mathbb{N}})$. To tightly characterize $\mathbf{C}_t$, this paper proposes the **layer-peeled decomposition**, which decomposes the covariance matrix $\mathbf{C}_t$ into the sum of the following dynamics.

**Definition D.3** (Layer 0 Dynamics)**.** Let

$$\tilde{\mathbf{C}}_t = \mathbf{A}\tilde{\mathbf{C}}_{t-1}\mathbf{A}^\top + \begin{pmatrix} \eta^2\mathbf{R} & \eta q\mathbf{R} \\ \eta q\mathbf{R} & q^2\mathbf{R} \end{pmatrix}, \quad \tilde{\mathbf{C}}_0 = \mathbf{O}, \tag{17}$$

**Definition D.4** (Layer $\ell$ Dynamics)**.** Let

$$\tilde{\mathbf{C}}_t^{(\ell)} = \mathbf{A}\tilde{\mathbf{C}}_{t-1}^{(\ell)}\mathbf{A}^\top + (\mathcal{B} - \tilde{\mathcal{B}}) \circ \tilde{\mathbf{C}}_{t-1}^{(\ell-1)}, \quad \tilde{\mathbf{C}}_0^{(\ell)} = \mathbf{O}. \tag{18}$$

We call (17) the layer 0 dynamics and call (18) layer $\ell$ dynamics. The following lemma shows that $\mathbf{C}_t = \tilde{\mathbf{C}}_t + \sum_{\ell=1}^{\infty} \tilde{\mathbf{C}}_t^{(\ell)}$.

**Lemma D.10.** *We have*

$$\mathbf{C}_t = \tilde{\mathbf{C}}_t + \sum_{\ell=1}^{\infty} \tilde{\mathbf{C}}_t^{(\ell)}.$$

*We also have $\tilde{\mathbf{C}}_t^{(\ell)} = \mathbf{O}$ for $\ell \geqslant t$, so the above display is a finite sum.*

*Proof.* We construct a series of dynamics (indexed by $\ell$) as follows:

$$\mathbf{C}_t^{(\ell)} = \mathcal{B} \circ \mathbf{C}_{t-1}^{(\ell)} + (\mathcal{B} - \tilde{\mathcal{B}}) \circ \tilde{\mathbf{C}}_{t-1}^{(\ell-1)}, \quad \mathbf{C}_0^{(\ell)} = \mathbf{O}. \tag{19}$$

Recall that $\tilde{\mathcal{B}} = \mathbf{A} \otimes \mathbf{A}$, so $\mathbf{C}_t = \tilde{\mathbf{C}}_t + \mathbf{C}_t^{(1)}$, $\mathbf{C}_t^{(\ell)} = \tilde{\mathbf{C}}_t^{(\ell)} + \mathbf{C}_t^{(\ell+1)}$, and

$$\mathbf{C}_t^{(\ell)} = \mathbf{O}, \quad \tilde{\mathbf{C}}_t^{(\ell)} = \mathbf{O}, \quad \text{for all } 0 \leqslant t \leqslant \ell,$$

so we have the following decomposition:

$$\begin{aligned} \mathbf{C}_t^{(1)} &= \tilde{\mathbf{C}}_t^{(1)} + \mathbf{C}_t^{(2)} \\ &= \tilde{\mathbf{C}}_t^{(1)} + \tilde{\mathbf{C}}_t^{(2)} + \mathbf{C}_t^{(3)} \\ &\cdots \\ &= \sum_{\ell=1}^{\infty} \tilde{\mathbf{C}}_t^{(\ell)}. \qquad \text{(note that } \tilde{\mathbf{C}}_t^{(\ell)} = \mathbf{O} \text{ for all } \ell \geqslant t) \end{aligned} \tag{20}$$

This completes the proof. □

### D.2.2 Core Auxiliary Dynamics

In this section, we analyze the following dynamics $\tilde{\mathbf{\Pi}}_t(\mathbf{M}) \in \mathbb{R}^{2d \times 2d}$,

$$\tilde{\mathbf{\Pi}}_t(\mathbf{M}) = \mathbf{A}\tilde{\mathbf{\Pi}}_{t-1}(\mathbf{M})\mathbf{A}^\top + \begin{pmatrix} \eta^2\mathbf{M} & \eta q\mathbf{M} \\ \eta q\mathbf{M} & q^2\mathbf{M} \end{pmatrix}, \quad \tilde{\mathbf{\Pi}}_0 = \mathbf{O}, \tag{21}$$

where $\mathbf{M}$ is any PSD matrix. We also write $\tilde{\mathbf{\Pi}}_t$ for notational simplicity. We focus on the update of the diagonal[4] of $\tilde{\mathbf{\Pi}}_t$, which is

$$\tilde{\mathbf{\Pi}}_{t,i} = \mathbf{A}_i\tilde{\mathbf{\Pi}}_{t-1,i}\mathbf{A}_i^\top + m_i \begin{pmatrix} \eta^2 & \eta q \\ \eta q & q^2 \end{pmatrix},$$

where $m_i$ denotes the $i$-th diagonal element of $\mathbf{M}$.

**Upper Bound of $L_{\text{var}}(\{\tilde{\mathbf{\Pi}}_t(\mathbf{M})\}_{t\in\mathbb{N}})$.** We bound $L_{\text{var}}(\{\tilde{\mathbf{\Pi}}_t(\mathbf{M})\}_{t\in\mathbb{N}})$ from above. The first step is to derive a matrix form.

[4]To be more precise, $\tilde{\mathbf{\Pi}}_{t,i}$ is the $2 \times 2$ submatrix of $\tilde{\mathbf{\Pi}}_{t,i}$ formed by the entries $(i,i)$, $(d+i,i)$, $(i,d+i)$ and $(d+i,d+i)$ of $\tilde{\mathbf{\Pi}}_t$.

**Lemma D.11.** *We have* $L_{\mathrm{var}}(\{\tilde{\mathbf{\Pi}}_t(\mathbf{M})\}_{t\in\mathbb{N}}) = 4(A_{T/2} + B_{T/2})/T^2$*, where*

$$A_m = \left\langle \begin{pmatrix} \mathbf{\Sigma} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix}, \left(\sum_{k=0}^{T-m-1} \mathbf{A}^k\right) \tilde{\mathbf{\Pi}}_{m+1} \left(\sum_{k=0}^{T-m-1} \mathbf{A}^k\right)^\top \right\rangle,$$
$$B_m = \sum_{s=1}^{T-m-1} \left\langle \begin{pmatrix} \mathbf{\Sigma} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix}, \left(\sum_{k=0}^{s-1} \mathbf{A}^k\right) \begin{pmatrix} \eta^2\mathbf{M} & \eta q\mathbf{M} \\ \eta q\mathbf{M} & q^2\mathbf{M} \end{pmatrix} \left(\sum_{k=0}^{s-1} \mathbf{A}^k\right)^\top \right\rangle.$$

*Proof.* We show by induction that

$$L_m \overset{\text{def}}{=} \left\langle \begin{pmatrix} \mathbf{\Sigma} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix}, \sum_{s,t=m+1}^{T} \mathbf{A}^{\max\{s-t,0\}} \tilde{\mathbf{\Pi}}_{\min\{s,t\}} \left(\mathbf{A}^{\max\{t-s,0\}}\right)^\top \right\rangle = A_m + B_m.$$

The above display holds for $m = T-1$. Assuming the above display holds for $m \leqslant T-1$, we show that it holds for $m-1$. First, note that

$$\begin{aligned} L_{m-1} - L_m = &\left\langle \begin{pmatrix} \mathbf{\Sigma} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix}, \tilde{\mathbf{\Pi}}_m \right\rangle + \left\langle \begin{pmatrix} \mathbf{\Sigma} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix}, \left(\sum_{k=1}^{T-m} \mathbf{A}^k\right) \tilde{\mathbf{\Pi}}_m \right\rangle \\ &+ \left\langle \begin{pmatrix} \mathbf{\Sigma} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix}, \tilde{\mathbf{\Pi}}_m \left(\sum_{k=1}^{T-m} \mathbf{A}^k\right)^\top \right\rangle. \end{aligned} \tag{22}$$

Therefore, we have

$$\begin{aligned} &A_{m-1} - A_m \\ &= \left\langle \begin{pmatrix} \mathbf{\Sigma} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix}, \left(\mathbf{A}\sum_{k=0}^{T-m-1} \mathbf{A}^k + \mathbf{I}\right) \tilde{\mathbf{\Pi}}_m \left(\mathbf{A}\sum_{k=0}^{T-m-1} \mathbf{A}^k + \mathbf{I}\right)^\top \right\rangle \\ &\quad - \left\langle \begin{pmatrix} \mathbf{\Sigma} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix}, \left(\sum_{k=0}^{T-m-1} \mathbf{A}^k\right) \tilde{\mathbf{\Pi}}_{m+1} \left(\sum_{k=0}^{T-m-1} \mathbf{A}^k\right)^\top \right\rangle \\ &\overset{a}{=} (L_{m-1} - L_m) + \left\langle \begin{pmatrix} \mathbf{\Sigma} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix}, \left(\sum_{k=0}^{T-m-1} \mathbf{A}^k\right) \left(\tilde{\mathbf{\Pi}}_{m+1} - \mathbf{A}\tilde{\mathbf{\Pi}}_m\mathbf{A}^\top\right) \left(\sum_{k=0}^{T-m-1} \mathbf{A}^k\right)^\top \right\rangle \\ &\overset{b}{=} (L_{m-1} - L_m) - \left\langle \begin{pmatrix} \mathbf{\Sigma} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix}, \left(\sum_{k=0}^{T-m-1} \mathbf{A}^k\right) \begin{pmatrix} \eta^2\mathbf{M} & \eta q\mathbf{M} \\ \eta q\mathbf{M} & q^2\mathbf{M} \end{pmatrix} \left(\sum_{k=0}^{T-m-1} \mathbf{A}^k\right)^\top \right\rangle \\ &= (L_{m-1} - L_m) - (B_{m-1} - B_m), \end{aligned}$$

where we expand the product and apply (22) in $\overset{a}{=}$, and $\overset{b}{=}$ applies the update rule of $\tilde{\mathbf{\Pi}}_t$ in (21). Thus, $A_{m-1} + B_{m-1} = A_m + B_m + (L_{m-1} - L_m) = L_{m-1}$. □

We further derive a dimension-wise summation form of $L_{\mathrm{var}}(\{\tilde{\mathbf{\Pi}}_t(\mathbf{M})\}_{t\in\mathbb{N}})$.

**Lemma D.12.** *Let* $m_i$ *denote the* $i$*-th element of the diagonal of* $\mathbf{M}$*, then*

$$\begin{aligned} L_{\mathrm{var}}(\{\tilde{\mathbf{\Pi}}_t(\mathbf{M})\}_{t\in\mathbb{N}}) = &\frac{4}{T^2} \sum_{i=1}^{d} \sum_{k=0}^{T/2} \lambda_i m_i \left(\mathbf{A}_i^k \left(\mathbf{I} - \mathbf{A}_i^{T/2}\right) (\mathbf{I} - \mathbf{A}_i)^{-1} \begin{pmatrix} \eta \\ q \end{pmatrix}\right)_1^2 \\ &+ \frac{4}{T^2} \sum_{i=1}^{d} \sum_{k=1}^{T/2} \lambda_i m_i \left(\left(\mathbf{I} - \mathbf{A}_i^{T/2-k}\right) (\mathbf{I} - \mathbf{A}_i)^{-1} \begin{pmatrix} \eta \\ q \end{pmatrix}\right)_1^2. \end{aligned}$$

*Proof.* By Lemma D.11, we have $L_{\mathrm{var}}(\{\tilde{\mathbf{\Pi}}_t(\mathbf{M})\}_{t\in\mathbb{N}}) = 4(A_{T/2} + B_{T/2})/T^2$, where

$$A_{T/2} = \left\langle \begin{pmatrix} \mathbf{\Sigma} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix}, \left(\sum_{k=0}^{T/2-1} \mathbf{A}^k\right) \tilde{\mathbf{\Pi}}_{T/2+1} \left(\sum_{k=0}^{T/2-1} \mathbf{A}^k\right)^\top \right\rangle,$$

$$B_{T/2} = \sum_{k=1}^{T/2} \left\langle \begin{pmatrix} \mathbf{\Sigma} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix}, \left(\sum_{k=0}^{T/2-k-1} \mathbf{A}^k\right) \begin{pmatrix} \eta^2\mathbf{M} & \eta q\mathbf{M} \\ \eta q\mathbf{M} & q^2\mathbf{M} \end{pmatrix} \left(\sum_{k=0}^{T/2-k-1} \mathbf{A}^k\right)^\top \right\rangle.$$

Note that the matrices in $B_{T/2}$ are block-diagonal. Thus,

$$B_{T/2} = \sum_{i=1}^{d} \sum_{k=1}^{T/2} \lambda_i m_i \left( \left(\mathbf{I} - \mathbf{A}_i^{T/2-k}\right) (\mathbf{I} - \mathbf{A}_i)^{-1} \begin{pmatrix} \eta \\ q \end{pmatrix} \right)_1^2.$$

By the definition of $\tilde{\mathbf{\Pi}}$ in (21), we have

$$\tilde{\mathbf{\Pi}}_{T/2+1} = \sum_{k=0}^{T/2} \mathbf{A}^k \begin{pmatrix} \eta^2\mathbf{M} & \eta q\mathbf{M} \\ \eta q\mathbf{M} & q^2\mathbf{M} \end{pmatrix} \left(\mathbf{A}^k\right)^\top.$$

Thus, $A_{T/2}$ has the following form:

$$A_{T/2} = \sum_{i=1}^{d} \sum_{k=0}^{T/2} \lambda_i m_i \left( \mathbf{A}_i^k \left(\mathbf{I} - \mathbf{A}_i^{T/2}\right) (\mathbf{I} - \mathbf{A}_i)^{-1} \begin{pmatrix} \eta \\ q \end{pmatrix} \right)_1^2,$$

where $(\cdot)_1$ denotes the first element of a 2-dimensional vector. The result follows by applying $L_{\mathrm{var}}(\{\tilde{\mathbf{\Pi}}_t(\mathbf{M})\}_{t\in\mathbb{N}}) = 4(A_{T/2} + B_{T/2})/T^2$. □

The following lemma bounds $L_{\mathrm{var}}(\{\tilde{\mathbf{\Pi}}_t(\mathbf{M})\}_{t\in\mathbb{N}})$.

**Lemma D.13.** *Suppose we choose the hyperparameters as specified in Appendix C, then we have*

$$L_{\mathrm{var}}(\{\tilde{\mathbf{\Pi}}_t(\mathbf{M})\}_{t\in\mathbb{N}}) \leqslant \frac{64\,\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{M})}{T}.$$

*Proof.* By Lemma D.3 and Lemma D.5, we have

$$A_{T/2} = \sum_{i=1}^{d} \sum_{k=0}^{T/2} \frac{m_i}{\lambda_i} \left( \mathbf{A}_i^k \left(\mathbf{I} - \mathbf{A}_i^{T/2}\right) \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right)_1^2 \leqslant \sum_{i=1}^{d} \sum_{k=0}^{T/2} \frac{16 m_i}{\lambda_i} = \frac{8\,\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{M})}{T},$$

$$B_{T/2} = \sum_{i=1}^{d} \sum_{k=1}^{T/2} \lambda_i m_i \left( \left(\mathbf{I} - \mathbf{A}_i^{T/2-k}\right) \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right)_1^2 \leqslant \sum_{i=1}^{d} \sum_{k=0}^{T/2} \frac{16 m_i}{\lambda_i} = \frac{8\,\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{M})}{T}.$$

We complete the proof by $L_{\mathrm{var}}(\{\tilde{\mathbf{\Pi}}_t(\mathbf{M})\}_{t\in\mathbb{N}}) = 4(A_{T/2} + B_{T/2})/T^2 \leqslant 64\,\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{M})/T$. □

**Derivation of Stationary Covariance.** We derive the stationary state of $\tilde{\mathbf{\Pi}}_t(\mathbf{M})$. The following lemma shows that $\tilde{\mathbf{\Pi}}_t(\mathbf{M})$ is increasing.

**Lemma D.14.** *Let $\tilde{\mathbf{\Pi}}_t(\mathbf{M})$ be defined in* (21)*. Then*

$$\tilde{\mathbf{\Pi}}_0(\mathbf{M}) \preceq \tilde{\mathbf{\Pi}}_1(\mathbf{M}) \preceq \cdots \preceq \tilde{\mathbf{\Pi}}_t(\mathbf{M}) \preceq \tilde{\mathbf{\Pi}}_{t+1}(\mathbf{M}) \preceq \cdots$$

*Proof.* We expand the recursion of $\tilde{\mathbf{\Pi}}_t$ to obtain

$$\tilde{\mathbf{\Pi}}_t(\mathbf{M}) = \sum_{k=0}^{t-1} \mathbf{A}^k \begin{pmatrix} \eta^2\mathbf{M} & \eta q\mathbf{M} \\ \eta q\mathbf{M} & q^2\mathbf{M} \end{pmatrix} \left(\mathbf{A}^k\right)^\top.$$

Since it is a summation of PSD matrices, the conclusion follows. □

We define the stationary state of $\tilde{\mathbf{\Pi}}$ as follows:

$$\tilde{\mathbf{\Pi}}_\infty(\mathbf{M}) = \sum_{k=0}^{\infty} \mathbf{A}^k \begin{pmatrix} \eta^2\mathbf{M} & \eta q\mathbf{M} \\ \eta q\mathbf{M} & q^2\mathbf{M} \end{pmatrix} \left(\mathbf{A}^k\right)^\top, \tag{23}$$

so $\tilde{\mathbf{\Pi}}_t(\mathbf{M}) \preceq \tilde{\mathbf{\Pi}}_\infty(\mathbf{M})$ for all $t \in \mathbb{N}$. Note that

$$\tilde{\mathbf{\Pi}}_\infty(\mathbf{M}) = \mathbf{A}\tilde{\mathbf{\Pi}}_\infty(\mathbf{M})\mathbf{A}^\top + \begin{pmatrix} \eta^2\mathbf{M} & \eta q\mathbf{M} \\ \eta q\mathbf{M} & q^2\mathbf{M} \end{pmatrix}. \tag{24}$$

The following lemma explicitly calculates $\tilde{\mathbf{\Pi}}_\infty(\mathbf{M})$.

