A stochastic gradient algorithm for non-separable optimization with convergence guarantee

Yingzhou Li*, Ruofan Wu†

Abstract

We study non-separable objectives in which the loss depend on dataset-level quantities. We introduce an SGD-style framework that employs two batch-gradient constructs: the ideal per-batch gradient 'G' and a cached surrogate 'H' for cases where full-data terms are expensive. Notably, in the sample-wise separable case, our method reduces to standard mini-batch SGD. Our main contribution is a unified local convergence theory: under mild smoothness and Jacobian-boundedness assumptions, we prove local linear convergence under local strong convexity and local $O(1/k)$ sublinear convergence under local convexity for both 'G'-driven and 'H'-driven updates. Crucially, these guarantees hold for fixed step sizes within explicitly characterized ranges; we provide explicit bounds showing how cache staleness, surrogate approximation error, batch size, and step size influence the convergence constants and allowable step-size ranges.

Keywords: Non-separable optimization, stochastic gradient descent, constant step size, local convergence theory

1 Introduction

Modern deep learning systems are trained by minimizing an objective function that quantifies the fit between the model's behavior and the observed training data. The dominant paradigm assumes that this objective decomposes into independent per-sample losses, an assumption that underpins stochastic gradient descent (SGD) and its variants. In this paper, we study the setting where this separability fails and the objective couples all samples through a global operator. As a point of departure, the standard setting is expressed as a sample-wise separable finite sum,

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \ell(\theta; x_i, y_i),$$

where each term depends only on a single sample (x_i, y_i), and mini-batch SGD provides unbiased gradient estimates. In contrast, the objective we consider takes the coupled form

$$\mathcal{L}(\theta) = F\left(\{\phi(\theta; x_i, y_i)\}_{i=1}^{N}\right),$$

where $\phi(\theta; x_i, y_i)$ denotes the per-sample model output and F is a coupling operator that depends on the whole dataset. A growing class of modern learning objectives—including those arising in

*School of Mathematical Sciences, Fudan University; Shanghai Key Laboratory for Contemporary Applied Mathematics, Fudan University, yingzhouli@fudan.edu.cn

†School of Mathematical Sciences, Fudan University, rfwu22@m.fudan.edu.cn

contrastive representation learning, deep clustering, and losses involving global dataset statistics—falls into this non-separable category. In such settings, the mini-batch gradient

$$\tilde{g}_k = \nabla_\theta F\left(\{\phi(\theta_k; x_j, y_j)\}_{j\in\mathcal{B}_k}\right)$$

is generally a biased estimator of the true gradient: a small batch cannot faithfully represent the global coupling operator, leading to both relation truncation and skewed batch statistics.

Training deep neural networks is typically approached through iterative stochastic first-order methods, with SGD and its adaptive variants—Adagrad [1], ADADELTA [2], Adam [3]—forming the practical backbone. These methods estimate gradients from randomly sampled mini-batches, which drastically reduce per-iteration cost relative to full-batch approaches and make large-scale optimization tractable [1, 2, 3, 4, 5, 6]. Variance-reduction methods (e.g., SAG [4], SVRG [5], SAGA [6]) can accelerate convergence by reducing stochastic gradient noise, yet their theoretical guarantees remain restricted to separable objectives. When the objective is non-separable, the mini-batch gradient becomes inherently biased—a structural error driven by relation truncation and skewed batch statistics, as noted above. Crucially, this bias is distinct from ordinary stochastic variance, persisting even in highly expressive regimes, and is not addressed by variance-reduction techniques that assume a decomposable objective. The non-separable case therefore remains inadequately served by general-purpose stochastic algorithms. While a few specialized methods have been proposed for particular non-decomposable settings [7, 8, 9], they do not extend to general coupled objectives.

Even for separable objectives, fixed-step SGD is theoretically limited to convergence to a neighborhood of optima, as highlighted in standard analyses [10], which leaves a gap between empirical success and formal guarantees. The nonconvex, high-dimensional, and often nonsmooth nature of deep-learning objectives poses substantial theoretical and practical challenges. In particular, vanilla SGD enjoys the advantages of simplicity and low per-iteration cost, but convergence to the exact optimum with a fixed step size typically require restrictive conditions such as interpolation (i.e., the existence of a model that fits all training samples simultaneously). Classical analyses for general nonconvex problems often demand diminishing step sizes to ensure asymptotic convergence to stationary points, yielding slow sublinear rates that limit practical efficiency [11, 12, 13, 14, 15, 16, 17, 10, 18, 19]. Variance-reduction methods offer an alternative route: by correcting the stochastic gradient noise, they enable exact convergence under constant step sizes for separable finite-sum problems, but at the cost of additional memory or periodic full-gradient computations [5, 6]. Modern machine learning (ML) and artificial intelligence (AI) systems increasingly rely on highly over-parameterized models, particularly deep neural networks (DNNs). In this regime, models are sufficiently expressive to represent optimal solutions and can nearly perfectly fit training data even when the number of parameters far exceeds the sample size [20, 21, 22, 23]. A large body of work documents the "double descent" phenomenon: after a critical model-complexity threshold, further over-parameterization can improve generalization and fundamentally alter the geometry of the loss landscape [24, 25, 26]. Recent theoretical progress has shown that under certain structural conditions—most prominently interpolation—fixed-step SGD can enjoy much stronger behavior. Bassily et al. [27] proved exponential convergence rates for SGD on over-parameterized nonconvex models under interpolation, where a global minimizer fits all training samples exactly. Under gradient-dominance (Polyak–Łojasiewicz) conditions, Vaswani et al. [28] obtained linear convergence rates without requiring strong convexity. Liu et al. [29] showed that implicit regularization in over-parameterized regimes can bias SGD toward flat minima, enabling fast convergence without explicit step-size decay. These results illustrate the exceptional efficiency of vanilla SGD when interpolation or related structural assumptions hold. However, many practical settings violate strict interpolation: examples include noisy labels, explicit regularization (e.g., weight decay), and inherently non-interpolable tasks like unsupervised clustering. In such non-interpolating regimes, vanilla SGD's speed and guarantees often deteriorate.