**Lemma D.15.** *Let $m_i$ denote the $i$-th diagonal element of $\mathbf{M}$, then*

$$\tilde{\mathbf{\Pi}}_{\infty,i}(\mathbf{M}) = m_i \begin{pmatrix} * & * \\ * & \dfrac{(1+c)(q-c\eta) + c\eta(q+c\eta)\lambda_i}{(2+2c-(q+c\eta)\lambda_i)(1-c+c\eta\lambda_i)\lambda_i} \end{pmatrix}$$

*Proof.* We note that (24) is a system of linear equations for $\mathbf{M}$. The desired result follows from direct calculations. □

The following bounds $\left\langle \begin{pmatrix} \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{\Sigma} \end{pmatrix}, \tilde{\mathbf{\Pi}}_\infty(\mathbf{M}) \right\rangle$.

**Lemma D.16.** *Suppose we choose the hyperparameter as specified in Appendix C, then we have*

$$\left\langle \begin{pmatrix} \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{\Sigma} \end{pmatrix}, \tilde{\mathbf{\Pi}}_\infty(\mathbf{M}) \right\rangle \leqslant \frac{\operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{M})}{4\tilde{\kappa}} + 4\eta \operatorname{tr}\mathbf{M}.$$

*Proof.* We have

$$\begin{aligned} \left\langle \begin{pmatrix} \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{\Sigma} \end{pmatrix}, \tilde{\mathbf{\Pi}}_\infty(\mathbf{M}) \right\rangle &\leqslant \sum_{i=1}^{d} \lambda_i \left(\tilde{\mathbf{\Pi}}_\infty(\mathbf{M})\right)_{22} \\ &\leqslant \sum_{i=1}^{d} \lambda_i m_i \cdot \frac{(1+c)(q-c\eta) + c\eta(q+c\eta)\lambda_i}{(2+2c-(q+c\eta)\lambda_i)(1-c+c\eta\lambda_i)\lambda_i}. \end{aligned}$$

Since $\eta \geqslant \theta\gamma$, and

$$q = \frac{\eta+\theta\gamma}{1+\theta} \leqslant \frac{2\eta}{1+\theta} = 2c\eta,$$

so we have

$$2 + 2c - (q+c\eta)\lambda_i \geqslant 2 + 2c - (2+c)\eta\lambda_i \geqslant 1.$$

For the first term,

$$\begin{aligned} \frac{(1+c)(q-c\eta)}{(1-c+c\eta\lambda_i)\lambda_i} &= \frac{1+c}{(1-c+c\eta\lambda_i)\lambda_i}\left(q - \frac{(1+c)\eta}{2} + \frac{(1-c)\eta}{2}\right) \\ &\leqslant \frac{1+c}{c\eta\lambda_i^2}\left(q - \frac{(1+c)\eta}{2}\right) + \frac{(1+c)\eta}{2\lambda_i} \\ &= \frac{2\theta\gamma}{(1-\theta^2)\eta\lambda_i^2} + \frac{\eta}{(1+\theta)\lambda_i} \overset{a}{=} \frac{1}{8(1-\theta^2)\tilde{\kappa}\lambda_i^2} + \frac{\eta}{(1+\theta)\lambda_i} \\ &\overset{b}{\leqslant} \frac{1}{4\tilde{\kappa}\lambda_i} + \frac{\eta}{\lambda_i} \end{aligned}$$

where $\overset{a}{=}$ uses $\eta = 16\tilde{\kappa}\theta\gamma$, and $\overset{b}{\leqslant}$ uses $1-\theta^2 \geqslant 1/2$. For the second term,

$$\frac{c\eta(q+c\eta)\lambda_i}{(1-c+c\eta\lambda_i)\lambda_i} \overset{a}{\leqslant} \frac{q+c\eta}{\lambda_i} = \frac{(3+\theta)\eta}{(1+\theta)\lambda_i} \leqslant \frac{3\eta}{\lambda_i},$$

where $\overset{a}{\leqslant}$ uses $c\eta\lambda_i \leqslant 1 - c + c\eta\lambda_i$. Combining the results, we obtain

$$\left\langle \begin{pmatrix} \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{\Sigma} \end{pmatrix}, \tilde{\mathbf{\Pi}}_\infty(\mathbf{M}) \right\rangle \leqslant \sum_{i=1}^{d} \frac{m_i}{4\tilde{\kappa}\lambda_i} + 4\eta m_i = \frac{\operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{M})}{4\tilde{\kappa}} + 4\eta \operatorname{tr}\mathbf{M}.$$

This completes the proof. □

### D.2.3 Decomposition of Noise Covariance

Define

$$\mathbf{R} = \mathbb{E}_{\mathbf{a},b\sim\mathcal{D}}\left(\varepsilon^2\mathbf{a}\mathbf{a}^\top\right). \tag{25}$$

where $\varepsilon$ is defined in (9). In this subsection, we decompose the $\mathbf{\Sigma}^{-1}$-norm of the stochastic gradient, which is $\operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{R})$, into a term $\operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{Q})$ corresponding to the noise at $\mathbf{x}^*$, and an additional term proportional to excess risk $F(\tilde{\mathbf{y}})-F(\mathbf{x}^*)$ in Lemma D.18. We derive a similar decomposition of the 2-norm of the stochastic gradient $\operatorname{tr}(\mathbf{R})$ in Lemma D.19. These results are key to deriving statistical complexity that is localized to $\mathbf{x}^*$. i.e., obtaining statistical term that only depends on $\mathbf{Q}$, which is the covariance of the stochastic gradient at $\mathbf{x}^*$.

We need the following lemma to bound the norm of the gradient of $F$.

**Lemma D.17.** *Suppose Assumption 1 holds. Then for any $\mathbf{x}\in\mathbb{R}^d$, we have*

$$\|\nabla F(\mathbf{x})\|^2_{\mathbf{\Sigma}^{-1}} \leqslant 2L_\ell(F(\mathbf{x})-F(\mathbf{x}^*)),$$

*where $\mathbf{x}^*$ is a minimizer of $F$.*

*Proof.* Let $\mathbf{u}=\mathbf{\Sigma}^{-1}\nabla F(\mathbf{x})/L_\ell$. Since $\nabla^2 F\preceq L_\ell\mathbf{\Sigma}$, so

$$F(\mathbf{x}^*)\leqslant F(\mathbf{x}-\mathbf{u})\leqslant F(\mathbf{x})-\langle\nabla F(\mathbf{x}),\mathbf{u}\rangle+\frac{1}{2}\|\mathbf{u}\|^2_{L_\ell\mathbf{\Sigma}}=F(\mathbf{x})-\frac{1}{2L_\ell}\|\nabla F(\mathbf{x})\|^2_{\mathbf{\Sigma}^{-1}}.$$

The result follows by rearranging the terms. □

With Lemma D.17, we can bound $\operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{R})$ as follows.

**Lemma D.18** (Noise Upper Bound)**.** *Suppose Assumptions 3 and 5 hold. Then*

$$\operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{R})\leqslant 5\operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{Q})+5L_\ell(L+\tilde{\kappa})(F(\tilde{\mathbf{y}})-F(\mathbf{x}^*)).$$

*Proof.* Note that by (9), we have

$$\begin{aligned}\varepsilon^2\mathbf{a}\mathbf{a}^\top \preceq\ & 5\left(\ell'(\mathbf{a}^\top\mathbf{x}^*,b)\right)^2\mathbf{a}\mathbf{a}^\top+\frac{5}{2}\left(\ell'(\mathbf{a}^\top\tilde{\mathbf{y}},b)-\ell'(\mathbf{a}^\top\mathbf{x}^*,b)\right)^2\mathbf{a}\mathbf{a}^\top\\ &+\frac{5}{2}\left(\mathbf{a}^\top(\tilde{\mathbf{y}}-\tilde{\mathbf{x}}^*_+)\right)^2\mathbf{a}\mathbf{a}^\top,\end{aligned} \tag{26}$$

where we use $(a+b+c)^2\leqslant 5a^2+5b^2/2+5c^2/2$. Taking inner product with $\mathbf{\Sigma}^{-1}$ and take expectation, the first term is bounded by $\operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{Q})$, and the second term is bounded by $2L_\ell L(F(\tilde{\mathbf{y}})-F(\mathbf{x}^*))$ by Assumption 5, since $\ell'(\mathbf{a}^\top\tilde{\mathbf{y}},b)\mathbf{a}$ is the stochastic gradient of $F(\tilde{\mathbf{y}})$. We bound the third term as follows

$$\mathbb{E}\left(\mathbf{a}^\top(\tilde{\mathbf{y}}-\tilde{\mathbf{x}}^*_+)\right)^2\|\mathbf{a}\|^2_{\mathbf{\Sigma}^{-1}}\overset{a}{\leqslant}\tilde{\kappa}\|\tilde{\mathbf{y}}-\tilde{\mathbf{x}}^*_+\|^2_{\mathbf{\Sigma}}=\tilde{\kappa}\|\nabla F(\tilde{\mathbf{y}})\|^2_{\mathbf{\Sigma}^{-1}}\overset{b}{\leqslant}2L_\ell\tilde{\kappa}(F(\tilde{\mathbf{y}})-F(\mathbf{x}^*)),$$

where $\overset{a}{\leqslant}$ uses Assumption 3, and $\overset{b}{\leqslant}$ uses Lemma D.17. The result follows by combining the three bounds. □

We also need the upper bound of $\operatorname{tr}\mathbf{R}$. The proof is similar to the proof of Lemma D.18.

**Lemma D.19.** *Suppose Assumptions 3 and 4 hold. Then*

$$\operatorname{tr}\mathbf{R}\leqslant 5\operatorname{tr}\mathbf{Q}+5L_\ell(B+R^2)(F(\tilde{\mathbf{y}})-F(\mathbf{x}^*)).$$

*Proof.* We bound the trace of (26). The first term is bounded by $\operatorname{tr}\mathbf{Q}$. The second term is bounded by

$$\mathbb{E}\left(\ell'(\mathbf{a}^\top\tilde{\mathbf{y}},b)-\ell'(\mathbf{a}^\top\mathbf{x}^*,b)\right)^2\|\mathbf{a}\|^2\leqslant 2L_\ell B(F(\tilde{\mathbf{y}})-F(\mathbf{x}^*)),$$

where we use Assumption 4. The third term is bounded by

$$R^2\|\tilde{\mathbf{y}}-\tilde{\mathbf{x}}^*_+\|^2_{\mathbf{\Sigma}}=R^2\|\nabla F(\tilde{\mathbf{y}})\|^2_{\mathbf{\Sigma}^{-1}}\overset{a}{\leqslant}2L_\ell R^2(F(\tilde{\mathbf{y}})-F(\mathbf{x}^*)),$$

where we use Assumption 3. The result follows by combining the three bounds. □

### D.2.4 Bound of Layer 0 Dynamics $\tilde{\mathbf{C}}_t$

By the definition of $\tilde{\mathbf{\Pi}}$ in (21), we have

$$\tilde{\mathbf{C}}_t = \tilde{\mathbf{\Pi}}_t(\mathbf{R}).$$

Therefore, by applying Lemma D.13 and Lemma D.16 in Appendix D.2.2, we have the following lemmas.

**Lemma D.20.** *Suppose we choose the hyperparameters as specified in Appendix C. Then we have*

$$L_{\mathrm{var}}\left(\left\{\tilde{\mathbf{C}}_t\right\}_{t\in\mathbb{N}}\right) \leqslant \frac{320\,\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{Q})}{T} + \frac{320(L+\tilde{\kappa})(F(\tilde{\mathbf{y}})-F(\mathbf{x}^*))}{T}.$$

**Lemma D.21.** *Suppose we choose the hyperparameters as specified in Appendix C, then we have*

$$\left\langle \begin{pmatrix} \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{\Sigma} \end{pmatrix}, \tilde{\mathbf{C}}_\infty \right\rangle \leqslant \frac{5\,\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{Q}) + 5(L+\tilde{\kappa})(F(\tilde{\mathbf{y}})-F(\mathbf{x}^*))}{4\tilde{\kappa}} + 20\eta\,\mathrm{tr}\,\mathbf{Q} + 20\eta(B+R^2)(F(\tilde{\mathbf{y}})-F(\mathbf{x}^*))$$

### D.2.5 Bound of Layer $\ell$ Dynamics $\tilde{\mathbf{C}}_t^{(\ell)}$

The following lemma derives a decomposition of $\mathbb{E}\|\mathbf{v}_k\|_{\mathbf{H}}^2$.

**Lemma D.22.** *We have*

$$\mathbb{E}\|\mathbf{v}_k\|_{\mathbf{H}}^2 = L_{\mathrm{var}}\left(\left\{\tilde{\mathbf{C}}_t\right\}_{t\in\mathbb{N}}\right) + \sum_{\ell=1}^{\infty} L_{\mathrm{var}}\left(\left\{\tilde{\mathbf{C}}_t^{(\ell)}\right\}_{t\in\mathbb{N}}\right),$$

*Proof.* By Lemma D.10, we have

$$\mathbf{C}_t = \tilde{\mathbf{C}}_t + \sum_{\ell=1}^{\infty} \tilde{\mathbf{C}}_t^{(\ell)}.$$

The desired result follows from the linearity of $L_{\mathrm{var}}$ in Lemma D.7. □

Define the $\ell$-level stationary state $\tilde{\mathbf{C}}_\infty^{(\ell)}$ as follows:

$$\tilde{\mathbf{C}}_\infty^{(\ell)} = \begin{pmatrix} \tilde{\mathbf{C}}_{\infty,11}^{(\ell)} & \tilde{\mathbf{C}}_{\infty,12}^{(\ell)} \\ \tilde{\mathbf{C}}_{\infty,21}^{(\ell)} & \tilde{\mathbf{C}}_{\infty,22}^{(\ell)} \end{pmatrix} \overset{\mathrm{def}}{=} \tilde{\mathbf{\Pi}}_\infty(\mathbf{R}^{(\ell)}), \quad \text{where } \mathbf{R}^{(\ell)} = (\mathcal{M}-\tilde{\mathcal{M}}) \circ \tilde{\mathbf{C}}_{\infty,22}^{(\ell-1)}.$$

We set $\tilde{\mathbf{C}}_\infty^{(0)} = \tilde{\mathbf{C}}_\infty \overset{\mathrm{def}}{=} \tilde{\mathbf{\Pi}}_\infty(\mathbf{R})$ defined in (23). One can verify $\tilde{\mathbf{C}}_t^{(\ell)} \preceq \tilde{\mathbf{C}}_\infty^{(\ell)}$. Thus,

$$L_{\mathrm{var}}\left(\left\{\tilde{\mathbf{C}}_t^{(\ell)}\right\}_{t\in\mathbb{N}}\right) \leqslant L_{\mathrm{var}}\left(\left\{\tilde{\mathbf{\Pi}}_t(\mathbf{R}^{(\ell)})\right\}_{t\in\mathbb{N}}\right)$$

We apply the results in Appendix D.2.2 for $\mathbf{M} = \mathbf{R}^{(\ell)}$ to obtain the following lemmas.

**Lemma D.23.** *We have*

$$L_{\mathrm{var}}\left(\left\{\tilde{\mathbf{C}}_t^{(\ell)}\right\}_{t\in\mathbb{N}}\right) \leqslant L_{\mathrm{var}}\left(\left\{\tilde{\mathbf{\Pi}}_t(\mathbf{R}^{(\ell)})\right\}_{t\in\mathbb{N}}\right) \leqslant \frac{64\,\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{R}^{(\ell)})}{T}.$$

**Lemma D.24.** *We have*

$$\left\langle \begin{pmatrix} \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{\Sigma} \end{pmatrix}, \tilde{\mathbf{C}}_t^{(\ell)} \right\rangle \leqslant \left\langle \begin{pmatrix} \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{\Sigma} \end{pmatrix}, \tilde{\mathbf{C}}_\infty^{(\ell)} \right\rangle \leqslant \frac{\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{R}^{(\ell)})}{4\tilde{\kappa}} + 4\eta\,\mathrm{tr}\,\mathbf{R}^{(\ell)}.$$

We define

$$\sigma^{(\ell)} = \left\langle \begin{pmatrix} \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{\Sigma} \end{pmatrix}, \tilde{\mathbf{C}}_\infty^{(\ell)} \right\rangle.$$

The following lemma is an implication of Assumption 3, which is the first step to relate $\sigma^{(\ell-1)}$ to $\sigma^{(\ell)}$.

**Lemma D.25.** *Suppose Assumption 3 holds, then we have*

$$\operatorname{tr}\mathbf{R}^{(\ell)} \leqslant R^2\sigma^{(\ell-1)}, \quad \operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{R}^{(\ell)}) \leqslant \tilde{\kappa}\sigma^{(\ell-1)}.$$

*Proof.* By the definition of $\mathbf{R}^{(\ell)}$,

$$\mathbf{R}^{(\ell)} = (\mathcal{M} - \tilde{\mathcal{M}}) \circ \tilde{\mathbf{C}}_{\infty,22}^{(\ell-1)} = \mathbb{E}\left(\|\mathbf{a}\|^2_{\tilde{\mathbf{C}}_{\infty,22}^{(\ell-1)}}\mathbf{a}\mathbf{a}^\top\right) - \mathbf{\Sigma}\tilde{\mathbf{C}}_{\infty,22}^{(\ell-1)}\mathbf{\Sigma}.$$

We drop the last term, which is a PSD matrix, so

$$\begin{aligned}
\operatorname{tr}\mathbf{R}^{(\ell)} &\leqslant \mathbb{E}\left(\|\mathbf{a}\|^2\|\mathbf{a}\|^2_{\tilde{\mathbf{C}}_{\infty,22}^{(\ell-1)}}\right) = \left\langle \mathbb{E}\left(\|\mathbf{a}\|^2\mathbf{a}\mathbf{a}^\top\right), \tilde{\mathbf{C}}_{\infty,22}^{(\ell-1)}\right\rangle \\
&\overset{a}{\leqslant} \left\langle R^2\mathbf{\Sigma}, \tilde{\mathbf{C}}_{\infty,22}^{(\ell-1)}\right\rangle = R^2\left\langle \begin{pmatrix}\mathbf{O} & \mathbf{O}\\ \mathbf{O} & \mathbf{\Sigma}\end{pmatrix}, \tilde{\mathbf{C}}_{\infty}^{(\ell-1)}\right\rangle \overset{b}{\leqslant} R^2\sigma^{(\ell-1)}. \\
\operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{R}^{(\ell)}) &\leqslant \mathbb{E}\left(\|\mathbf{a}\|^2_{\mathbf{\Sigma}^{-1}}\|\mathbf{a}\|^2_{\tilde{\mathbf{C}}_{\infty,22}^{(\ell-1)}}\right) = \left\langle \mathbb{E}\left(\|\mathbf{a}\|^2_{\mathbf{\Sigma}^{-1}}\mathbf{a}\mathbf{a}^\top\right), \tilde{\mathbf{C}}_{\infty,22}^{(\ell-1)}\right\rangle \\
&\overset{a}{\leqslant} \left\langle \tilde{\kappa}\mathbf{\Sigma}, \tilde{\mathbf{C}}_{\infty,22}^{(\ell-1)}\right\rangle = \tilde{\kappa}\left\langle \begin{pmatrix}\mathbf{O} & \mathbf{O}\\ \mathbf{O} & \mathbf{\Sigma}\end{pmatrix}, \tilde{\mathbf{C}}_{\infty}^{(\ell-1)}\right\rangle \overset{b}{\leqslant} \tilde{\kappa}\sigma^{(\ell-1)}.
\end{aligned}$$

where $\overset{a}{\leqslant}$ is from Assumption 3, and $\overset{b}{\leqslant}$ uses Lemma D.24. □

Combine Lemma D.24 and Lemma D.25, we relate $\sigma^{(\ell-1)}$ to $\sigma^{(\ell)}$ as follows.

**Lemma D.26.** *We have $\sigma^{(\ell)} \leqslant \sigma^{(\ell-1)}/2$ for all $\ell \geqslant 1$.*

*Proof.* Note that

$$\sigma^{(\ell)} \overset{a}{\leqslant} \frac{\operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{R}^{(\ell)})}{4\tilde{\kappa}} + 4\eta\operatorname{tr}\mathbf{R}^{(\ell)} \overset{a}{\leqslant} \frac{1}{4\tilde{\kappa}}\cdot\tilde{\kappa}\sigma^{(\ell-1)} + 4\eta\cdot R^2\sigma^{(\ell-1)} \overset{b}{\leqslant} \frac{1}{2}\sigma^{(\ell-1)},$$

where $\overset{a}{\leqslant}$ uses Lemma D.24, $\overset{b}{\leqslant}$ applies Lemma D.25, and $\overset{c}{\leqslant}$ uses $\eta \leqslant 1/(16R^2)$. □

The follow lemma bounds $L_{\mathrm{var}}(\{\tilde{\mathbf{C}}_t^{(\ell)}\}_{t\in\mathbb{N}})$ for $\ell \geqslant 1$

**Lemma D.27.** *We have*

$$\begin{aligned}
L_{\mathrm{var}}\left(\left\{\tilde{\mathbf{C}}_t^{(\ell)}\right\}_{t\in\mathbb{N}}\right) \leqslant \frac{64\tilde{\kappa}}{2^{\ell-1}T}\Bigg(&\frac{5\operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{Q}) + 5L_\ell(L+\tilde{\kappa})(F(\tilde{\mathbf{y}}) - F(\mathbf{x}^*))}{4\tilde{\kappa}} \\
&+20\eta\operatorname{tr}\mathbf{Q} + 20\eta L_\ell(B+R^2)(F(\tilde{\mathbf{y}}) - F(\mathbf{x}^*))\Bigg).
\end{aligned}$$

*Proof.* By Lemma D.21, we have

$$\begin{aligned}
\sigma^{(0)} \overset{\text{def}}{=} \left\langle \begin{pmatrix}\mathbf{O} & \mathbf{O}\\ \mathbf{O} & \mathbf{\Sigma}\end{pmatrix}, \tilde{\mathbf{C}}_\infty\right\rangle \leqslant &\frac{5\operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{Q}) + 5L_\ell(L+\tilde{\kappa})(F(\tilde{\mathbf{y}}) - F(\mathbf{x}^*))}{4\tilde{\kappa}} \\
&+20\eta L_\ell\operatorname{tr}\mathbf{Q} + 20\eta(B+R^2)(F(\tilde{\mathbf{y}}) - F(\mathbf{x}^*)).
\end{aligned}$$

By Lemma D.26, we have

$$\begin{aligned}
L_{\mathrm{var}}\left(\left\{\tilde{\mathbf{C}}_t^{(\ell)}\right\}_{t\in\mathbb{N}}\right) \leqslant L_{\mathrm{var}}\left(\left\{\tilde{\mathbf{\Pi}}_t(\mathbf{R}^{(\ell)})\right\}_{t\in\mathbb{N}}\right) &\leqslant \frac{64\operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{R}^{(\ell)})}{T} \leqslant \frac{64\tilde{\kappa}\sigma^{(\ell)}}{T} \leqslant \frac{64\tilde{\kappa}\sigma^{(0)}}{2^{\ell-1}T} \\
&\leqslant \frac{64\tilde{\kappa}}{2^{\ell-1}T}\Bigg(\frac{5\operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{Q}) + 5L_\ell(L+\tilde{\kappa})(F(\tilde{\mathbf{y}}) - F(\mathbf{x}^*))}{4\tilde{\kappa}} \\
&\quad +20\eta\operatorname{tr}\mathbf{Q} + 20\eta L_\ell(B+R^2)(F(\tilde{\mathbf{y}}) - F(\mathbf{x}^*))\Bigg),
\end{aligned}$$

This completes the proof. □

Finally, we are ready to bound $\sum_{\ell=1}^{\infty} L_{\mathrm{var}}(\{\tilde{\mathbf{C}}_t^{(\ell)}\}_{t\in\mathbb{N}})$.

**Lemma D.28.** *Suppose Assumptions 3, 4 and 5 hold, then we have*

$$\sum_{\ell=1}^{\infty} L_{\mathrm{var}}\left(\left\{\tilde{\mathbf{C}}_t^{(\ell)}\right\}_{t\in\mathbb{N}}\right) \leqslant \frac{640\,\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{Q}) + 640L_\ell(L+\tilde{\kappa})(F(\tilde{\mathbf{y}})-F(\mathbf{x}^*))}{T} + \frac{2560\eta\tilde{\kappa}\,\mathrm{tr}\,\mathbf{Q} + 2560\eta\tilde{\kappa}L_\ell(B+R^2)(F(\tilde{\mathbf{y}})-F(\mathbf{x}^*))}{T}.$$

*Proof.* We apply Lemma D.26,

$$\begin{aligned}\sum_{\ell=1}^{\infty} L_{\mathrm{var}}\left(\left\{\tilde{\mathbf{C}}_t^{(\ell)}\right\}_{t\in\mathbb{N}}\right) &\leqslant \sum_{\ell=1}^{\infty}\frac{64\tilde{\kappa}}{2^{\ell-1}T}\left(\frac{5\,\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{Q}) + 5L_\ell(L+\tilde{\kappa})(F(\tilde{\mathbf{y}})-F(\mathbf{x}^*))}{4\tilde{\kappa}}\right.\\ &\quad\left.+20\eta\,\mathrm{tr}\,\mathbf{Q} + 20\eta L_\ell(B+R^2)(F(\tilde{\mathbf{y}})-F(\mathbf{x}^*))\right)\\ &\leqslant \frac{640\,\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{Q}) + 640L_\ell(L+\tilde{\kappa})(F(\tilde{\mathbf{y}})-F(\mathbf{x}^*))}{T}\\ &\quad+\frac{2560\eta\tilde{\kappa}\,\mathrm{tr}\,\mathbf{Q} + 2560\eta\tilde{\kappa}L_\ell(B+R^2)(F(\tilde{\mathbf{y}})-F(\mathbf{x}^*))}{T}.\end{aligned}$$

This completes the proof. □

### D.2.6 Proof of Lemma D.6

*Proof of Lemma D.6.* By the definition of $\boldsymbol{\eta}_t^{\mathrm{var}}$, we have $\mathbb{E}\boldsymbol{\eta}_t = \boldsymbol{\eta}_0 = \mathbf{0}$. Thus, by the definition of $\mathbf{v}$ in Definition D.1, we have $\mathbb{E}\mathbf{v} = \mathbf{0}$. To bound $\mathbb{E}\|\mathbf{v}\|_{\boldsymbol{\Sigma}}^2$, we apply Lemma D.22,

$$\mathbb{E}\|\mathbf{v}_k\|_{\mathbf{H}}^2 = L_{\mathrm{var}}\left(\left\{\tilde{\mathbf{C}}_t\right\}_{t\in\mathbb{N}}\right) + \sum_{\ell=1}^{\infty} L_{\mathrm{var}}\left(\left\{\tilde{\mathbf{C}}_t^{(\ell)}\right\}_{t\in\mathbb{N}}\right).$$

For the bound of $\mathbf{v}_k$, we have

$$\begin{aligned}\mathbb{E}\|\mathbf{v}_k\|_{\boldsymbol{\Sigma}}^2 &= L_{\mathrm{var}}\left(\left\{\tilde{\mathbf{C}}_t\right\}_{t\in\mathbb{N}}\right) + \sum_{\ell=1}^{\infty} L_{\mathrm{var}}\left(\left\{\tilde{\mathbf{C}}_t^{(\ell)}\right\}_{t\in\mathbb{N}}\right)\\ &\overset{a}{\leqslant} \frac{320\,\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{Q})}{T} + \frac{320L_\ell(L+\tilde{\kappa})(F(\tilde{\mathbf{y}})-F(\mathbf{x}^*))}{T}\\ &\quad+\frac{640\,\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{Q}) + 640L_\ell(L+\tilde{\kappa})(F(\tilde{\mathbf{y}})-F(\mathbf{x}^*))}{T}\\ &\quad+\frac{2560\eta\tilde{\kappa}\,\mathrm{tr}\,\mathbf{Q} + 2560\eta\tilde{\kappa}L_\ell(B+R^2)(F(\tilde{\mathbf{y}})-F(\mathbf{x}^*))}{T}.\end{aligned}$$

where $\overset{a}{\leqslant}$ applies Lemma D.20 and Lemma D.28. This completes the proof. □

## D.3 Bias Upper Bound

The goal of this section is to prove the following lemma.

**Lemma D.29** (Bias Upper Bound)**.** *Suppose Assumptions 1, 2, 3, 4 and 5 hold. Let* $\mathbf{r}$ *defined in Definition D.1. Then we have*

$$\mathbb{E}\|\mathbf{r}\|_{\boldsymbol{\Sigma}}^2 \leqslant \frac{L_\ell}{8\alpha}\|\nabla F(\tilde{\mathbf{y}}_{k-1})\|_{\boldsymbol{\Sigma}^{-1}}^2,$$

*where the expectation is taken with respect to the samples drawn in the* $k$*-th outer iteration.*

The proof is deferred to the end of this subsection.

Recall the definition of $\boldsymbol{\eta}_t^{\text{bias}} \in \mathbb{R}^{2d}$ in (12):

$$\boldsymbol{\eta}_t^{\text{bias}} = \widehat{\mathbf{A}}_t \boldsymbol{\eta}_{t-1}^{\text{bias}}, \quad \boldsymbol{\eta}_0^{\text{bias}} = \begin{pmatrix} \tilde{\mathbf{y}} - \tilde{\mathbf{x}}_+^* \\ \tilde{\mathbf{y}} - \tilde{\mathbf{x}}_+^* \end{pmatrix}.$$

Let

$$\boldsymbol{\eta}_t^{\text{bias}} = \begin{pmatrix} \mathbf{x}_t^{\text{bias}} \\ \mathbf{y}_t^{\text{bias}} \end{pmatrix}, \quad \text{where } \mathbf{x}_t^{\text{bias}}, \mathbf{y}_t^{\text{bias}} \in \mathbb{R}^d$$

For simplicity, we drop the superscript of $\mathbf{x}_t^{\text{bias}}, \mathbf{y}_t^{\text{bias}}, \mathbf{z}_t^{\text{bias}}$ in this subsection. The bias iteration can be written in the following form:

$$\begin{aligned} \mathbf{y}_{t-1} &= \frac{1}{1+\theta}\mathbf{x}_{t-1} + \frac{\theta}{1+\theta}\mathbf{z}_{t-1}, \\ \mathbf{x}_t &= \mathbf{y}_{t-1} - \eta \mathbf{a}_t \mathbf{a}_t^\top (\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*), \\ \mathbf{z}_t &= \theta \mathbf{y}_{t-1} + (1-\theta)\mathbf{z}_{t-1} - \gamma \mathbf{a}_t \mathbf{a}_t^\top (\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*), \end{aligned}$$

where $\mathbf{x}_0 = \mathbf{z}_0 = \tilde{\mathbf{y}} - \tilde{\mathbf{x}}_+^*$. The following lemma shows that $\|\mathbf{x}_t - \tilde{\mathbf{x}}_+^*\|$ converges exponentially fast.