All of these advances—variance reduction and interpolation-based analysis alike—are derived under the assumption that the objective is sample-wise separable [27, 28, 29, 30, 31, 32, 33, 34]. Because they rest on the separability assumption, they do not resolve the structural bias that arises when the objective depends on global sample interactions. Crucially, non-separable objectives with global sample coupling are inherently non-interpolable in the mini-batch setting: because the loss depends on the full dataset, a model that interpolates a single mini-batch does not necessarily interpolate the entire training set. Consequently, even if one could construct an unbiased gradient estimator, the fast-convergence theory for fixed-step SGD would remain inapplicable. Taken together, the non-separable setting suffers from a compounding of two failures: the standard mini-batch gradient is structurally biased, and the fast-convergence theory for fixed-step SGD is inapplicable. To address these issues, we propose a novel algorithmic approach. Under the over-parameterized regime, our proposed algorithm admits exact local convergence with a constant step size, while retaining the computational simplicity of mini-batch SGD. Our main contributions are:

- We introduce an algorithm that generalizes standard mini-batch SGD to sample-coupled objectives by correcting the inherent bias in batch-wise gradient estimates. Under the sample-wise separable case, the method reduces to standard mini-batch SGD.
- We prove that, for objectives that are locally convex near an optimum, the algorithm converges locally to the exact optimum with a constant step size, without the step-size decay that standard SGD requires to suppress gradient noise.
- We establish fast local convergence rates: locally linear convergence in the presence of local strong convexity, and locally sublinear convergence for merely locally convex objectives, both under constant step sizes.

The remainder of the paper is organized as follows. In Section 2 we present the algorithm. Section 3 gives a local convergence analysis for the ideal per-batch gradient G, and Section 4 presents analogous results for the cached surrogate H. Section 5 reports numerical experiments that illustrate the empirical behavior of the method. We conclude with final remarks and directions for future work in Section 6.

2 A novel stochastic gradient algorithm

2.1 Background

Building on the problem setting introduced in Section 1, this section reviews the mini-batch SGD framework and formalizes the structural bias that arises when global sample coupling prevents unbiased gradient estimation. Most neural network training problems are optimization problems: one minimizes the loss function $\mathcal{L}(\theta)$ to obtain optimal parameters θ. Classical (full-batch) gradient descent updates parameters by

$$\theta_{k+1} = \theta_k - \eta \nabla_\theta \mathcal{L}(\theta_k),$$

but this is computationally prohibitive for the large datasets typical in deep learning. To address this, stochastic gradient descent [35], and in particular its mini-batch variant, uses randomly sampled subsets $\mathcal{B}_k$ to approximate the true gradient. The parameter update at step k is

$$\theta_{k+1} = \theta_k - \eta_k \frac{1}{|\mathcal{B}_k|} \sum_{i \in \mathcal{B}_k} \nabla_\theta \ell(\theta_k; x_i, y_i),$$

where x_i and y_i denote the input features and target (or label) of the i-th sample, respectively, $\ell(\cdot)$ denotes the per-example loss, $|\mathcal{B}_k|$ is the mini-batch size (typically 32–512), and η_k is the learning rate. This approach offers three practical properties [36, 37]:

- *Hardware efficiency*: mini-batches enable parallel processing (e.g., on GPUs/TPUs) and higher hardware utilization;
- *Variance reduction*: averaging over a batch produces a more stable gradient estimator than single-sample updates;
- *Implicit regularization*: controlled stochasticity often improves generalization.

The standard mini-batch SGD framework assumes a decomposable objective $\mathcal{L}(\theta) = \mathbb{E}_{(x,y)}[\ell(\theta; x, y)]$, so that per-batch gradients provide unbiased estimates of the full gradient. Modern objectives, however, cannot be decomposed into independent per-sample losses; as introduced in Section 1, they instead couple the entire dataset through a global operator of the form

$$\mathcal{L}(\theta) = F\left(\{\phi(\theta; x_i, y_i)\}_{i=1}^N\right),$$

where F is a coupling operator (for example, a sorting or a statistic over the dataset). In such settings the mini-batch gradient

$$\tilde{g}_k = \nabla_\theta F\left(\{\phi(\theta_k; x_j, y_j)\}_{j\in\mathcal{B}_k}\right)$$

is generally a biased estimator of $\nabla_\theta \mathcal{L}$, with the bias depending on $|\mathcal{B}_k|$ and on the sensitivity of F [38].

This issue is particularly acute in unsupervised learning, where objectives inherently involve relational coupling between samples [37, 39]. Given unlabeled data $\mathcal{X} = \{x_i\}_{i=1}^N$, losses typically take the form:

$$\mathcal{L}(\theta) = F\Big(\{\phi_\theta(x_i), \{\phi_\theta(x_j)\}_{j\in\mathcal{R}(x_i)}\}_{i=1}^N\Big)$$

where ϕ_θ is a representation encoder and $\mathcal{R}(x_i)$ denotes a relational set (for example, neighbors used in contrastive learning). The corresponding mini-batch gradient commonly used in practice is

$$\tilde{g}_k = \nabla_\theta F\left(\{\phi_\theta(x_j), \{\phi_\theta(x_l)\}_{l\in\mathcal{R}(x_j)\cap\mathcal{B}_k}\}_{j\in\mathcal{B}_k}\right)$$

which gives rise to two challenges beyond the usual stochastic gradient variance [40]:

- *Relation truncation*: the intersection $\mathcal{R}(x_j) \cap \mathcal{B}_k$ may omit important long-range dependencies, degrading the fidelity of the batch-wise objective;
- *Biased statistics*: dataset-level operators F (e.g., distribution-matching or global moments) produce skewed statistics when evaluated on small batches, resulting in biased gradient estimates.