**Lemma D.30.** *Suppose Assumption 3 holds and let the hyperparameter be chosen as specified in Appendix C, then we have*

$$\mathbb{E}\|\mathbf{x}_T - \tilde{\mathbf{x}}_+^*\|^2 + \frac{\theta}{2\gamma}\mathbb{E}\|\mathbf{z}_T - \tilde{\mathbf{x}}_+^*\|_{\boldsymbol{\Sigma}^{-1}}^2 \leqslant (1-\theta)^T \left( \|\mathbf{x}_0 - \tilde{\mathbf{x}}_+^*\|^2 + \frac{\theta}{2\gamma}\|\mathbf{x}_0 - \tilde{\mathbf{x}}_+^*\|_{\boldsymbol{\Sigma}^{-1}}^2 \right). \quad (27)$$

*Proof.* Let $\mathbb{E}_t$ denote the expectation conditioned on $\{(\mathbf{a}_s, b_s)\}_{s=1}^t$. We consider the energy function $\|\mathbf{x}_t - \tilde{\mathbf{x}}_+^*\|^2/2$. By the update rule,

$$\begin{aligned} \frac{1}{2}\mathbb{E}_{t-1}\|\mathbf{x}_t - \tilde{\mathbf{x}}_+^*\|^2 &= \frac{1}{2}\|\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*\|^2 + \mathbb{E}_{t-1}\langle \mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*, \mathbf{x}_t - \mathbf{y}_{t-1}\rangle + \frac{1}{2}\mathbb{E}_{t-1}\|\mathbf{x}_t - \mathbf{y}_{t-1}\|^2 \\ &= \frac{1}{2}\|\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*\|^2 - \left\langle \mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*, \boldsymbol{\Sigma}(\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*) \right\rangle \\ &\quad + \frac{\eta^2}{2}\mathbb{E}_{t-1}\left\| \mathbf{a}_t \mathbf{a}_t^\top (\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*) \right\|^2 \\ &\overset{a}{\leqslant} \frac{1}{2}\|\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*\|^2 - \left( \eta - \frac{\eta^2 R^2}{2} \right) \|\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*\|_{\boldsymbol{\Sigma}}^2 \\ &\overset{b}{\leqslant} \frac{1}{2}\|\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*\|^2 - \frac{\eta}{2}\|\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*\|_{\boldsymbol{\Sigma}}^2, \end{aligned} \quad (28)$$

where $\overset{a}{\leqslant}$ uses $\mathbb{E}\left(\|\mathbf{a}\|^2 \mathbf{a}\mathbf{a}^\top\right) \preceq R^2\boldsymbol{\Sigma}$, and $\overset{b}{\leqslant}$ uses $\eta \leqslant 1/R^2$. Then, we have to bound $\|\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*\|^2/2$. By the convexity of $\|\cdot\|$, we have

$$\frac{1}{2}\|\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*\|^2 \leqslant \langle \mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*, \mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^* \rangle - \frac{1}{2}\|\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*\|^2, \quad (29)$$

$$\frac{1}{2}\|\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*\|^2 \leqslant \frac{1}{2}\|\mathbf{x}_{t-1} - \tilde{\mathbf{x}}_+^*\|^2 + \langle \mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*, \mathbf{y}_{t-1} - \mathbf{x}_{t-1} \rangle. \quad (30)$$

Let $\mathbf{x} = (1-\theta)\mathbf{x}_{t-1} + \theta \mathbf{x}_+^*$, then $\theta \times (29) + (1-\theta) \times (30)$ yields

$$\frac{1}{2}\|\mathbf{y}_{t-1} - \mathbf{x}_{t-1}\|^2 \leqslant \frac{1}{2}(1-\theta)\|\mathbf{x}_{t-1} - \tilde{\mathbf{x}}_+^*\|^2 + \langle \mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*, \mathbf{y}_{t-1} - \mathbf{x} \rangle - \frac{\theta}{2}\|\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*\|^2.$$

Let $\mathbf{w}_{t-1} = \theta \mathbf{y}_{t-1} + (1-\theta)\mathbf{z}_{t-1}$, and note that

$$\mathbb{E}(\mathbf{x}_t - \mathbf{y}_{t-1}) = -\eta \boldsymbol{\Sigma}(\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*),$$

$$\mathbf{y}_{t-1} - \mathbf{x} = \theta(\mathbf{w}_k - \tilde{\mathbf{x}}_+^*), \quad \mathbf{x}_t - \mathbf{y}_{t-1} = \frac{\eta}{\gamma}(\mathbf{z}_t - \mathbf{w}_{t-1}).$$

Therefore, we have

$$\begin{aligned}
&\frac{1}{2}\|\mathbf{y}_{t-1}-\mathbf{x}_{t-1}\|^2\\
\leqslant&\frac{1}{2}(1-\theta)\|\mathbf{x}_{t-1}-\tilde{\mathbf{x}}_+^*\|^2+\frac{\theta}{\gamma}\mathbb{E}\langle\mathbf{w}_{t-1}-\mathbf{z}_{t-1},\mathbf{w}_{t-1}-\tilde{\mathbf{x}}_+^*\rangle_{\mathbf{\Sigma}^{-1}}-\frac{\theta}{2}\|\mathbf{y}_{t-1}-\tilde{\mathbf{x}}_+^*\|^2\\
=&\frac{1}{2}(1-\theta)\|\mathbf{x}_{t-1}-\tilde{\mathbf{x}}_+^*\|^2-\frac{\theta}{2}\|\mathbf{y}_{t-1}-\tilde{\mathbf{x}}_+^*\|^2\\
&+\frac{\theta}{2\gamma}\left(\|\mathbf{w}_{t-1}-\tilde{\mathbf{x}}_+^*\|_{\mathbf{\Sigma}^{-1}}^2+\mathbb{E}_{t-1}\|\mathbf{z}_t-\mathbf{w}_{t-1}\|_{\mathbf{\Sigma}^{-1}}^2-\mathbb{E}_{t-1}\|\mathbf{z}_t-\tilde{\mathbf{x}}_+^*\|_{\mathbf{\Sigma}^{-1}}^2\right)\\
\overset{a}{\leqslant}&\frac{1}{2}(1-\theta)\|\mathbf{x}_{t-1}-\tilde{\mathbf{x}}_+^*\|^2-\frac{\theta}{2}\|\mathbf{y}_{t-1}-\tilde{\mathbf{x}}_+^*\|^2+\frac{\theta^2}{2\gamma}\|\mathbf{y}_{t-1}-\tilde{\mathbf{x}}_+^*\|_{\mathbf{\Sigma}^{-1}}^2\\
&+\frac{\theta}{2\gamma}\left((1-\theta)\|\mathbf{z}_{t-1}-\tilde{\mathbf{x}}_+^*\|_{\mathbf{\Sigma}^{-1}}^2+\mathbb{E}_{t-1}\|\mathbf{z}_t-\mathbf{w}_{t-1}\|_{\mathbf{\Sigma}^{-1}}^2-\mathbb{E}_{t-1}\|\mathbf{z}_t-\tilde{\mathbf{x}}_+^*\|_{\mathbf{\Sigma}^{-1}}^2\right).
\end{aligned}$$

Apply Assumption 3, we have

$$\begin{aligned}
\frac{1}{2}\|\mathbf{y}_{t-1}-\mathbf{x}_{t-1}\|^2\leqslant&\frac{1}{2}(1-\theta)\|\mathbf{x}_{t-1}-\tilde{\mathbf{x}}_+^*\|^2+\frac{\tilde{\kappa}\theta\gamma}{2}\|\mathbf{y}_{t-1}-\tilde{\mathbf{x}}_+^*\|_{\mathbf{\Sigma}}^2-\frac{\theta(\mu\gamma-\theta)}{2\mu\gamma}\|\mathbf{y}_{t-1}-\tilde{\mathbf{x}}_+^*\|^2\\
&+\frac{\theta}{2\gamma}\left((1-\theta)\|\mathbf{z}_{t-1}-\tilde{\mathbf{x}}_+^*\|_{\mathbf{\Sigma}^{-1}}^2-\mathbb{E}_{t-1}\|\mathbf{z}_t-\tilde{\mathbf{x}}_+^*\|_{\mathbf{\Sigma}^{-1}}^2\right).
\end{aligned}$$

Combining the above inequality and (28), we have

$$\begin{aligned}
&\frac{1}{2}\mathbb{E}_{t-1}\|\mathbf{x}_t-\tilde{\mathbf{x}}_+^*\|^2\\
\leqslant&\frac{1}{2}(1-\theta)\|\mathbf{x}_{t-1}-\tilde{\mathbf{x}}_+^*\|^2+\frac{\theta}{2\gamma}\left((1-\theta)\|\mathbf{z}_{t-1}-\tilde{\mathbf{x}}_+^*\|_{\mathbf{\Sigma}^{-1}}^2-\mathbb{E}_{t-1}\|\mathbf{z}_t-\tilde{\mathbf{x}}_+^*\|_{\mathbf{\Sigma}^{-1}}^2\right)\\
&-\frac{\eta-\tilde{\kappa}\theta\gamma}{2}\|\mathbf{y}_{t-1}-\tilde{\mathbf{x}}_+^*\|_{\mathbf{\Sigma}}^2-\frac{\theta(\gamma\mu-\theta)}{2\mu\gamma}\|\mathbf{y}_{t-1}-\tilde{\mathbf{x}}_+^*\|^2.
\end{aligned}$$

By the hyperparameter choice, we have $\eta-\tilde{\kappa}\theta\gamma\geqslant 0$ and $\gamma\mu-\theta\geqslant 0$, so

$$\mathbb{E}_{t-1}\|\mathbf{x}_t-\tilde{\mathbf{x}}_+^*\|^2+\frac{\theta}{2\gamma}\mathbb{E}_{t-1}\|\mathbf{z}_t-\tilde{\mathbf{x}}_+^*\|_{\mathbf{\Sigma}^{-1}}^2\leqslant(1-\theta)\left(\|\mathbf{x}_{t-1}-\tilde{\mathbf{x}}_+^*\|^2+\frac{\theta}{2\gamma}\|\mathbf{z}_{t-1}-\tilde{\mathbf{x}}_+^*\|_{\mathbf{\Sigma}^{-1}}^2\right).$$

Thus, we take the expectation and obtain

$$\mathbb{E}\|\mathbf{x}_T-\tilde{\mathbf{x}}_+^*\|^2+\frac{\theta}{2\gamma}\mathbb{E}\|\mathbf{z}_T-\tilde{\mathbf{x}}_+^*\|_{\mathbf{\Sigma}^{-1}}^2\leqslant(1-\theta)^T\left(\|\mathbf{x}_0-\tilde{\mathbf{x}}_+^*\|^2+\frac{\theta}{2\gamma}\|\mathbf{x}_0-\tilde{\mathbf{x}}_+^*\|_{\mathbf{\Sigma}^{-1}}^2\right).$$

This completes the proof. □

We are ready to prove Lemma D.29.

*Proof of Lemma D.29.* By Lemma D.30, we have

$$\begin{aligned}
\mathbb{E}\|\mathbf{r}_k\|_{\mathbf{\Sigma}}^2&\leqslant\lambda_{\max}(\mathbf{\Sigma})(1-\theta)^T\left(\|\mathbf{x}_0-\mathbf{x}_+^*\|^2+\frac{\theta}{2\gamma}\|\mathbf{x}_0-\mathbf{x}_+^*\|_{\mathbf{\Sigma}^{-1}}^2\right)\\
&\overset{a}{\leqslant}\frac{L_\ell}{8\alpha}\|\nabla F(\tilde{\mathbf{y}}_{k-1})\|_{\mathbf{\Sigma}^{-1}}^2,
\end{aligned}$$

where $\overset{a}{\leqslant}$ follows from the choice of $T$ and $\theta$ in Appendix C. □

### D.4 Proof of Lemma D.1

*Proof of Lemma D.1.* Recall that $\mathbf{H}=L_\ell\mathbf{\Sigma}$. The desired results follow by combining Lemma D.29 and Lemma D.6. □

### D.5 Auxiliary Lemmas

**Lemma D.31.** *Suppose Assumption 3 holds, then $R^2 \geqslant \operatorname{tr}\mathbf{\Sigma}$ and $\tilde{\kappa} \geqslant d$.*

*Proof.* For the first statement, note that

$$(\operatorname{tr}\mathbf{\Sigma})^2 = \left(\mathbb{E}\|\mathbf{a}\|^2\right)^2 \leqslant \mathbb{E}\|\mathbf{a}\|^4 \leqslant R^2 \operatorname{tr}\mathbf{\Sigma}.$$

For the second statement, note that

$$d^2 = \left(\mathbb{E}\|\mathbf{a}\|_{\mathbf{\Sigma}^{-1}}^2\right)^2 \leqslant \mathbb{E}\|\mathbf{a}\|_{\mathbf{\Sigma}^{-1}}^4 \leqslant \tilde{\kappa} d.$$

The results follow by cancelling $\operatorname{tr}\mathbf{\Sigma}$ or $d$. □

**Lemma D.32.** *For $k \geqslant 0$, we have $x^k\left[1 + k(1-x)\right] \leqslant 1$ for $x \in [0,1]$.*

*Proof.* Let $f(x) = x^k\left[1 + k(1-x)\right]$. The derivative of $f(x)$ is $f'(x) = k(k+1)x^{k-1}(1-x) \geqslant 0$ for $x \in [0,1]$. Thus, we have $f(x) \leqslant f(1) = 1$. □

## E Part II: Analysis of Outer Loop

In this section, we analyze the outer loop of Algorithm 1. In Section E.1, we restate Lemma D.1, under which we analyze the outer loop.

### E.1 Excess Risk Bound of Outer Loop

Let $\mathcal{F}_k$ denote the filtration generated by the samples of the first $k$ inner loops. By Lemma D.1, we write the update of the outer loop as follows:

$$\tilde{\mathbf{y}}_{k-1} = \tilde{\mathbf{x}}_{k-1} + \beta(\tilde{\mathbf{x}}_{k-1} - \tilde{\mathbf{x}}_{k-2}), \tag{31}$$

$$\tilde{\mathbf{x}}_k = \tilde{\mathbf{y}}_{k-1} - L_\ell h_k \mathbf{H}^{-1}\nabla F(\tilde{\mathbf{y}}_{k-1}) + h_k \mathbf{r}_k + h_k \mathbf{v}_k. \tag{32}$$

where $\mathbf{r}_k$ and $\mathbf{v}_k$ satisfies

$$\begin{aligned}
\mathbb{E}(\|\mathbf{r}_k\|_{\mathbf{H}}^2|\mathcal{F}_{k-1}) &\leqslant \frac{L_\ell h_k}{8\alpha}\|\nabla F(\tilde{\mathbf{y}}_{k-1})\|_{\mathbf{H}^{-1}}^2, \quad \mathbb{E}(\mathbf{v}_k|\mathcal{F}_{k-1}) = 0,\\
\mathbb{E}(\|\mathbf{v}_k\|_{\mathbf{H}}^2|\mathcal{F}_{k-1}) &\leqslant \frac{320L_\ell\left(3L_\ell \operatorname{tr}(\mathbf{H}^{-1}\mathbf{Q}) + 8\eta\tilde{\kappa}\operatorname{tr}\mathbf{Q}\right)}{T}\\
&\quad + \frac{160L_\ell^2(6L + \tilde{\kappa}(7+16\eta B))(F(\tilde{\mathbf{y}}) - F(\mathbf{x}^*))}{T}\\
&\overset{\text{def}}{=} L_\ell\sigma^2 + L_{\text{eff}}L_\ell^2(F(\tilde{\mathbf{y}}_{k-1}) - F(\mathbf{x}^*)),
\end{aligned}$$

where

$$\sigma^2 = \frac{320\left(3L_\ell \operatorname{tr}(\mathbf{H}^{-1}\mathbf{Q}) + 8\eta\tilde{\kappa}\operatorname{tr}\mathbf{Q}\right)}{T}, \quad L_{\text{eff}} = \frac{160(6L + \tilde{\kappa}(7+16\eta B))(F(\tilde{\mathbf{y}}) - F(\mathbf{x}^*))}{T}.$$

**Lemma E.1.** *Assume we have*

$$h_k \leqslant \min\left\{\frac{1}{6L_\ell}, \frac{T^2}{288\alpha L_\ell L_{\text{eff}}^2}\right\}, \quad \theta_k = \sqrt{\frac{L_\ell h_k}{2\alpha}}, \quad \beta_k = \frac{1-\theta_k}{1+\theta_k}.$$

*Let $\tilde{\mathbf{z}}_k = \tilde{\mathbf{y}}_k + (\tilde{\mathbf{y}}_k - \tilde{\mathbf{x}}_k)/\theta_k$, then we have*

$$\begin{aligned}
&\mathbb{E}_{k-1}F(\tilde{\mathbf{x}}_k) - F(\mathbf{x}^*) + \frac{2\theta_k^2}{3L_\ell h_k}\mathbb{E}_{k-1}\|\tilde{\mathbf{z}}_k - \mathbf{x}^*\|_{\mathbf{H}}^2\\
&\leqslant \left(1 - \frac{\theta_k}{2}\right)\left(F(\tilde{\mathbf{x}}_{k-1}) - F(\mathbf{x}^*) + \frac{2\theta_k^2}{3L_\ell h_k}\|\tilde{\mathbf{z}}_{k-1} - \mathbf{x}^*\|_{\mathbf{H}}^2\right) + \frac{3}{2}h_k\sigma^2.
\end{aligned}$$

*Proof.* Since $\nabla^2 F \preceq \mathbf{H}$, we have

$$F(\tilde{\mathbf{x}}_k) \leqslant F(\tilde{\mathbf{y}}_{k-1}) + \langle \nabla F(\tilde{\mathbf{y}}_{k-1}), \tilde{\mathbf{x}}_k - \tilde{\mathbf{y}}_{k-1}\rangle + \frac{1}{2}\|\tilde{\mathbf{x}}_k - \tilde{\mathbf{y}}_{k-1}\|_{\mathbf{H}}^2.$$

Take expectation conditioned on $\mathcal{F}_{k-1}$ and use (32) twice, we have

$$\begin{aligned}
&\mathbb{E}_{k-1}\langle \nabla F(\tilde{\mathbf{y}}_{k-1}), \tilde{\mathbf{x}}_k - \tilde{\mathbf{y}}_{k-1}\rangle \\
=& -\frac{1}{L_\ell h_k}\mathbb{E}_{k-1}\langle \mathbf{H}(\tilde{\mathbf{x}}_k - \tilde{\mathbf{y}}_{k-1}), \tilde{\mathbf{x}}_k - \tilde{\mathbf{y}}_{k-1}\rangle + \frac{1}{L_\ell}\mathbb{E}_{k-1}\langle \mathbf{H}(\mathbf{r}_k + \mathbf{v}_k), \tilde{\mathbf{x}}_k - \tilde{\mathbf{y}}_{k-1}\rangle \\
=& -\frac{1}{L_\ell h_k}\mathbb{E}_{k-1}\|\tilde{\mathbf{x}}_k - \tilde{\mathbf{y}}_{k-1}\|_{\mathbf{H}}^2 - h_k\mathbb{E}_{k-1}\langle \nabla F(\tilde{\mathbf{y}}_{k-1}), \mathbf{r}_k\rangle + \frac{h_k}{L_\ell}\mathbb{E}_{k-1}\|\mathbf{r}_k + \mathbf{v}_k\|_{\mathbf{H}}^2.
\end{aligned} \tag{33}$$