While reducing the step size can mitigate the effect of stochasticity, it slows convergence. To achieve fast convergence with constant step sizes, we therefore seek stochastic gradient algorithms that produce (or closely approximate) unbiased gradient estimates in the presence of inter-sample coupling.

2.2 Setup

In the following, for convenience and consistency we express the loss as:

$$\mathcal{L}(\theta) = F\big(Y(\theta)\big)$$

where $Y(\theta) \in \mathbb{R}^{N\times d}$ is the output matrix whose rows are sample-wise outputs, and F is a nonlinear function. To facilitate analysis, we focus on domains where F is locally convex in Y, a condition often met in neural-network optimization landscapes.

In traditional mini-batch training one evaluates the model on a subset (batch) $\mathcal{B}_k$ and uses the resulting batch-based quantity to form an update. To represent this formally we use a masking interpretation: let $\mathrm{mask}_{\mathcal{B}} : \mathbb{R}^{N\times d} \to \mathbb{R}^{N\times d}$ denote the operator that keeps rows indexed by $\mathcal{B}$ and ignores (i.e., masks out) other rows. We then denote the batch-evaluated object by $Y_{\mathcal{B}} \coloneqq \mathrm{mask}_{\mathcal{B}}(Y)$. Note that when F is sample-wise separable (i.e., $F(Y) = \sum_n f(Y_n)$) this mask is equivalent to zeroing out the non-batch rows; however, when F depends on global statistics (e.g., $Y^\top Y$ or distribution-matching operators) zeroing out the non-batch rows is not equivalent to true batch evaluation. In such cases, we treat the batch operator as "apply F only to the submatrix formed by the batch rows" (i.e., evaluate F on the dataset restricted to $\mathcal{B}$), and denote this by $F_{\mathcal{B}}(Y)$.

By the chain rule the full-parameter gradient can be written compactly using a contraction over the sample and output indices. For notational clarity, we denote this contraction by $\times$. Concretely, for an $N \times d$ matrix A and an $N \times d \times s$ tensor B, we define

$$A \times B \in \mathbb{R}^s, \quad (A \times B)_m = \sum_{n=1}^{N}\sum_{j=1}^{d} A_{nj}B_{njm},$$

so that

$$\nabla_\theta \mathcal{L} = \nabla_Y F(Y(\theta)) \times \nabla_\theta Y(\theta),$$

Equivalently, writing per-sample Jacobians $\nabla_\theta Y_n(\theta) \in \mathbb{R}^{d\times s}$ and the n-th row $\nabla_Y F_n \in \mathbb{R}^d$, we have the componentwise expansion

$$\nabla_\theta \mathcal{L} = \sum_{n=1}^{N} (\nabla_Y F)_n \cdot \nabla_\theta Y_n(\theta)$$

where the dot denotes multiplication of a row vector with a $d \times s$ Jacobian yielding an s-vector.

In the mini-batch formulation, the gradient commonly computed in implementations can be written as

$$\tilde{g}_k = \nabla_\theta F\left(Y_{\mathcal{B}_k}(\theta_k)\right) = \nabla_Y F\left(Y_{\mathcal{B}_k}(\theta_k)\right) \times \nabla_\theta Y(\theta_k).$$

When $\nabla_Y F\left(Y_{\mathcal{B}_k}(\theta_k)\right)$ is nonzero only on rows indexed by $\mathcal{B}_k$ the contraction reduces to summation over $n \in \mathcal{B}_k$. In practical implementations one only computes the per-sample Jacobians $\nabla_\theta Y_n(\theta_k)$ for $n \in \mathcal{B}_k$ (i.e., the $|\mathcal{B}_k| \times d$ block), and automatic differentiation yields these Jacobians exactly for the current batch; thus the potential source of statistical bias in $\tilde{g}_k$ is the batch-evaluated output-gradient $\nabla_Y F(Y_{\mathcal{B}_k})$ (or more generally the batch operator $F(Y_{\mathcal{B}_k})$) rather than the per-sample Jacobians. When F is non-separable or depends on global statistics, it is typical that $\mathbb{E}_{\mathcal{B}}[\nabla_Y F(Y_{\mathcal{B}})] \neq \nabla_Y F(Y)$, and hence $\tilde{g}$ is biased. Revisiting the example from Section 2.1, consider

$$\mathcal{L}(\theta) = F\Big(\{\phi_\theta(x_i), \{\phi_\theta(x_j)\}_{j\in\mathcal{R}(x_i)}\}_{i=1}^{N}\Big)$$

where ϕ_θ is a representation encoder and $\mathcal{R}(x_i)$ denotes a relational set. The practical mini-batch evaluation replaces the global interactions by their batch-restricted counterparts, leading to a batch gradient of the form

$$\tilde{g}_k = \nabla_Y F\left(\{\phi_\theta(x_j), \{\phi_\theta(x_l)\}_{l\in\mathcal{R}(x_j)\cap\mathcal{B}_k}\}_{j\in\mathcal{B}_k}\right) \times \nabla_\theta Y(\theta_k),$$

with the output-gradient factor alone carrying the relation truncation and skewed statistics.

The per-sample Jacobian slices $\nabla_\theta Y_n$ for $n \in \mathcal{B}$ are deterministic given θ and are computed exactly by automatic differentiation, hence they do not introduce statistical bias; therefore the bias of the mini-batch estimator is determined primarily by the output-gradient $\nabla_Y F$ (or the batch operator $F_{\mathcal{B}}$). To mitigate this bias while preserving per-iteration efficiency, we next introduce notation and approximations that allow us to construct batch gradients that more faithfully estimate the true parameter gradient.

2.3 Ideal and Surrogate Stochastic Gradient Descent

We will use the following names and symbols for the gradient estimators that appear throughout the paper. The ideal per-batch gradient, denoted G, corresponds to selecting the rows of the full output-gradient $\nabla_Y F$ associated with the current batch and contracting them with the per-sample Jacobians; G is unbiased but typically requires access to global data-dependent quantities. The cached surrogate gradient, denoted H, replaces expensive global terms in $\nabla_Y F$ by cheap surrogates computed from the cached outputs $\tilde{Y}$; H is computationally efficient and reduces to G when the cached quantities are exact.

Define the ideal batch-evaluated gradient (using the masked full output) as

$$G = \nabla_{\mathcal{B}} F\left(Y(\theta)\right) \times \nabla_\theta Y(\theta),$$

where $\nabla_{\mathcal{B}} F$ is the $N \times d$ matrix that equals $\nabla_Y F$ on rows indexed by the batch $\mathcal{B}$ and is zero elsewhere.

Since $\nabla_\theta Y$ can be viewed as the vertical concatenation of per-sample Jacobians,

$$\nabla_\theta Y = \begin{bmatrix} \nabla_\theta Y_1 \\ \nabla_\theta Y_2 \\ \vdots \\ \nabla_\theta Y_N \end{bmatrix}, \quad \nabla_\theta Y_n = \frac{\partial Y_n}{\partial \theta} \in \mathbb{R}^{d \times s}.$$

Here $\nabla_\theta Y_n$ is the Jacobian of the n-th sample's output $Y_n \in \mathbb{R}^d$ with respect to parameters θ. Thus the parameter gradient can be written as

$$\nabla_\theta \mathcal{L} = \nabla_Y F(Y(\theta)) \times \nabla_\theta Y = \sum_{n=1}^{N} (\nabla_Y F)_n \cdot \nabla_\theta Y_n(\theta)$$

where $(\nabla_Y F)_n$ is the n-th row of $\nabla_Y F$ and $\nabla_\theta Y_n(\theta)$ is the Jacobian of the n-th output. Therefore,

$$G = \nabla_{\mathcal{B}} F\left(Y(\theta)\right) \times \nabla_\theta Y(\theta) = \sum_{n \in \mathcal{B}} (\nabla_Y F)_n \cdot \nabla_\theta Y_n(\theta)$$

which matches the computational pattern of standard mini-batch gradients and, by construction, unbiasedly estimates $\nabla_\theta \mathcal{L}$.

Computing the exact rows $(\nabla_Y F)_n$ is non-trivial because the n-th row generally couples to all other rows of Y through the global operator F. To address this we introduce an approximate output-gradient $\widetilde{\nabla_Y F}$ that depends on a cached approximation $\tilde{Y}$ and on the current Y:

$$\widetilde{\nabla_Y F} = \widetilde{\nabla_Y F}(\tilde{Y}, Y),$$

and define the practical batch estimator

$$H = \widetilde{\nabla_{\mathcal{B}} F}\left(Y(\theta)\right) \times \nabla_\theta Y(\theta) = \sum_{n \in \mathcal{B}} (\widetilde{\nabla_Y F})_n \cdot \nabla_\theta Y_n(\theta),$$

where $\widetilde{\nabla_{\mathcal{B}} F}$ equals $\widetilde{\nabla_Y F}$ on rows in $\mathcal{B}$ and is zero elsewhere. The approximation $\widetilde{\nabla_Y F}(\tilde{Y}, Y)$ is constructed by evaluating computationally inexpensive terms at the current Y while substituting $\tilde{Y}$ for the expensive-to-evaluate global terms. Concretely:

- $\tilde{Y}$ is a cached approximation of Y maintained across iterations; on iteration k we update the cached rows for the current batch and keep other rows unchanged:

$$(\tilde{Y}_{\text{new}})_n = \begin{cases} (Y_{\text{new}})_n & \text{if } n \in \mathcal{B} \\ (\tilde{Y}_{\text{old}})_n & \text{if } n \notin \mathcal{B} \end{cases}$$

- The approximation replaces only the expensive parts of $\nabla_Y F$ by expressions evaluated at $\tilde{Y}$; inexpensive or local terms are evaluated at the current Y.

- This design allows incremental updates of global quantities (e.g., $\tilde{Y}^\top \tilde{Y}$) using low-cost rank-$|\mathcal{B}|$ updates. For instance, if $\nabla_Y F(Y) = AY + BY^\top Y$ with A sparse or diagonal, then AY can be evaluated at the current Y while $BY^\top Y$ is approximated using $\tilde{Y}$. The cached Gram matrix admits the cheap update

$$\tilde{Y}_{\text{new}}^\top \tilde{Y}_{\text{new}} = \tilde{Y}_{\text{old}}^\top \tilde{Y}_{\text{old}} - \sum_{n\in\mathcal{B}} (\tilde{Y}_{\text{old}})_n^\top (\tilde{Y}_{\text{old}})_n + \sum_{n\in\mathcal{B}} (\tilde{Y}_{\text{new}})_n^\top (\tilde{Y}_{\text{new}})_n.$$

  Hence one can update $\widetilde{\nabla_{\mathcal{B}} F}$ without recomputing quantities over $\mathcal{B}^c$ (more precisely over $\mathcal{B}_{\text{new}}^c \cap \mathcal{B}_{\text{old}}^c$).