Note that we also have

$$\mathbb{E}_{k-1}\langle \nabla F(\tilde{\mathbf{y}}_{k-1}), \tilde{\mathbf{x}}_k - \tilde{\mathbf{y}}_{k-1}\rangle = -L_\ell h_k \|\nabla F(\tilde{\mathbf{y}}_{k-1})\|_{\mathbf{H}^{-1}}^2 + h_k\langle \nabla F(\tilde{\mathbf{y}}_{k-1}), \mathbf{r}_k\rangle \tag{34}$$

Therefore, replace $\langle \nabla F(\tilde{\mathbf{y}}_{k-1}), \tilde{\mathbf{x}}_k - \tilde{\mathbf{y}}_{k-1}\rangle$ by $(3/4)\times(33) + (1/4)\times(34)$ to obtain

$$\begin{aligned}
\mathbb{E}_{k-1}F(\tilde{\mathbf{x}}_k) \leqslant& F(\tilde{\mathbf{y}}_{k-1}) - \frac{3 - 2L_\ell h_k}{4L_\ell h_k}\|\tilde{\mathbf{x}}_k - \tilde{\mathbf{y}}_{k-1}\|_{\mathbf{H}}^2 + \frac{3h_k}{2L_\ell}\left(\mathbb{E}_{k-1}\|\mathbf{r}_k\|_{\mathbf{H}}^2 + \mathbb{E}_{k-1}\|\mathbf{v}_k\|_{\mathbf{H}}^2\right) \\
&-\frac{h_k}{2}\mathbb{E}_{k-1}\langle \nabla F(\tilde{\mathbf{y}}_{k-1}), \mathbf{r}_k\rangle - \frac{L_\ell h_k}{4}\|\nabla F(\tilde{\mathbf{y}}_{k-1})\|_{\mathbf{H}^{-1}}^2. \\
\overset{a}{\leqslant}& F(\tilde{\mathbf{y}}_{k-1}) - \frac{3 - 2L_\ell h_k}{4L_\ell h_k}\|\tilde{\mathbf{x}}_k - \tilde{\mathbf{y}}_{k-1}\|_{\mathbf{H}}^2 - \frac{L_\ell h_k}{8}\|\nabla F(\tilde{\mathbf{y}}_{k-1})\|_{\mathbf{H}^{-1}}^2 \\
&+\frac{3h_k\sigma^2}{2} + \frac{3L_{\text{eff}}L_\ell h_k(F(\tilde{\mathbf{y}}_{k-1}) - F(\mathbf{x}^*))}{2},
\end{aligned}$$

where $\overset{a}{\leqslant}$ uses $\mathbb{E}_{k-1}\langle \nabla F(\tilde{\mathbf{y}}_{k-1}), \mathbf{r}_k\rangle \leqslant L_\ell\|\nabla F(\tilde{\mathbf{y}}_{k-1})\|_{\mathbf{H}^{-1}}^2/4$. Thus,

$$\begin{aligned}
\mathbb{E}_{k-1}F(\tilde{\mathbf{x}}_k) - F(\mathbf{x}^*) \leqslant& \left(1 + \frac{3L_{\text{eff}}L_\ell h_k}{2}\right)(F(\tilde{\mathbf{y}}_{k-1}) - F(\mathbf{x}^*)) + \frac{3h_k\sigma^2}{2} \\
&-\frac{3 - 2L_\ell h_k}{4L_\ell h_k}\|\tilde{\mathbf{x}}_k - \tilde{\mathbf{y}}_{k-1}\|_{\mathbf{H}}^2 - \frac{L_\ell h_k}{8}\|\nabla F(\tilde{\mathbf{y}}_{k-1})\|_{\mathbf{H}^{-1}}^2.
\end{aligned} \tag{35}$$

Let $\beta_k = (1-\theta_k)/(1+\theta_k)$, $\tilde{\mathbf{z}}_k = \tilde{\mathbf{y}}_k + (\tilde{\mathbf{y}}_k - \tilde{\mathbf{x}}_k)/\theta_k$, $\tilde{\mathbf{w}}_k = \tilde{\mathbf{y}}_k + (1-\theta_k)(\tilde{\mathbf{y}}_k - \tilde{\mathbf{x}}_k)$, where $\theta_k$ will be chosen later. By Assumption 1, we have $\nabla^2 F \geqslant \mathbf{H}/\alpha$, so

$$F(\tilde{\mathbf{y}}_{k-1}) \leqslant F(\mathbf{x}^*) + \langle \nabla F(\tilde{\mathbf{y}}_{k-1}), \tilde{\mathbf{y}}_{k-1} - \mathbf{x}^*\rangle - \frac{1}{2\alpha}\|\tilde{\mathbf{y}}_{k-1} - \mathbf{x}^*\|_{\mathbf{H}}^2, \tag{36}$$

$$F(\tilde{\mathbf{y}}_{k-1}) \leqslant F(\tilde{\mathbf{x}}_{k-1}) + \langle \nabla F(\tilde{\mathbf{y}}_{k-1}), \tilde{\mathbf{y}}_{k-1} - \tilde{\mathbf{x}}_{k-1}\rangle - \frac{1}{2\alpha}\|\tilde{\mathbf{y}}_{k-1} - \tilde{\mathbf{x}}_{k-1}\|_{\mathbf{H}}^2. \tag{37}$$

Let $\tilde{\mathbf{x}} = (1-\theta_k)\tilde{\mathbf{x}}_{k-1} + \theta_k\mathbf{x}^*$, then $\theta_k \times (36) + (1-\theta_k)\times(37)$ yields

$$\begin{aligned}
F(\tilde{\mathbf{y}}_{k-1}) - F(\mathbf{x}^*) \leqslant& (1-\theta_k)(F(\tilde{\mathbf{x}}_{k-1}) - F(\mathbf{x}^*)) + \langle \nabla F(\tilde{\mathbf{y}}_{k-1}), \tilde{\mathbf{y}}_{k-1} - \tilde{\mathbf{x}}\rangle \\
&-\frac{\theta_k}{2\alpha}\|\tilde{\mathbf{y}}_{k-1} - \mathbf{x}^*\|_{\mathbf{H}}^2 - \frac{1-\theta_k}{2\alpha}\|\tilde{\mathbf{y}}_{k-1} - \tilde{\mathbf{x}}_{k-1}\|_{\mathbf{H}}^2 \\
\overset{a}{\leqslant}& (1-\theta_k)(F(\tilde{\mathbf{x}}_{k-1}) - F(\mathbf{x}^*)) + \langle \nabla F(\tilde{\mathbf{y}}_{k-1}), \tilde{\mathbf{y}}_{k-1} - \tilde{\mathbf{x}}\rangle \\
&-\frac{\theta_k}{4\alpha}\|\tilde{\mathbf{y}}_{k-1} - \mathbf{x}^*\|_{\mathbf{H}}^2 - \frac{1}{4\alpha}\|\tilde{\mathbf{y}}_{k-1} - \tilde{\mathbf{x}}\|_{\mathbf{H}}^2,
\end{aligned}$$

where $\overset{a}{\leqslant}$ uses the convexity of $\|\cdot\|_{\mathbf{H}}^2$ as follows

$$\theta_k\|\tilde{\mathbf{y}}_{k-1} - \mathbf{x}^*\|_{\mathbf{H}}^2 + (1-\theta_k)\|\tilde{\mathbf{y}}_{k-1} - \tilde{\mathbf{x}}_{k-1}\|_{\mathbf{H}}^2 \geqslant \|\tilde{\mathbf{y}}_{k-1} - \tilde{\mathbf{x}}\|_{\mathbf{H}}^2.$$

Note that $\nabla F(\tilde{\mathbf{y}}_{k-1}) = \mathbf{H}\mathbb{E}_{k-1}(\tilde{\mathbf{y}}_{k-1} - \tilde{\mathbf{x}}_k)/(L_\ell h_k) + \mathbf{H}\mathbb{E}_{k-1}\mathbf{r}_k/L_\ell$. Therefore,

$$\begin{aligned}
&\langle \nabla F(\tilde{\mathbf{y}}_{k-1}), \tilde{\mathbf{y}}_{k-1} - \tilde{\mathbf{x}}\rangle = \frac{1}{L_\ell h_k}\mathbb{E}_{k-1}\langle \mathbf{H}(\tilde{\mathbf{y}}_{k-1} - \tilde{\mathbf{x}}_k), \tilde{\mathbf{y}}_{k-1} - \tilde{\mathbf{x}}\rangle + \frac{1}{L_\ell}\mathbb{E}_{k-1}\langle \mathbf{H}\mathbf{r}_k, \tilde{\mathbf{y}}_{k-1} - \tilde{\mathbf{x}}\rangle \\
&= \frac{\theta_k^2}{L_\ell h_k}\mathbb{E}_{k-1}\langle \mathbf{H}(\tilde{\mathbf{w}}_{k-1} - \tilde{\mathbf{z}}_k), \tilde{\mathbf{w}}_{k-1} - \mathbf{x}^*\rangle + \frac{1}{L_\ell}\mathbb{E}_{k-1}\langle \mathbf{H}\mathbf{r}_k, \tilde{\mathbf{y}}_{k-1} - \tilde{\mathbf{x}}\rangle \\
&\leqslant \frac{\theta_k^2}{2L_\ell h_k}\left(\|\tilde{\mathbf{w}}_{k-1} - \mathbf{x}^*\|_{\mathbf{H}}^2 + \mathbb{E}_{k-1}\|\tilde{\mathbf{z}}_k - \tilde{\mathbf{w}}_{k-1}\|_{\mathbf{H}}^2 - \mathbb{E}_{k-1}\|\tilde{\mathbf{z}}_k - \mathbf{x}^*\|_{\mathbf{H}}^2\right) \\
&\quad + \frac{\alpha}{L_\ell^2}\mathbb{E}_{k-1}\|\mathbf{r}_k\|_{\mathbf{H}}^2 + \frac{1}{4\alpha}\|\tilde{\mathbf{y}}_{k-1} - \tilde{\mathbf{x}}\|_{\mathbf{H}}^2.
\end{aligned}$$

Thus, by applying $\alpha\mathbb{E}_{k-1}\|\mathbf{r}_k\|_{\mathbf{H}}^2 \leqslant L_\ell^3 h_k \|\nabla F(\tilde{\mathbf{y}}_{k-1})\|_{\mathbf{H}^{-1}}^2/4$, we have

$$\begin{aligned}
&F(\tilde{\mathbf{y}}_{k-1}) - F(\mathbf{x}^*) \\
&\leqslant (1-\theta_k)(F(\tilde{\mathbf{x}}_{k-1}) - F(\mathbf{x}^*)) + \frac{\alpha}{L_\ell^2}\mathbb{E}_{k-1}\|\mathbf{r}_k\|_{\mathbf{H}}^2 - \frac{\theta_k}{4\alpha}\|\tilde{\mathbf{y}}_{k-1} - \mathbf{x}^*\|_{\mathbf{H}}^2 \\
&\quad + \frac{\theta_k^2}{2L_\ell h_k}\left(\|\tilde{\mathbf{w}}_{k-1} - \mathbf{x}^*\|_{\mathbf{H}}^2 + \mathbb{E}_{k-1}\|\tilde{\mathbf{z}}_k - \tilde{\mathbf{w}}_{k-1}\|_{\mathbf{H}}^2 - \mathbb{E}_{k-1}\|\tilde{\mathbf{z}}_k - \mathbf{x}^*\|_{\mathbf{H}}^2\right) \\
&\leqslant (1-\theta_k)(F(\tilde{\mathbf{x}}_{k-1}) - F(\mathbf{x}^*)) + \frac{\alpha}{L_\ell^2}\mathbb{E}_{k-1}\|\mathbf{r}_k\|_{\mathbf{H}}^2 - \left(\frac{\theta_k}{4\alpha} - \frac{\theta_k^3}{2h_k}\right)\|\tilde{\mathbf{y}}_{k-1} - \mathbf{x}^*\|_{\mathbf{H}}^2 \\
&\quad + \frac{\theta_k^2}{2L_\ell h_k}\left((1-\theta_k)\|\tilde{\mathbf{z}}_{k-1} - \mathbf{x}^*\|_{\mathbf{H}}^2 + \mathbb{E}_{k-1}\|\tilde{\mathbf{z}}_k - \tilde{\mathbf{w}}_{k-1}\|_{\mathbf{H}}^2 - \mathbb{E}_{k-1}\|\tilde{\mathbf{z}}_k - \mathbf{x}^*\|_{\mathbf{H}}^2\right).
\end{aligned} \tag{38}$$

Finally, combine above bound and (35), and use $3L_{\text{eff}}L_\ell h_k \leqslant \theta_k/4 \leqslant 1/3$ and Lemma D.1, we have

$$\begin{aligned}
\mathbb{E}_{k-1}F(\tilde{\mathbf{x}}_k) - F(\mathbf{x}^*) &\leqslant (1-\theta_k/2)(F(\tilde{\mathbf{x}}_k) - F(\mathbf{x}^*)) + \frac{3}{2}h_k\sigma^2 + \frac{4\alpha}{3}\mathbb{E}_{k-1}\|\mathbf{r}_k\|_{\mathbf{H}}^2 \\
&\quad + \frac{2\theta_k^2}{3L_\ell h_k}\left((1-\theta_k)\|\tilde{\mathbf{z}}_{k-1} - \mathbf{x}^*\|_{\mathbf{H}}^2 - \mathbb{E}_{k-1}\|\tilde{\mathbf{z}}_k - \mathbf{x}^*\|_{\mathbf{H}}^2\right) \\
&\quad - \frac{1 - 6L_\ell h_k}{12L_\ell h_k}\mathbb{E}_{k-1}\|\tilde{\mathbf{x}}_k - \tilde{\mathbf{y}}_{k-1}\|_{\mathbf{H}^{-1}}^2 - \frac{h_k}{8}\|\nabla F(\tilde{\mathbf{y}}_{k-1})\|_{\mathbf{H}^{-1}}^2.
\end{aligned}$$

We complete the proof by arranging the terms. □

For convenience, we analyze the following recursion,

$$L_k \leqslant (1 - a\theta_k)L_{k-1} + c\theta_k^2,$$

where $\theta_k \leqslant 1/d$. The following lemma solves the recursion.

**Lemma E.2.** *Suppose we choose*

$$\theta_k = \begin{cases} \theta_{\max}, & k \leqslant K/2, \\ \dfrac{2}{a(2/(a\theta_{\max}) + k - K/2)}, & k > K/2. \end{cases}$$

*Then we have*

$$L_K \leqslant (1 - a\theta_{\max})^{K/2} L_0 + \frac{12c}{a^2 K}.$$

*Proof.* We first consider the case $K \leqslant 4d/a$. Note that we have

$$L_k \leqslant (1 - a\theta_k)L_{k-1} + c\theta_k\theta_{\max}.$$

By rearranging the terms, we have

$$L_k - \frac{c\theta_{\max}}{a} \leqslant (1 - a\theta_{\max})\left(L_{k-1} - \frac{c}{a}\right).$$

Thus, we solve the recursion,

$$L_K \leqslant (1 - a\theta_{\max})^K L_0 + \frac{c\theta_{\max}}{a} \leqslant (1 - a\theta_{\max})^K L_0 + \frac{4c}{a^2 K}.$$

For the case $K > 4/(a\theta_{\max})$, we have the following initial condition:

$$L_{K/2} \leqslant (1 - a\theta_{\max})^{K/2} L_0 + \frac{c\theta_{\max}}{a}.$$

By the choice of $\theta$, we have for $t > K/2$,

$$L_k \leqslant \frac{2/(a\theta_{\max}) + k - K/2 - 2}{2/(a\theta_{\max}) + k - K/2} L_{k-1} + \frac{4c}{a^2(2/(a\theta_{\max}) + k - K/2)^2}.$$

Multiply both side by $(2/(a\theta_{\max}) + k - K/2)^2$, we have

$$(2/(a\theta_{\max}) + k - K/2)^2 L_k \leqslant (2/(a\theta_{\max}) + k - K/2 - 1)^2 L_{k-1} + \frac{4c}{a^2}.$$

We solve the recursion,

$$\begin{aligned} L_K &\leqslant \frac{16}{a^2\theta_{\max}^2 K^2} L_{K/2} + \frac{8c}{a^2 K} \leqslant \frac{16}{a^2\theta_{\max} K^2} \left[(1 - a\theta_{\max})^{K/2} L_0 + \frac{c\theta_{\max}}{a}\right] + \frac{8c}{a^2 K}, \\ &\leqslant (1 - a\theta_{\max})^{K/2} L_0 + \frac{16c}{a^3\theta_{\max} K^2} + \frac{8c}{a^2 K} \leqslant (1 - a\theta_{\max})^{K/2} L_0 + \frac{12c}{a^2 K}. \end{aligned}$$

We complete the proof by combining the two cases. □

### E.2 Proof of Theorem 5.1

**Theorem E.1** (Full version of Theorem 5.1)**.** *Suppose we choose the hyperparameters as specified in Appendix C. Let $\tilde{\mathbf{x}}_K$ denote the output of Algorithm 1 after $K$ outer iterations; each consisting of $T \geqslant \tilde{\Omega}(\sqrt{\tilde{\kappa}/(\mu\eta)})$ inner updates. Then we have*

$$\begin{aligned} \mathbb{E}F(\tilde{\mathbf{x}}_K) - F(\mathbf{x}^*) \leqslant \frac{5}{3} \exp\left(-\frac{K}{8\sqrt{3\alpha} + 24\sqrt{2}\alpha L_{\text{eff}}/T}\right) \left(F(\tilde{\mathbf{x}}_0) - F(\mathbf{x}^*)\right) \\ + \frac{46080\alpha \operatorname{tr}(\mathbf{H}^{-1}\mathbf{Q})}{KT} + \frac{122880\alpha\eta\tilde{\kappa} \operatorname{tr}\mathbf{Q}}{L_\ell KT}. \end{aligned}$$

*Remark* 6. Suppose Assumptions 1, 2, and 3 holds, then we have $L_{\text{eff}} \lesssim \alpha\tilde{\kappa}$. Therefore, Theorem E.1 implies Theorem 5.1.