Note that whereas the standard mini-batch gradient is obtained directly by automatic differentiation, our proposed gradient requires additional processing. Concretely, one can obtain the parameter update by differentiating the surrogate objective

$$\frac{1}{2}\text{tr}(Y_{\mathcal{B}}(\theta)^\top \nabla_{\mathcal{B}} F) \quad \left(\text{or} \quad \frac{1}{2}\text{tr}(Y_{\mathcal{B}}(\theta)^\top \widetilde{\nabla_{\mathcal{B}} F})\right),$$

treating $Y_{\mathcal{B}}(\theta)$ as a function of θ while holding $\nabla_{\mathcal{B}} F$ (or $\widetilde{\nabla_{\mathcal{B}} F}$) fixed.

Although the proposed batch gradient can be unbiased, it may still converge only to a neighborhood of optimal points, similar to the SGD analysis [10]. In the next section we show that under additional conditions the estimator can converge asymptotically to the exact optimum.

2.4 Example of the gradient algorithm

In this section, we discuss the relationship between the proposed method and existing algorithms. Under specific conditions the proposed estimator coincides with prior methods.

- Sample-wise separable losses. If the loss decomposes per sample,

$$\mathcal{L}(\theta) = \sum_{n=1}^{N} f_n(\mathbf{y}_n(\theta)),$$

  then the Jacobian $\nabla_Y F$ is block-diagonal and the parameter gradient decomposes into independent per-sample terms. In this case the mini-batch update equals the ideal per-batch gradient 'G', so the cached surrogate 'H' is unnecessary and the method reduces to standard mini-batch SGD.

- Relation to SpecNet2 [41]. For the SpecNet2 objective, practical batch-gradient schemes (local/batch-restricted, full-data, or neighborhood-based) correspond to different choices of per-row output-gradients in our framework. Concretely:

  - Full-data evaluation (use rows computed from the entire dataset) yields the ideal gradient 'G'.
  - Batch-restricted / naive evaluation (use only rows from the current batch) yields the usual mini-batch gradient, which can be biased.
  - Neighbor/localized evaluation (use cached or neighborhood-based rows) corresponds to the surrogate 'H'.

Actually, our approach is inspired by these revision and approximation strategies: we approximate costly global terms by cached or localized surrogates while leveraging per-sample Jacobians computed by automatic differentiation. The same principles apply to the gradient schemes Chen et al. propose for SpecNet1.

3 Local convergence analysis for the ideal per-batch gradient

In this section, we assume that F is locally convex around Y^*. Within this local region, we analyze the convergence of the ideal per-batch gradient. Specifically, it will be shown that when F is locally strongly convex around Y^*, the algorithm exhibits local linear convergence; whereas when F is merely locally convex around Y^*, it exhibits local sublinear convergence.

To formalize this analysis, the section proceeds in two parts. Section 3.1 establishes the required notation and standing assumptions; Section 3.2 then states and proves the convergence theorems under these assumptions.

3.1 Notation and standing assumptions

In this section, we adopt the same notation as in sections 2.2 and 2.3 (in particular $Y \in \mathbb{R}^{N\times d}$, $\nabla_Y F$, per-sample Jacobians $\nabla_\theta Y_n$, and the definition $G = \nabla_{\mathcal{B}} F(Y) \times \nabla_\theta Y$). For later reference, define

$$z_i := \left((\nabla_Y F)_i \cdot \nabla_\theta Y_i\right)^\top,$$

so that the (transposed) batch gradient used by the algorithm is $G^\top = \sum_{i\in\mathcal{B}} z_i$. If we use θ^+ to represent the updated parameters, with a fixed step size $\eta > 0$, the parameter update is

$$\theta^+ = \theta - \eta \cdot G^\top = \theta - \eta \cdot \sum_{i\in\mathcal{B}} z_i,$$

and to simplify the subsequent proofs, we assume all mini-batches have the same size $b = |\mathcal{B}|$. The results could be extended to variable batch sizes.

Having specified notational conventions, we now state the standing assumptions used throughout the convergence analysis.

Assumptions.

- (A_1) $\|\nabla_Y F\|_{\mathrm{F}}$ locally uniformly bounded by B_0.
- (A_2) The map $Y \mapsto \nabla_Y F(Y)$ is L-Lipschitz in Frobenius norm; i.e. for all Y, Y' in the local region,
$$\|\nabla_Y F(Y) - \nabla_Y F(Y')\|_{\mathrm{F}} \le L\|Y - Y'\|_{\mathrm{F}}.$$
(Equivalently, $\nabla_Y F$ has Lipschitz constant L.)
- (A_3) $\|\nabla_\theta Y\|_{\mathrm{F}}$ is locally uniformly bounded by B_1.
- (A_4) $\|\nabla^2_\theta Y\|_{\mathrm{F}}$ is locally uniformly bounded by B_2.
- (A_5) The smallest eigenvalue of $\nabla_\theta Y \cdot (\nabla_\theta Y)^\top$ admits a locally uniform positive lower bound:
$$\lambda_{\min}\left(\nabla_\theta Y \cdot (\nabla_\theta Y)^\top\right) \ge \lambda_{\min} > 0.$$

Remark 3.1. Assumption A_5 requires a uniform positive lower bound on the smallest eigenvalue of $\nabla_\theta Y(\theta) \cdot \nabla_\theta Y(\theta)^\top$ (equivalently the smallest nonzero singular value of the empirical Jacobian) in the optimization region of interest. We emphasize three points. First, Assumption A_5 is a local non-degeneracy condition: it is assumed to hold along the optimization path or within the neighborhood where the analysis applies, not necessarily globally over the entire parameter space. Second, similar non-degeneracy has been rigorously or empirically observed in several overparameterized regimes; for example, NTK-style results establish spectral stability for sufficiently wide networks at initialization and for short training times (see [42, 43] and follow-ups). We therefore expect Assumption A_5 to hold approximately for wide architectures under standard initialization and training regimes, though extending such guarantees to long training horizons and finite widths may require additional assumptions. Third, Assumption A_5 is empirically verifiable: one can monitor the smallest singular value of a small empirical Jacobian (e.g., computed on a fixed held-out subset) during training. If this quantity approaches zero for a particular model, the theoretical guarantees relying on Assumption A_5 may not apply and empirical evaluation should be used.

3.2 Main theorems and lemmas

The lemmas and propositions stated below serve as the principal technical components for the convergence analysis of the algorithm. For brevity, we present their statements and brief description here and defer all proofs to Appendix A. Our main convergence results are as follows. Theorem 3.4 establishes linear convergence of the algorithm under local strong convexity, and Theorem 3.6 establishes $O(1/k)$ sublinear convergence under local convexity.

The following lemma relates the per-step actual gradient estimator to the full gradient, providing a norm inequality that will be used repeatedly in the single-step bounds.

Lemma 3.1. *If Assumption A_3 holds, then $\|G\|_{\mathrm{F}} \le B_1 \|\nabla_Y F\|_{\mathrm{F}}$.*

Then, we explain the single-step descent situation through Lemma 3.2 and Lemma 3.3.

Lemma 3.2. *If Assumptions A_1-A_4 hold, denote $Y^+ = Y(\theta^+)$, then there exists a constant $c_2 \ge 0$ such that*

$$F(Y^+) \le F(Y) - \eta \cdot \sum_{i=1}^{N} \langle z_i, G^\top \rangle + c_2 \cdot \eta^2 \|\nabla_Y F\|_{\mathrm{F}}^2.$$

Specifically, one may take $c_2 = \frac{1}{2}\left(LB_1^2 + B_0 B_2\right) c_1$, where $c_1 = B_1^2$.

Lemma 3.3. *If Assumptions A_1-A_5 hold, while $\eta < \frac{|\mathcal{B}|}{N} \cdot \frac{\lambda_{\min}}{c_2}$, there exists a constant $c_3 > 0$, such that*

$$\mathbb{E}_{\mathcal{B}}[F(Y^+)] \le F(Y) - c_3 \cdot \eta \|\nabla_Y F\|_{\mathrm{F}}^2.$$

Specifically, one may take $c_3 = \left(\frac{|\mathcal{B}|}{N}\lambda_{\min} - c_2\eta\right)$.

Remark 3.2. Up to and including Lemma 3.3 we have not used local convexity of F in Y. Hence Lemma 3.3 already yields the standard stochastic stationarity guarantee.

Summing the inequality in Lemma 3.3 over $k = 0, \ldots, T$ and taking total expectation gives

$$\sum_{k=0}^{T} \mathbb{E}\big[\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2\big] = \frac{1}{c_3 \cdot \eta} \cdot \left(F(Y(\theta_0)) - \mathbb{E}\big[F(Y(\theta_{T+1}))\big]\right) \le \frac{1}{c_3 \cdot \eta} \cdot F(Y(\theta_0)).$$

Thus

$$\min_{k=0,\cdots,T} \mathbb{E}\big[\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2\big] \le \frac{F(Y(\theta_0))}{c_3 \cdot \eta \cdot (T+1)},$$

which implies $\min_{k \le T} \mathbb{E}\big[\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}\big] = O(T^{-1/2})$. The detailed step-by-step derivation is given in Appendix A.4.

With the preceding lemmas in place, we now state the main convergence results.

Theorem 3.4. *Suppose F is locally strongly convex around Y^* with strong convexity constant $c_Y > 0$. If Assumptions A_1-A_5 hold and the step size satisfies*

$$\eta < \min\left\{ \frac{|\mathcal{B}|}{N} \cdot \frac{\lambda_{\min}}{c_2},\ \frac{1}{2c_3 c_Y} \right\},$$

then while the iterates remain in a neighborhood of Y^* where the local assumptions hold, the expected optimality gap converges linearly to zero: there exist constants $C > 0$ and $\rho \in (0,1)$ such that

$$\mathbb{E}\big[F(Y(\theta_k))\big] - F(Y^*) \le C\rho^k,$$

where $\mathbb{E}\big[F(Y(\theta_k))\big]$ denotes $\mathbb{E}_{\mathcal{B}_0,\cdots,\mathcal{B}_{k-1}}\big[F(Y(\theta_k))\big]$ and Y^* is the minimum point of F.

Corollary 3.5. *Under the assumptions of Theorem 3.4, while the iterates remain in the neighborhood of Y^*, the optimality gap $F(Y(\theta_k)) - F(Y^*)$ converges to zero almost surely at a linear rate. More precisely, for almost every sample path ω, there exist $\rho(\omega) \in (0,1)$ and $k_0(\omega) \in \mathbb{N}$ such that*

$$F(Y(\theta_k)) - F(Y^*) \le \rho(\omega)^k, \quad \forall k \ge k_0(\omega).$$

Theorem 3.6. *Suppose F is locally convex around Y^*. If Assumptions A_1-A_5 hold and the step size satisfies $\eta < \frac{|\mathcal{B}|}{N} \cdot \frac{\lambda_{\min}}{c_2}$, then while the iterates remain in a neighborhood of Y^* where the local assumptions hold, the expected optimality gap converges sublinearly to zero at a rate of $O(1/k)$: there exists a constant $C > 0$ such that*

$$\mathbb{E}\big[F(Y(\theta_k))\big] - F(Y^*) \le \frac{C}{k},$$

where $\mathbb{E}\big[F(Y(\theta_k))\big]$ denotes $\mathbb{E}_{\mathcal{B}_0,\cdots,\mathcal{B}_{k-1}}\big[F(Y(\theta_k))\big]$ and Y^* is a minimum point of F.