*Proof.* We first apply Lemma E.2. Let

$$a = \frac{1}{2}, \quad c = \frac{\alpha\sigma^2}{L_\ell}, \quad \frac{1}{\theta_{\max}} = \min\left\{\sqrt{\frac{1}{12\alpha}}, \frac{L_\ell T}{12\sqrt{2}\alpha L_{\text{eff}}}\right\},$$

then we have

$$\mathbb{E}F(\tilde{\mathbf{x}}_K) - F(\mathbf{x}^*) \leqslant \frac{5}{3} \exp\left(-\frac{K}{8\sqrt{3\alpha} + 24\sqrt{2}\alpha L_{\text{eff}}/T}\right) \left(\mathbb{E}F(\tilde{\mathbf{x}}_K) - F(\mathbf{x}^*)\right) + \frac{48\alpha\sigma^2}{L_\ell K}.$$

Note that by Lemma D.1, we have

$$\sigma^2 = \frac{960 L_\ell \operatorname{tr}(\mathbf{H}^{-1}\mathbf{Q}) + 2560\eta\tilde{\kappa} \operatorname{tr}\mathbf{Q}}{T}.$$

This completes the proof. □

### E.3 Proof of Corollary 5.1

*Proof of Corollary 5.1.* Note that when

$$n \geqslant \tilde{\mathcal{O}}\left(\left(\sqrt{\alpha\kappa\tilde{\kappa}} + \alpha^2\tilde{\kappa}\right) + \frac{\alpha \operatorname{tr}(\mathbf{H}^{-1}\mathbf{Q})}{\varepsilon} + \left(\frac{\alpha^2\tilde{\kappa}^2 \operatorname{tr}\mathbf{Q}}{L_\ell \mu \varepsilon}\right)^{1/3}\right).$$

By Theorem 5.1, the optimization error and statistical error is smaller that $\mathcal{O}(\varepsilon)$,

$$\mathbb{E}F(\tilde{\mathbf{x}}_K) - F(\mathbf{x}^*) \leqslant \mathcal{O}(\varepsilon) + \frac{\alpha\eta\tilde{\kappa}\operatorname{tr}\mathbf{Q}}{L_\ell n}.$$

Since we have

$$n \geqslant \tilde{\Omega}\left(\left(\frac{\alpha^2\tilde{\kappa}^2 \operatorname{tr}\mathbf{Q}}{L_\ell \mu \varepsilon}\right)^{1/3}\right) \implies \tilde{\Theta}\left(\frac{\alpha\tilde{\kappa}}{\mu n^2}\right) \leqslant \tilde{\Theta}\left(\frac{n\varepsilon L_\ell}{\alpha\tilde{\kappa}\mathbf{Q}}\right),$$

$$n \geqslant \tilde{\Omega}(\sqrt{\alpha\kappa\tilde{\kappa}}) \implies \tilde{\Theta}\left(\frac{\alpha\tilde{\kappa}}{\mu n^2}\right) \leqslant \frac{1}{R^2}.$$

Thus, the choice of $\eta$ is feasible. Finally, note that the choice of $\eta \leqslant \tilde{\Theta}\left(\frac{n\varepsilon L_\ell}{\alpha\tilde{\kappa}\operatorname{tr}\mathbf{Q}}\right)$ implies that $\frac{\alpha\eta\tilde{\kappa}\operatorname{tr}\mathbf{Q}}{L_\ell n} \lesssim \varepsilon$. This completes the proof. □

## F Excess Risk of ERM

In this section, we present the excess risk bound of ERM for the generalized linear prediction by applying the local Rademacher technique. We consider ERM of the following form:

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{N}\sum_{i=1}^{N} \ell(\mathbf{a}_i^\top \mathbf{x}, b_i),$$

where $\{(\mathbf{a}_i, b_i)\}_{i=1}^N$ are i.i.d. random variables from the distribution $\mathcal{D}$. We make the following assumptions.

**Assumption 6.** *We make the following assumptions on the loss function $\ell$ and data distribution $\mathcal{D}$:*

1. *The loss function $\ell(z, y)$ is $L_0$-Lipschitz with respect to the first argument, i.e., for all $z_1, z_2 \in [-1, 1]$,*
$$\|\ell(z_1, y) - \ell(z_2, y)\| \leqslant L_0\|z_1 - z_2\| \quad \mathcal{D}\text{-a.s.}$$
2. *The loss function $\ell(\cdot, \cdot)$ is $\mu_\ell$-strongly convex with respect to the first argument.*
3. *The data distribution $\mathcal{D}$ satisfies that $\mathbf{\Sigma}^{-1/2}\mathbf{a}$ is $\sigma_a^2$-sub-Gaussian.*

Let the function class

$$\mathcal{F}_\delta = \left\{\mathbf{a} \mapsto \frac{\mathbf{a}^\top\mathbf{x}}{AX}\mathbb{I}_{\{\|\mathbf{a}^\top\mathbf{x}\|\leqslant AX\}} : \frac{1}{AX}\sqrt{\mathbb{E}_{a\sim\mathcal{D}}(\mathbf{a}^\top\mathbf{x})^2\mathbb{I}_{\{\|\mathbf{a}^\top\mathbf{x}\|\leqslant AX\}}} \leqslant \delta\right\},$$

where we choose

$$A = C\sigma_a d\sqrt{\log\frac{1}{\delta}}, \quad X = 2\|\mathbf{x}^*\|_{\mathbf{\Sigma}},$$

and $C$ is a sufficiently large constant such that $\mathbb{E}_{\mathbf{a}\sim\mathcal{D}}\mathbf{a}\mathbf{a}^\top\mathbb{I}_{\|\mathbf{\Sigma}^{-1/2}\mathbf{a}\|\leqslant A} \succeq \mathbf{\Sigma}/2$, and $\|\mathbf{\Sigma}^{-1/2}\mathbf{a}_i\| \leqslant A$ for all $i \in [N]$ with probability at lest $1 - N\delta$.

The following theorem bounds the excess risk of ERM.

**Theorem F.1** (Excess risk bound of ERM)**.** *Suppose Assumption 6 holds. Let $N \geqslant 441L_0^2 d\log(1/\delta)/\mu_\ell^2$ and $\sigma_0^2 = L_0^2 d/L_\ell$ Then with probability at least $1 - (N+1)\delta$, we have*

$$L(\widehat{\mathbf{x}}) - L(\mathbf{x}^*) \leqslant \frac{\alpha\sigma_0^2}{\mu_\ell N}$$

Furthermore, we show that $\sigma_0$ is lower bounded by $\|\mathbf{H}^{-1/2}\mathbf{Q}\mathbf{H}^{-1/2}\|d \geqslant \mathrm{tr}(\mathbf{H}^{-1}\mathbf{Q})$.

**Lemma F.1.** *Suppose Assumption 6 holds, then we have* $\sigma_0^2 \geqslant \|\mathbf{H}^{-1/2}\mathbf{Q}\mathbf{H}^{-1/2}\|d \geqslant \mathrm{tr}(\mathbf{H}^{-1}\mathbf{Q})$.

*Proof.* Since $\|\ell(\mathbf{a}^\top\mathbf{x}^*)\| \leqslant L_0$, we have

$$\mathbf{Q} = \mathbb{E}_{\mathbf{a},\ell\sim\mathcal{D}}\left(\ell'(\mathbf{a}^\top\mathbf{x}^*)\right)^2\mathbf{a}\mathbf{a}^\top \preceq L_0\mathbf{\Sigma} = \frac{L_0\mathbf{H}}{L_\ell}.$$

Therefore, we have $L_0 \geqslant L_\ell\|\mathbf{H}^{-1/2}\mathbf{Q}\mathbf{H}^{-1/2}\|$, which is equivalent to $\sigma_0^2 \geqslant \|\mathbf{H}^{-1/2}\mathbf{Q}\mathbf{H}^{-1/2}\|d \geqslant \mathrm{tr}(\mathbf{H}^{-1}\mathbf{Q})$. □

### F.1 Proof of Theorem F.1

The following lemma bounds the Rademacher complexity of $\mathcal{F}_\delta$.

**Lemma F.2.** *Suppose Assumption 6 holds, then we have*

$$\mathcal{R}_N(\mathcal{F}_\delta) \leqslant 2\delta\sqrt{\frac{d}{N}}.$$

*Proof.* Note that $\mathcal{F}_\delta \subset \bar{\mathcal{F}}_\delta$, where

$$\bar{\mathcal{F}}_\delta = \left\{\mathbf{a} \mapsto \frac{\mathbf{a}^\top\mathbf{x}}{AX}\mathbb{I}_{\{\|\mathbf{a}^\top\mathbf{x}\|\leqslant AX\}} : \frac{\|\mathbf{x}-\mathbf{x}^*\|_{\mathbf{\Sigma}}}{2AX} \leqslant \delta\right\}$$

By definition, we have

$$\begin{aligned}
&\mathcal{R}_N(\mathcal{F}_\delta) \leqslant \mathcal{R}_N(\bar{\mathcal{F}}_\delta)\\
&= \mathbb{E}_{\mathbf{a},\boldsymbol{\varepsilon}}\sup_{\|\mathbf{x}-\mathbf{x}^*\|_{\mathbf{\Sigma}}\leqslant 2AX\delta}\frac{1}{N}\sum_{i=1}^N\frac{\varepsilon_i\mathbf{a}_i^\top\mathbf{x}}{AX}\mathbb{I}_{\{|\mathbf{a}_i^\top\mathbf{x}|\leqslant AX\}} \leqslant \frac{1}{AX}\mathbb{E}_{\mathbf{a},\boldsymbol{\varepsilon}}\sup_{\|\mathbf{h}\|_{\mathbf{\Sigma}}\leqslant 2AX\delta}\frac{1}{N}\sum_{i=1}^N\varepsilon_i\mathbf{a}_i^\top\mathbf{h}\mathbb{I}_{\{\|\mathbf{h}\|\leqslant AX\}}\\
&\leqslant \frac{\delta}{AX}\sqrt{\mathbb{E}_{\mathbf{a},\boldsymbol{\varepsilon}}\left\|\frac{1}{N}\sum_{i=1}^N\varepsilon_i\mathbf{a}_i\right\|^2} \leqslant 2\delta\sqrt{\frac{d}{N}}.
\end{aligned}$$

This completes the proof. □

*Proof of Theorem F.1.* We consider a constrained ERM as follows:

$$\tilde{\mathbf{x}} = \operatorname*{arg\,min}_{\|\mathbf{x}-\mathbf{x}^*\|_{\mathbf{\Sigma}}\leqslant 1}\frac{1}{N}\sum_{i=1}^N\ell_i(\mathbf{a}_i^\top\mathbf{x}).$$

Let $\tilde{\ell}(z) = \ell(AXz)$, so $\tilde{\ell}$ is $L_0AX$-Lipschitz and $\mu_\ell A^2X^2$-strongly convex, and the objective can be written as:

$$\min_{\|\mathbf{x}-\mathbf{x}^*\|_{\mathbf{\Sigma}}\leqslant X}\frac{1}{N}\sum_{i=1}^N\tilde{\ell}_i(\mathbf{a}_i^\top\mathbf{x}/(AX))$$

Let $\delta_n = 2\sqrt{\frac{d}{N}\log\frac{1}{\delta}}$, then $\delta_N^2 \geqslant \mathcal{R}_N(\mathcal{F}_{\delta_N})$. We apply the localization technique (Theorem 14.20 in Wainwright (2019)) to obtain that

$$\mathbb{E}\left(\frac{(\mathbf{a}^\top(\tilde{\mathbf{x}}-\mathbf{x}^*))^2\mathbb{I}_{\{\|\mathbf{a}^\top\mathbf{x}\|\leqslant AX\}}}{A^2X^2}\middle|\mathbf{a}_1,\ldots,\mathbf{a}_N\right) \leqslant \frac{441L_0^2d}{\mu_\ell^2A^2X^2N}\log\frac{1}{\delta},$$

$$L(\tilde{\mathbf{x}}) - L(\mathbf{x}^*) \leqslant \frac{220L_0^2d}{\mu_\ell N}\log\frac{1}{\delta},$$

with probability at least $1-\delta$. Thus, we have $\|\tilde{\mathbf{x}}-\mathbf{x}^*\|_{\mathbf{\Sigma}} \leqslant 1$ by $N \geqslant 441L_0^2d\log(1/\delta)/\mu_\ell^2$. Therefore, we have $\widehat{\mathbf{x}} = \tilde{\mathbf{x}}$ by the strong convexity. □

# G WORST-CASE OPTIMALITY OF STATISTICAL TERM

In this section, we show that the statistical term in Theorem 5.1 cannot be further improved without additional assumptions on the smoothness of $\ell \sim \mathcal{D}$.

**Data Generating Process.** We consider the following statistical model:

$$\mathbf{a} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}), \quad b \sim \mathcal{N}(\mathbf{a}^\top \mathbf{x}^*, \sigma^2), \tag{39}$$

where $\mathbf{x}^*$ is the ground truth. We consider the following problem class.

**Definition G.1** (Problem Class)**.** The problem class $\mathcal{P}_n$ is defined as $\mathcal{P}_n = \{\mu_{\mathbf{x}^*}^{\otimes n} : \mathbf{x}^* \in \mathbb{R}^d\}$, where $\mu_{\mathbf{x}^*}$ denotes the distribution in (39), and $\mu_{\mathbf{x}^*}^{\otimes n}$ denotes the distribution of $\{(\mathbf{a}_i, b_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mu_{\mathbf{x}^*}$.

**Loss Function.** Let $\delta > 0$, we define the following function:

$$\ell_\delta(z) = \begin{cases} \mu_\ell z^2/2 + (L_\ell - \mu_\ell)\delta z - (L_\ell - \mu_\ell)\delta^2/2, & z < -\delta. \\ L_\ell z^2/2, & \|z\| \leqslant \delta, \\ \mu_\ell z^2/2 + (L_\ell - \mu_\ell)\delta z - (L_\ell - \mu_\ell)\delta^2/2, & z > \delta. \end{cases}$$

The population loss function is defined as

$$F_\delta(\mathbf{x}) = \mathbb{E}_{\mathbf{a}, b \sim \mathcal{D}}\, \ell_\delta(\mathbf{a}^\top \mathbf{x} - b).$$

Let $F_\delta^* = F_\delta(\mathbf{x}^*)$ We denote the minimax risk by

$$\mathcal{R}_n(\mathcal{P}_n) = \inf_{\widehat{\mathbf{x}}} \sup_{\mathcal{D}_n \in \mathcal{P}_n} \mathbb{E}_{\{(\mathbf{a}_i, b_i)\}_{i=1}^n \sim \mathcal{D}_n} F_\delta\left(\widehat{\mathbf{x}}(\{(\mathbf{a}_i, b_i)\}_{i=1}^n)\right) - F_\delta(\mathbf{x}^*(\mathcal{D})), \tag{40}$$

where $\inf$ is taken over all estimators $\widehat{\mathbf{x}} : (\mathbb{R}^d \times \mathbb{R})^n \to \mathbb{R}^d$.

The following theorem establishes the worst-case optimality of the statistical term.

**Theorem G.1.** *Consider the problem class $\mathcal{P}_n$ defined in Definition G.1. For sufficiently small $\delta > 0$, we have the following lower bound of the minimax risk $\mathcal{R}_n(\mathcal{P}_n)$ defined in* (40)*:*

$$\mathcal{R}_n(\mathcal{P}_n) \geqslant \frac{\alpha \operatorname{tr}(\mathbf{H}^{-1}\mathbf{Q}_\delta)}{2n}.$$

## G.1 PROOF OF THEOREM G.1

The following lemma bounds the minimax risk of $\mathcal{P}_n$.

**Lemma G.1.** *Consider the problem class $\mathcal{P}_n$, we have the following lower bound of the minimax risk $\mathcal{R}_n(\mathcal{P}_n)$ defined in* (40)*:*

$$\mathcal{R}_n(\mathcal{P}_n) \geqslant \frac{\mu_\ell d}{n} + \frac{\mu_\ell \sigma^2}{n} - F_\delta^*$$

*Proof.* We apply Theorem 1 in Mourtada (2022) to obtain:

$$\inf_{\widehat{\mathbf{x}}} \sup_{\mathcal{D}_n \in \mathcal{P}_n} \mathbb{E}_{\{(\mathbf{a}_i, b_i)\}_{i=1}^n \sim \mathcal{D}_n} \|\widehat{\mathbf{x}}(\{(\mathbf{a}_i, b_i)\}_{i=1}^n) - \mathbf{x}^*\|_{\mathbf{\Sigma}}^2 = \frac{1}{n}\mathbb{E} \operatorname{tr}\left(\mathbf{\Sigma}\widehat{\mathbf{\Sigma}}_n^{-1}\right),$$

where $\widehat{\mathbf{\Sigma}}_n = \frac{1}{n}\sum_{i=1}^n \mathbf{a}_i \mathbf{a}_i^\top$ is the sample covariance matrix. By the operator convexity of the matrix inverse, we have $\mathbb{E} \operatorname{tr}(\mathbf{\Sigma}\widehat{\mathbf{\Sigma}}_n^{-1}) \geqslant d$. Finally, note that $\ell_\delta$ is $\mu_\ell$-strongly convex, so we have

$$F_\delta(\mathbf{x}) \geqslant \mathbb{E}_{\mathbf{a}, b \sim \mathcal{D}} \frac{\mu_\ell}{2}(\mathbf{a}^\top \mathbf{x} - \mathbf{a}^\top \mathbf{x}^* + \varepsilon)^2 = \frac{\mu_\ell}{2}\|\mathbf{x} - \mathbf{x}^*\|_{\mathbf{\Sigma}}^2 + \frac{\mu_\ell \sigma^2}{2}.$$

Therefore, we have

$$\begin{aligned} \mathcal{R}_n(\mathcal{P}_n) &\geqslant \inf_{\widehat{\mathbf{x}}} \sup_{\mathcal{D}_n \in \mathcal{P}} \mathbb{E}_{\{(\mathbf{a}_i, b_i)\}_{i=1}^n \sim \mathcal{D}_n} \frac{\mu_\ell}{2} \|\widehat{\mathbf{x}}(\{(\mathbf{a}_i, b_i)\}_{i=1}^n) - \mathbf{x}^*(\mathcal{D})\|_{\mathbf{\Sigma}}^2 + \frac{\mu_\ell \sigma^2}{2} - F_\delta^* \\ &\geqslant \frac{\mu_\ell d}{n} + \frac{\mu_\ell \sigma^2}{n} - F_\delta^*, \end{aligned}$$

This completes the proof. □

**Algorithm 2** **S**tochastic **A**ccelerated **D**ata-Dependent **A**lgorithm for **Non**-**S**trongly **C**onvex Objectives (SADA-NonSC)

**Input:** Initialization $\tilde{\mathbf{x}}_0 = \tilde{\mathbf{x}}_{-1}$, regularization parameters $\{h_k\}_{k=1}^K$, step sizes $\eta$, $\gamma$, and momentum parameters $\{\beta_k\}_{k=1}^K$, $\theta$, target accuracy $\varepsilon$, radius $D$, $M$

**for** $k = 1, 2, \ldots, K$ **do**

  $\tilde{\mathbf{y}}_{k-1} \leftarrow \tilde{\mathbf{x}}_{k-1} + \beta_k(\tilde{\mathbf{x}}_{k-1} - \tilde{\mathbf{x}}_{k-2})$ ▷ Extraplotation

  $\mathbf{x}_0 \leftarrow \tilde{\mathbf{y}}_{k-1}$, $\mathbf{z}_0 \leftarrow \tilde{\mathbf{y}}_{k-1}$

  **for** $t = 1, 2, \ldots, T$ **do** ▷ Inner loop for solving subproblem (41)

    Sample fresh data $(\mathbf{a}_t, b_t) \sim \mathcal{D}$

    $\mathbf{y}_{t-1} \leftarrow \frac{1}{1+\theta}\mathbf{x}_{t-1} + \frac{\theta}{1+\theta}\mathbf{z}_{t-1}$

    $\widehat{\mathbf{g}}_t \leftarrow h_k \ell'(\mathbf{a}_t^\top \tilde{\mathbf{y}}_{k-1}, b_t)\mathbf{a}_t + \dfrac{h_k \varepsilon}{2M^2}\mathbf{a} + \left[\mathbf{a}_t^\top(\mathbf{y}_{t-1} - \tilde{\mathbf{y}}_{k-1})\right]\mathbf{a}_t + \dfrac{\varepsilon}{L_\ell D^2}(\mathbf{y}_{t-1} - \tilde{\mathbf{y}}_{k-1})$

    $\mathbf{x}_t \leftarrow \mathbf{y}_{t-1} - \eta\widehat{\mathbf{g}}_t$

    $\mathbf{z}_t \leftarrow \theta\mathbf{y}_{t-1} + (1-\theta)\mathbf{z}_{t-1} - \gamma\widehat{\mathbf{g}}_t$

  **end for**

  $\tilde{\mathbf{x}}_k \leftarrow \frac{2}{T}\sum_{t=T/2+1}^{T}\mathbf{x}_t$ ▷ Tail-averaging scheme

**end for**

**return** $\tilde{\mathbf{x}}_K$

Let $\mathbf{H} = L_\ell\mathbf{\Sigma}$, and $\mathbf{Q}_\delta = \mathbb{E}((\ell'(\mathbf{a}^\top\mathbf{x}^*))^2\mathbf{a}\mathbf{a}^\top)$ denote the second moment of the gradient covariance matrix; then the statistical term in Theorem 5.1 is

$$\frac{\alpha\,\mathrm{tr}(\mathbf{H}^{-1}\mathbf{Q}_\delta)}{n}.$$

In the following we prove Theorem G.1.

*Proof of Theorem G.1.* Let $\mathbf{Q} = \lim_{\delta\to 0}\mathbf{Q}_\delta = \mu_\ell^2\mathbf{\Sigma}$. Note that we have

$$\lim_{\delta\to 0} F_\delta^* = \lim_{\delta\to 0}\mathbb{E}\ell_\delta(\varepsilon) = \frac{\mu_\ell\sigma^2}{2},$$
$$\lim_{\delta\to 0}\mathbf{Q}_\delta = \lim_{\delta\to 0}\mathbb{E}(\ell'_\delta(\varepsilon))^2\mathbf{a}\mathbf{a}^\top = \mathbb{E}(\mu_\ell\varepsilon)^2\mathbf{a}\mathbf{a}^\top = \mu_\ell^2\mathbf{\Sigma}.$$

Combining the above results, we have

$$\lim_{\delta\to 0}\left\{\mathcal{R}_n(\mathcal{P}_n) - \frac{\alpha\,\mathrm{tr}(\mathbf{H}^{-1}\mathbf{Q}_\delta)}{2n}\right\} = \frac{\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{Q})}{\mu_\ell n} - \frac{\alpha\,\mathrm{tr}(\mathbf{H}^{-1}\mathbf{Q})}{2n} = \frac{\alpha\,\mathrm{tr}(\mathbf{H}^{-1}\mathbf{Q})}{2n} > 0.$$

Thus, there exists $\delta > 0$ such that

$$\inf_{\widehat{\mathbf{x}}}\sup_{\mathcal{D}\in\mathcal{P}} F_\delta(\widehat{\mathbf{x}}) - F_\delta(\mathbf{x}^*) - \frac{\alpha\,\mathrm{tr}(\mathbf{H}^{-1}\mathbf{Q}_\delta)}{2n} > 0.$$

Rearrange the terms to obtain the desired result. □

# H Extensions of SADA

## H.1 SADA for Non-strongly Convex Objectives

The algorithm SADA-WC is shown in Algorithm 2. The difference is that the inner loop solves the following subproblem:

$$\min_{\mathbf{x}\in\mathbb{R}^d}\ \mathbb{E}_{\mathbf{a},b\sim\mathcal{D}}h_k\left\langle\ell'(\mathbf{a}^\top\tilde{\mathbf{y}}_{k-1}, b)\mathbf{a} + \frac{\varepsilon}{2M^2}\mathbf{a}, \mathbf{x} - \tilde{\mathbf{y}}_{k-1}\right\rangle + \frac{1}{2}\|\mathbf{x} - \tilde{\mathbf{y}}_{k-1}\|^2_{\mathbf{\Sigma}+\frac{\varepsilon}{L_\ell D^2}\mathbf{I}}. \tag{41}$$

The proofs presented in Appendix D.2 and Appendix D.3 can be extended to this case with slight modifications to the proof of the noise bound. For the inner loop, we modify the definition of $\mathcal{M}$ and $\tilde{\mathcal{M}}$ in (13) to

$$\mathcal{M}_\varepsilon = \mathbb{E}\left(\mathbf{a}\mathbf{a}^\top + \frac{\varepsilon}{L_\ell D^2}\mathbf{I}\right)\otimes\left(\mathbf{a}\mathbf{a}^\top + \frac{\varepsilon}{L_\ell D^2}\mathbf{I}\right),\quad \tilde{\mathcal{M}}_\varepsilon = \left(\mathbf{\Sigma} + \frac{\varepsilon}{L_\ell D^2}\mathbf{I}\right)\otimes\left(\mathbf{\Sigma} + \frac{\varepsilon}{L_\ell D^2}\mathbf{I}\right).$$

**Algorithm 3** **S**tochastic **A**ccelerated **D**ata-Dependent **A**lgorithm with **U**nlabeled **D**ata (SADA-UD)

**Input:** Initialization $\tilde{\mathbf{x}}_0$, regularization parameters $\{h_k\}_{k=1}^K$, step sizes $\eta$, $\gamma$, and momentum parameters $\{\beta_k\}_{k=1}^K$, $\theta$, $\tilde{\mathbf{x}}_{-1} = \tilde{\mathbf{x}}_0$, $T_0 = \tilde{\Theta}(\sqrt{\tilde{\kappa}/(\mu\eta)})$

**for** $k = 1, 2, \ldots, K$ **do**

  $\tilde{\mathbf{y}}_{k-1} \leftarrow \tilde{\mathbf{x}}_{k-1} + \beta_k(\tilde{\mathbf{x}}_{k-1} - \tilde{\mathbf{x}}_{k-2})$ ▷ Extraplotation

  $\mathbf{x}_0 \leftarrow \tilde{\mathbf{y}}_{k-1}$, $\mathbf{z}_0 \leftarrow \tilde{\mathbf{y}}_{k-1}$

  **for** $t = 1, 2, \ldots, T$ **do** ▷ Inner loop for solving subproblem (4)

    $\mathbf{y}_{t-1} \leftarrow \frac{1}{1+\theta}\mathbf{x}_{t-1} + \frac{\theta}{1+\theta}\mathbf{z}_{t-1}$

    Sample fresh labeled data $(\mathbf{a}_t, b_t) \sim \mathcal{D}$

    **if** $t \leqslant T_0$ **then**

      Sample fresh $m$ unlabeled data $\{\mathbf{a}_{t,i}^{\text{ul}}\}_{i=1}^m \sim$ i.i.d.$\mathcal{D}$ ▷ Sample $m$ unlabeled data

      $$\widehat{\mathbf{g}}_t \leftarrow h_k\ell'(\mathbf{a}_t^\top \tilde{\mathbf{y}}_{k-1}, b_t)\mathbf{a}_t + \frac{1}{m+1}\left(\mathbf{a}_t\mathbf{a}_t^\top + \sum_{i=1}^m \mathbf{a}_{t,i}^{\text{ul}}(\mathbf{a}_{t,i}^{\text{ul}})^\top\right)(\mathbf{y}_{t-1} - \tilde{\mathbf{y}}_{k-1})$$

    **else**

      $$\widehat{\mathbf{g}}_t \leftarrow h_k\ell'(\mathbf{a}_t^\top \tilde{\mathbf{y}}_{k-1}, b_t)\mathbf{a}_t + \left[\mathbf{a}_t^\top(\mathbf{y}_{t-1} - \tilde{\mathbf{y}}_{k-1})\right]\mathbf{a}_t$$

    **end if**

    $\mathbf{x}_t \leftarrow \mathbf{y}_{t-1} - \eta\widehat{\mathbf{g}}_t$

    $\mathbf{z}_t \leftarrow \theta\mathbf{y}_{t-1} + (1-\theta)\mathbf{z}_{t-1} - \gamma\widehat{\mathbf{g}}_t$

  **end for**

  $\tilde{\mathbf{x}}_k \leftarrow \frac{2}{T}\sum_{t=T/2+1}^T \mathbf{x}_t$ ▷ Tail-averaging scheme

**end for**

**return** $\tilde{\mathbf{x}}_K$

Note that $\mathcal{M}_\varepsilon - \tilde{\mathcal{M}}_\varepsilon = \mathcal{M} - \tilde{\mathcal{M}}$, so the bounds in Appendix D hold by replacing the eigenvalue $\lambda_i$ by $\lambda_i + \frac{\varepsilon}{L_\ell D^2}$, hence the claimed bound.

### H.2 SADA WITH UNLABELED DATA

The algorithm SADA-UD is shown in Algorithm 3. The difference is that for each inner iteration, we sample a data pair $\mathbf{a}, b$ from $\mathcal{D}$ and an additional $m$ independent unlabeled data $\{\mathbf{a}_i^{\text{ul}}\}_{i=1}^m$ from the same distribution $\mathcal{D}$ without labels. Then, we use $\frac{1}{m+1}\left(\mathbf{a}\mathbf{a}^\top + \sum_{i=1}^m \mathbf{a}_i^{\text{ul}}(\mathbf{a}_i^{\text{ul}})^\top\right)$ as the estimator of $\mathbf{\Sigma}$.

We present a sketch of the analysis of Algorithm 3. Let

$$\widehat{\mathbf{\Sigma}} = \frac{1}{m+1}\left(\mathbf{a}\mathbf{a}^\top + \sum_{i=1}^m \mathbf{a}_i^{\text{ul}}(\mathbf{a}_i^{\text{ul}})^\top\right).$$

We extend Assumption 3 to the following lemma.

**Lemma H.1.** *Suppose Assumption 3 holds, then we have*

$$\mathbb{E}\widehat{\mathbf{\Sigma}}^2 \preceq R_m^2\mathbf{\Sigma}, \quad \mathbb{E}\widehat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1}\widehat{\mathbf{\Sigma}} \preceq \tilde{\kappa}_m\mathbf{\Sigma},$$

*where* $R_m^2 = \frac{R^2 + m\lambda_{\max}(\mathbf{\Sigma})}{m+1}$ *and* $\tilde{\kappa}_m = \frac{\tilde{\kappa}+m}{m+1}$.

*Remark* 7. If we set $\widehat{\mathbf{\Sigma}} = \mathbf{a}\mathbf{a}^\top$, then Lemma H.1 is identical to Assumption 3.

*Proof.* By definition, we have

$$\begin{aligned}
\mathbb{E}\widehat{\mathbf{\Sigma}}^2 &= \frac{1}{(m+1)^2}\mathbb{E}\left(\mathbf{a}\mathbf{a}^\top + \sum_{i=1}^m \mathbf{a}_i^{\text{ul}}(\mathbf{a}_i^{\text{ul}})^\top\right)\left(\mathbf{a}\mathbf{a}^\top + \sum_{i=1}^m \mathbf{a}_i^{\text{ul}}(\mathbf{a}_i^{\text{ul}})^\top\right) \\
&\overset{a}{=} \frac{1}{m+1}\mathbb{E}\|\mathbf{a}\|^2\mathbf{a}\mathbf{a}^\top + \frac{m}{m+1}\mathbf{\Sigma}^2 \\
&\overset{b}{\preceq} \frac{R^2 + m\lambda_{\max}(\mathbf{\Sigma})}{m+1}\mathbf{\Sigma} = R_m^2\mathbf{\Sigma},
\end{aligned}$$

where for $\overset{a}{=}$ we use $\mathbf{a}$ and $\mathbf{a}_i^{\text{ul}}$ are i.i.d. random variables, and $\overset{b}{\preceq}$ uses Assumption 3. For the second inequality, we have

$$\begin{aligned}\mathbb{E}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}\widehat{\boldsymbol{\Sigma}} &= \frac{1}{(m+1)^2}\mathbb{E}\left(\mathbf{a}\mathbf{a}^\top + \sum_{i=1}^m \mathbf{a}_i^{\text{ul}}(\mathbf{a}_i^{\text{ul}})^\top\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{a}\mathbf{a}^\top + \sum_{i=1}^m \mathbf{a}_i^{\text{ul}}(\mathbf{a}_i^{\text{ul}})^\top\right)\\ &\overset{a}{=} \frac{1}{m+1}\mathbb{E}\|\mathbf{a}\|^2_{\boldsymbol{\Sigma}^{-1}}\mathbf{a}\mathbf{a}^\top + \frac{m}{m+1}\boldsymbol{\Sigma}\\ &\overset{b}{\preceq} \frac{\tilde{\kappa}+m}{m+1}\boldsymbol{\Sigma} = \tilde{\kappa}_m^2\boldsymbol{\Sigma},\end{aligned}$$

where for $\overset{a}{=}$ we use $\mathbf{a}$ and $\mathbf{a}_i^{\text{ul}}$ are i.i.d. random variables, and $\overset{b}{\preceq}$ uses Assumption 3. □

For the inner loop, the stochastic gradient at $\mathbf{x}$ is

$$\ell'(\mathbf{a}^\top\tilde{\mathbf{y}}, b)\mathbf{a} + \widehat{\boldsymbol{\Sigma}}(\mathbf{x} - \tilde{\mathbf{y}}),$$

so the noise covariance $\mathbf{Q}$ is defined as

$$\mathbf{R} = \mathbb{E}\left(\ell'(\mathbf{a}^\top\tilde{\mathbf{y}}, b)\mathbf{a} - \widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}\nabla F(\tilde{\mathbf{y}})\right)\left(\ell'(\mathbf{a}^\top\tilde{\mathbf{y}}, b)\mathbf{a} + \widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}\nabla F(\tilde{\mathbf{y}})\right)^\top.$$

Follow the proof of Lemma D.18 and Lemma D.19, one has

$$\begin{aligned}\operatorname{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{R}) &\leqslant 5\operatorname{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{Q}) + 5L_\ell(L + \tilde{\kappa}_m)(F(\tilde{\mathbf{y}}) - F(\mathbf{x}^*)),\\ \operatorname{tr}\mathbf{R} &\leqslant 5\operatorname{tr}\mathbf{Q} + 5L_\ell(B + R_m^2)(F(\tilde{\mathbf{y}}) - F(\mathbf{x}^*)).\end{aligned}$$

The proofs can be directly extended to this setting by replacing Assumption 3 by Lemma H.1, and accordingly the quantity $R^2$, $\kappa$ and $\tilde{\kappa}$ will be replaced by $R_m^2$, $\kappa_m$ and $\tilde{\kappa}_m$, yielding the claimed result. Finally, note that $L \leqslant \alpha\tilde{\kappa}$, we derive the results in Section 8.