Corollary 3.7. *Under the assumptions of Theorem 3.6, while the iterates remain in the neighborhood of Y^*, the following hold:*

- *$F(Y(\theta_k)) - F(Y^*)$ converges to zero almost surely.*
- *The sequence $F(Y(\theta_k))$ admits almost surely convergent random subsequences at rate of $O(1/k)$, i.e., for almost every sample path ω, there exist a subsequence $k_n(\omega)$ and a constant $C(\omega)$ such that*

$$F(Y(\theta_{k_n})) - F(Y^*) \le \frac{C(\omega)}{k_n}, \quad \forall\ n \in \mathbb{N}.$$

Remark 3.3. By applying Markov's inequality to the $O(1/k)$ expectation bound of Theorem 3.6 and invoking the Borel–Cantelli lemma, one obtains deterministic (albeit sparse) subsequences along which almost sure polynomial rates arbitrarily close to $1/k$ hold. Concretely, for any $\epsilon > 0$ the deterministic subsequence $k_n(\epsilon) := \lfloor n^{2/\epsilon} \rfloor + 1$ satisfies, for almost every ω, the bound

$$F(Y(\theta_{k_n(\epsilon)}))(\omega) - F(Y^*) \le C(\omega,\epsilon) k_n(\epsilon)^{-1+\epsilon}, \quad \forall n \in \mathbb{N}.$$

for some path-dependent constant $C(\omega,\epsilon)$. This subsequence construction is mainly of theoretical interest; nevertheless, it offers a heuristic for practical verification of near-optimal iterates in long runs. Full discussion and proof are in Appendix A.9.

4 The local convergence for the cached surrogate gradient

In this section we show that approximating the ideal per-batch gradient by the cached surrogate gradient preserves the same local convergence behavior derived in Section 3. Specifically, we prove that the cached surrogate gradient inherits the same linear and $O(1/k)$ sublinear convergence guarantees, albeit under slightly more restrictive step-size conditions. The section is organized in two parts: Section 4.1 introduces additional notation and standing assumptions; Section 4.2 states and proves the corresponding convergence results.

4.1 Notation and standing assumptions

We adopt the notation and assumptions from Section 3 (subsections 3.1 and 2.3). Let θ_k denote the parameters at iteration k, and, for brevity, set

$$\theta \coloneqq \theta_k, \quad \theta^+ \coloneqq \theta_{k+1}, \quad \theta^- \coloneqq \theta_{k-1},$$

and correspondingly

$$Y \coloneqq Y(\theta), \quad Y^+ \coloneqq Y(\theta^+), \quad Y^- \coloneqq Y(\theta^-).$$

Furthermore, we have $\tilde{Y}$ at step k and $\tilde{Y}^+$ at step $k+1$ satisfying

$$(\tilde{Y})_n = \begin{cases} (Y)_n & \text{if } n \in \mathcal{B}^- \\ (\tilde{Y}^-)_n & \text{if } n \notin \mathcal{B}^- \end{cases}$$

and

$$(\tilde{Y}^+)_n = \begin{cases} (Y^+)_n & \text{if } n \in \mathcal{B} \\ (\tilde{Y})_n & \text{if } n \notin \mathcal{B} \end{cases},$$

where $\mathcal{B}^-$ is the mini-batch sampled at the step following θ^-, and $\mathcal{B}$ is the mini-batch sampled at the step following θ.

For notational convenience define

$$p_i = \left((\widetilde{\nabla_Y F})_i \cdot \nabla_\theta Y_i \right)^\top \in \mathbb{R}^s, \quad \widetilde{\nabla_Y F} = \widetilde{\nabla_Y F}(\tilde{Y}, Y).$$

Then $H^\top = \sum_{i \in \mathcal{B}} p_i$ and with fixed step size $\eta > 0$ the parameter update is

$$\theta^+ = \theta - \eta \cdot H^\top = \theta - \eta \cdot \sum_{i \in \mathcal{B}} p_i,$$

As in Section 3, we assume all mini-batches have a common size $b = |\mathcal{B}|$.

Treating $\widetilde{\nabla_Y F}(\tilde{Y}, Y)$ as a function of $\tilde{Y}$ with Y held fixed, we consider its first-order approximation. Using the integral form of Taylor's theorem, we have

$$\widetilde{\nabla_Y F}(\tilde{Y}, Y) = \widetilde{\nabla_Y F}(Y, Y) + \int_0^1 \mathcal{T}(Y + t(\tilde{Y} - Y))[\tilde{Y} - Y] \mathrm{d}t = \nabla_Y F + \int_0^1 \mathcal{T}(Y + t(\tilde{Y} - Y))[\tilde{Y} - Y] \mathrm{d}t,$$

where for each Θ, $\mathcal{T}(\Theta)$ is a fourth-order tensor and $\mathcal{T}(\Theta)[\Delta]$ denotes the contraction of this tensor with the matrix $\Delta = \tilde{Y} - Y$ (i.e. the linear operator induced by the tensor acting on matrices).