### H.3 SADA WITH MINI-BATCHING AND PARALLELIZATION

**Mini-batching.** The algorithm defaults to single-data updates but can be extended to a batch setting. Using a batch size of $B$, we can redefine $R^2$, $\kappa$ and $\tilde{\kappa}$ to $R_B^2 = \frac{R^2+(B-1)\lambda_{\max}(\boldsymbol{\Sigma})}{B}$, $\kappa_B = \frac{R_B^2}{\mu}$ and $\tilde{\kappa}_B = \frac{\tilde{\kappa}+(B-1)}{B}$, respectively. To obtain a solution with excess risk at most $\varepsilon$, the sample size $n$ is

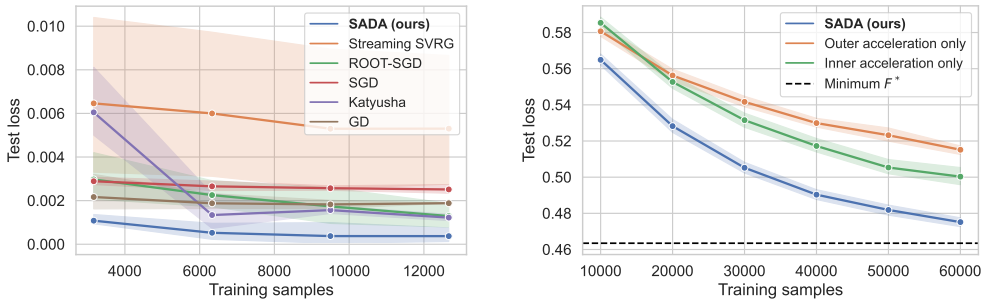
$$n \geqslant \tilde{\Theta}\left(B\left(\sqrt{\alpha\kappa_B\tilde{\kappa}_B} + \alpha^2\tilde{\kappa}_B\right) + \frac{\alpha\operatorname{tr}(\mathbf{H}^{-1}\mathbf{Q})}{\varepsilon} + \left(\frac{\alpha^2\tilde{\kappa}_B^2 B^2\operatorname{tr}\mathbf{Q}}{L_\ell\mu\varepsilon}\right)^{1/3}\right),$$

and the maximum batch-size $B_{\max} = n\sqrt{\mu\eta/(\alpha\kappa_B)}$, where we set $\eta = \tilde{\Theta}\left(\min\left\{\frac{1}{R_B^2}, \frac{n\varepsilon L_\ell}{\alpha\tilde{\kappa}_B\operatorname{tr}\mathbf{Q}}\right\}\right)$. and $K = \tilde{\Theta}(\sqrt{\alpha})$ and $T = n/K$. The proofs follow a similar argument to that used in the analysis with unlabeled data in Appendix H.2.

**Parallelization.** The inner loop of Algorithm 1 can be implemented in parallel when $T \geqslant T_0 = \tilde{\Theta}(\sqrt{\tilde{\kappa}/(\mu\eta)})$. The $T$ inner iterations can be partitioned into $T/T_0$ independent inner runs, and the outer iterate $\tilde{\mathbf{x}}_k$ is defined as the average of their outputs. The convergence guarantee in Theorem 5.1 holds under this parallel implementation.

## I ADDITIONAL EXPERIMENTS AND EXPERIMENTAL SETTING

This section gives the settings used to generate the sample-size–test-loss curves. The simulation selects the learning-rate multiplier on an independent validation split and reports the final loss on an independent test split; the real-data experiment follows the stated test-loss selection protocol. All searches are repeated independently for each sample budget and repeat.




(a) (b)

Figure 2: (a) Test loss versus the number of training samples on the MNIST dataset. (b) Ablation study on the simulated logistic regression task, comparing SADA with variants using only outer or inner acceleration.

## I.1 Additonal Experiments

### I.1.1 MNIST

We conduct experiments to compare the performance of the algorithms on the MNIST dataset. The experimental setting is detailed in Appendix I.2. We train a logistic regression model on the first two classes of MNIST. The result is shown in Figure 2(a). SADA achieves the lowest test loss among the compared methods as the number of training samples increases.

### I.1.2 Ablation Study

To investigate the contributions of the outer and inner acceleration in SADA, we compare SADA with two variants that retain only one of the two acceleration mechanisms. Specifically, *outer acceleration only* replaces the inner-loop solver by a gradient descent update. Hence, this algorithm is identical to Nesterov acceleration, while *inner acceleration only* removes the outer-loop acceleration and uses the 3-variable scheme, and the gradient is replaced by the stochastic gradient of the objective $F$. The experiment is conducted on the simulated logistic regression setting, as specified in Appendix I.2.1. The result is shown in Figure 2(b). The results show the necessity of incorporating both outer and inner acceleration.

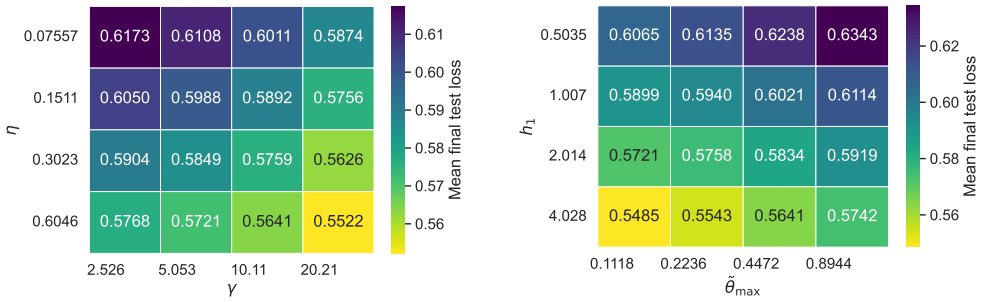
### I.1.3 Hyperparameter Sensitivity Analysis

We evaluate the hyperparameter sensitivity of SADA at sample budget $n = 10{,}000$, and report the final logistic test loss averaged over 20 independent runs in Figure 3. We perturb the inner-loop parameters $\eta$ and $\gamma$, and separately perturb the outer-loop parameters $h_k$ and $\tilde{\theta}_{\max}$. For a logarithmic search grid in which adjacent values differ by a factor of 10, any parameter value between the two candidates is within a factor of at most $\sqrt{10} \approx 3.16$ of its nearest candidate. From the experimental result, the loss is higher than the loss of the optimal hyperparameter by at most $\sim 0.05$. These results indicate that SADA does not require highly precise hyperparameter tuning and remains robust under the parameter mismatch from a coarse logarithmic search.

## I.2 Experimental Setting

### I.2.1 Numerical Simulation

**Logistic Regression.** We use the logistic loss with $d = 20$. Let $\lambda_k = k^{-4}$ for $k \in \{1, \dots, d\}$ and $\mathbf{\Sigma} = \mathrm{diag}(\lambda_1, \dots, \lambda_d)$. For every independent repeat, draw $\mathbf{a} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ and then draw binary label $b = \pm 1$ from $\mathbb{P}(b = 1 \mid \mathbf{a}) = 1/(1 + \exp(-\mathbf{a}^\top \mathbf{x}^*))$, where $x_k^* = 2/\sqrt{d\lambda_k}$. This setting is designed to require the algorithms to substantially reduce the initial error. We use $2{,}000$ initial training samples, an independent validation set of $400$ samples, and $10{,}000$ test samples. The sample budgets are $10{,}000$, $20{,}000$, $30{,}000$, $40{,}000$, $50{,}000$, and $60{,}000$. For each budget, we conduct 20




(a) Inner loop hyperparameters ($\eta$, $\gamma$)

(b) Outer loop hyperparameters ($h$, $\tilde{\theta}_{\max}$)

Figure 3: Sensitivity analysis of the hyperparameters of SADA. The result shows that the performance is not sensitive to the hyperparameters.

independent repetitions and report the corresponding $95\%$ confidence intervals computed across repetitions.

For streaming algorithms, including SADA, Streaming SVRG, ROOT-SGD, and SGD, each training sample is passed to the algorithm only once. In contrast, offline algorithms, including Katyusha and GD, are allowed to access the training samples repeatedly. For each method, sample budget, and repetition, we independently select the learning rate multiplier using the validation set. The search grid is $0.5, 1, 2, 4, 8, 16, 32, 64, 128$ for SADA, ROOT-SGD, and SGD, and $32, 64, 128, 256, 512, 1024$ for Streaming SVRG.

**Linear Regression.** The linear simulation uses the same covariance, ground truth $\mathbf{x}^*$, and sample size budgets. We draw label $b = \mathbf{a}^\top \mathbf{x}^* + 0.5\varepsilon$ with $\varepsilon \sim \mathcal{N}(0, 1)$ and evaluate the squared loss. Its Streaming SVRG grid is $\{32, 64, 128, 256, 512, 1024, 2048, 4096\}$; all other method grids and the single-pass/offline access rules are the same as the logistic regression task.

### I.2.2 Real Dataset

**CIFAR-10.** We use binary logistic regression on CIFAR-10. We retain the airplane and automobile classes. We use the official train–test split. This yields 10,000 training and 2,000 test images with dimension 3,073 for CIFAR-10. For preprocessing, we divide each pixel by 255, and append an intercept. We use three repeats. The training sample budgets are $\{2500, 5000, 7500, 10000\}$ for CIFAR-10. Each repeat independently permutes the official training set, and all methods and candidates use the same prefix of that permutation. For every budget and repeat, we perform grid search to select the optimal learning rate.

**MNIST.** For MNIST, we retain digits 0 and 1; for CIFAR-10, we retain airplane and automobile. We retain the official train–test split. This yields 12,665 training and 2,115 test images with dimension 785 for MNIST. The sample budgets are $\{3166, 6332, 9498, 12665\}$ for MNIST and $\{2500, 5000, 7500, 10000\}$ for CIFAR-10. We use the same preprocessing as for CIFAR-10.

### I.2.3 Algorithmic Details and Search Ranges

Except for SADA on CIFAR-10 and MNIST, we choose a base learning rate $\eta_0$ and search a multiplier $c_\eta$, hence the learning rate $\eta = c_\eta \eta_0$. In simulation, $\eta_0$ follows the corresponding paper. For SADA on CIFAR-10 and MNIST, the listed hyperparameters are searched directly.

- **SADA.** In simulation, we search $\eta$ multipliers in $\{0.5, 1, 2, 4, 8, 16, 32, 64, 128\}$ with $K = 5$ outer iterations, and $\gamma$ and $\theta$ computed from the covariance matrix based on (5) and (6). On CIFAR-10, the grid is $\eta \in \{5 \times 10^{-5}, 10^{-4}, 2 \times 10^{-4}, 5 \times 10^{-4}\}$, $\gamma \in \{0.005, 0.01, 0.02\}$, $\theta \in \{0.001, 0.01\}$, $\tilde{\theta} = 0.03$, $h = 2$, and $K \in \{1, 3, 10, 30\}$. On MNIST, we search $\eta \in \{0.001, 0.003, 0.006, 0.01\}$, $\gamma = 0.1$, $\theta = 0.001$, $\tilde{\theta} \in \{0.03, 0.2\}$, $h \in \{1, 2, 3, 4, 6, 8, 12, 16\}$, and $K \in \{1, 2, 3, 5, 8, 12, 20\}$.

- **Streaming SVRG.** We use the constant step-size and geometrically growing snapshot batches of Frostig et al. (2015), with batch-growth factor 3 and exponent $p = 2$. In simulation, the initial snapshot batch is 500, and the maximum inner length is 400; the multiplier grids are $\{32, 64, 128, 256, 512, 1024\}$ for logistic prediction and $\{32, 64, 128, 256, 512, 1024, 2048, 4096\}$ for linear prediction. On CIFAR-10, we use an initial snapshot batch of 20, maximum inner length 400, and grid $\{32, 128, 512, 2048\}$. On MNIST, we search two configurations: initial batch 20 and maximum inner length 400 with grid $\{32, 64, 96, 128, 192, 256, 384, 512, 768, 1024, 1536, 2048\}$.
- **ROOT-SGD.** Following Li et al. (2022), we use the recursive same-sample gradient estimator. The learning rate multiplier grids are $c_\eta \in \{0.5, 1, 4, 8, 16, 32, 64, 128\}$ in simulation, $\{0.5, 1, 4, 8, 16\}$ on CIFAR-10 and MNIST.
- **SGD.** We use one fresh sample per update with the decaying step $\eta_t = \eta/(1 + \rho t)$. In simulation, $\rho = 3 \times 10^{-4}$ and the multiplier grid is $\{0.5, 1, 2, 4, 8, 16, 32, 64, 128\}$. On CIFAR-10 and MNIST, $\rho = 0.003$, with learning rate multiplier $c_\eta \in \{1, 2, 4, 8\}$, respectively.
- **Katyusha.** Katyusha is an offline finite-sum baseline from Allen-Zhu (2018): it computes a full gradient and may reuse the stored training samples. In simulation, we use regularization $10^{-4}$, inner length 160, and grid $\{0.5, 1, 2, 4, 8, 16, 32, 64, 128\}$. On both CIFAR-10 and MNIST, we use the strongly convex proximal form with $\mu = 10^{-4}$, $\tau_2 = 1/2$, $\tau_1 = \min\{\sqrt{m\mu/(3L)}, 1/2\}$, and $m = 2n$, where $L$ bounds component smoothness. CIFAR-10 and MNIST use 20 epochs and grid $\{1, 2, 4\}$. The snapshot is the geometrically weighted epoch average prescribed by the strongly convex algorithm.
- **GD.** Each update reuses the complete training prefix to compute a full gradient, so gradient evaluations may exceed the number of unique samples on the horizontal axis. We use regularization $10^{-4}$ throughout. The simulation multiplier grid is $\{0.5, 1, 2, 4, 8, 16, 32, 64, 128\}$. On CIFAR-10, we use 400 steps with grid $\{1, 2, 4\}$. On MNIST, we use 1200 steps with grid $c_\eta \in \{16, 24, 32, 40, 48, 64\}$ at every budget.

For all methods, the plots report mean final test loss with 95% confidence intervals over repeats.

# J Additional Discussion

## J.1 Doubling Trick

The dependence of $\eta$ on the pre-known sample size $n$ can be removed using the standard doubling trick (Hazan & Kale, 2014). Specifically, instead of fixing a single horizon $n$ in advance, we divide the run into stages with geometrically increasing lengths, e.g., $1, 2, 4, 8, \ldots$. At the beginning of each stage, we restart SADA and set $\eta$ according to the horizon of that stage. Thus, the algorithm can be stopped at any time without knowing the final sample size in advance. Since the number of stages up to time $n$ is only $\mathcal{O}(\log n)$, this introduces at most a logarithmic restart overhead, while preserving the same order of the resulting bound.

## J.2 Hyperparameter Choice

One way to remove the dependence of the hyperparameter on unknown problem-dependent parameters is to perform a geometric search over the possible scales and restart the algorithm when moving to a new candidate. For example, suppose $D$ is chosen to upper bound $\|\mathbf{x}^*\|$ for SADA-NonSC. We can start from a conservative value $D_0$ and run SADA; if the resulting solution satisfies $\|\widehat{\mathbf{x}}\| \le D/2$, then the current radius is unnecessarily large, and we replace $D$ by $D/2$ and restart. The same procedure applies to $M$ and $\|\mathbf{x}^*\|_{\mathbf{H}}$. For a hyperparameter such as $\eta$, we can search over a geometric grid $\{\eta_0 2^j : j \in \mathbb{Z}\}$ rather than requiring its value to be known exactly in advance. Consequently, only a logarithmic number of scales need to be explored between an initial conservative value and the appropriate problem-dependent scale, leading to only logarithmic overhead from the restarts.

In practice, another simple approach is to use a small pilot sample to estimate quantities such as $\kappa$, $\tilde{\kappa}$, and $\mathrm{tr}\,\mathbf{Q}$, and then set the hyperparameters according to (5) and (6). To support the practical applicability, we perform a sensitivity analysis on the hyperparameters $\eta$, $\mu$, $h_k$, $\theta_{\max}$ in Appendix I.1.3. We find that our algorithm is not very sensitive to the hyperparameters.