Complementing Assumptions A_1-A_5 established from Section 3, we introduce the following additional standing assumption for the analysis in this section:

- (A_6) The tensor operator norm (Frobenius-induced) of $\mathcal{T}$ is locally uniformly bounded: there exists $B_3 > 0$ such that for all relevant Θ,

  $$\|\mathcal{T}(\Theta)\|_{\mathrm{F}} \leq B_3.$$

  In particular, this bound controls the magnitude of the linearization error $\mathcal{T}(\Theta)[\tilde{Y} - Y]$ via

  $$\|\mathcal{T}(\Theta)[\tilde{Y} - Y]\|_{\mathrm{F}} \leq B_3 \|[\tilde{Y} - Y]\|_{\mathrm{F}}.$$

4.2 Main theorems and lemmas

The next group of lemmas are analogues of Lemmas 3.1–3.3 for the surrogate gradient $\widetilde{\nabla_Y F}$ and the corresponding batch quantity H. Full details are given in Appendix A.10–A.15. Our main convergence results for the surrogate gradient are as follows. Theorem 4.7 establishes linear convergence under local strong convexity, and Theorem 4.9 establishes $O(1/k)$ sublinear convergence under local convexity.

Lemma 4.1. *If Assumption A_3 holds, then $\|H\|_{\mathrm{F}} \leq B_1 \|\widetilde{\nabla_Y F}\|_{\mathrm{F}}$.*

Lemma 4.2. *Using c_2 from Lemma 3.2, if Assumptions A_1-A_4 hold, we similarly have*

$$F(Y^+) \leq F(Y) - \eta \cdot \sum_i \langle z_i, H^\top \rangle + c_2 \cdot \eta^2 \|\widetilde{\nabla_Y F}\|_{\mathrm{F}}^2,$$

where $z_i = ((\nabla_Y F)_i \cdot \nabla_\theta Y_i)^\top$ is as defined in Section 3.1.

Lemma 4.3. *If Assumptions A_1-A_4 hold, then for any constant $s > 0$, we have*

$$\mathbb{E}\big[\|\tilde{Y}^+ - Y^+\|_{\mathrm{F}}^2\big] \leq (1 + \frac{1}{s}) \left(1 - \frac{|\mathcal{B}|}{N}\right) \|\tilde{Y} - Y\|_{\mathrm{F}}^2 \ + \eta^2 \cdot (1+s) B_1^4 \|\widetilde{\nabla_Y F}\|_{\mathrm{F}}^2.$$

Lemma 4.4. *If Assumptions A_1-A_4 hold, denote*

$$\mu := \sqrt{1 - \frac{|\mathcal{B}|}{N}} \in (0,1).$$

Then for every $k \geq 1$,

$$\mathbb{E}\|\tilde{Y}(\theta_k) - Y(\theta_k)\|_{\mathrm{F}}^2 \leq B_1^4 \cdot \frac{\eta^2}{\mu(1-\mu)} \sum_{i=1}^{k-1} \mu^i \mathbb{E}\|\widetilde{\nabla_Y F}\left(\tilde{Y}(\theta_{k-i}), Y(\theta_{k-i})\right)\|_{\mathrm{F}}^2,$$

Lemma 4.5. *If Assumptions A_1, A_2, A_3, A_4 and A_6 hold, denote*

$$c_4 := 4B_1^2 L(1 + B_1^2 L) \quad \textit{and} \quad \mu := \sqrt{1 - \frac{|\mathcal{B}|}{N}} \in (0,1)$$

while the step size satisfies

$$\eta < \min\left\{ \frac{1}{c_4} \cdot \frac{|\mathcal{B}|}{2N},\ \frac{\sqrt{2\mu(1-\mu)}|\mathcal{B}|}{8B_1^2 B_3 (2N - |\mathcal{B}|)},\ 1 \right\},$$

for every iteration k, we have

$$\frac{1}{4}\mathbb{E}\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2 \leq \mathbb{E}\|\widetilde{\nabla_Y F}\left(\tilde{Y}(\theta_k), Y(\theta_k)\right)\|_{\mathrm{F}}^2 \leq 4\mathbb{E}\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2, \tag{4.1}$$

$$(1 - c_4 \cdot \eta)\mathbb{E}\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2 \leq \mathbb{E}\|\nabla_Y F(Y(\theta_{k+1}))\|_{\mathrm{F}}^2 \leq (1 + c_4 \cdot \eta)\mathbb{E}\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2. \tag{4.2}$$

Lemma 4.6. *If Assumptions A_1-A_6 hold, denote*

$$\nu := \frac{\mu}{1 - c_4 \cdot \eta} \quad \textit{and} \quad c_5 := 4c_2 + \frac{|\mathcal{B}|}{2N} B_1^4 + \frac{2|\mathcal{B}| \cdot B_1^4 B_3^2}{N\mu(1-\mu)(1-\nu)},$$

then $\nu < 1$. Moreover, while

$$\eta < \min\left\{ \frac{1}{c_4} \cdot \frac{|\mathcal{B}|}{2N},\ \frac{\sqrt{2\mu(1-\mu)}|\mathcal{B}|}{8B_1^2 B_3 (2N - |\mathcal{B}|)},\ 1,\ \frac{|\mathcal{B}|}{c_5 N} \lambda_{\min} \right\},$$

there exists a constant $c_6 > 0$, such that

$$\mathbb{E}\big[F(Y(\theta_{k+1}))\big] \le \mathbb{E}\big[F(Y(\theta_k))\big] - c_6 \cdot \eta \mathbb{E}\big[\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2\big].$$

Specifically, one may take $c_6 = \Big(\frac{|\mathcal{B}|}{N}\lambda_{\min} - c_5\eta\Big)$.

Remark 4.1. By the same summation argument used in Remark 3.2, the descent inequality of Lemma 4.6 implies

$$\min_{k=0,\cdots,T} \mathbb{E}\big[\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}\big] \le \sqrt{\frac{F(Y(\theta_0))}{c_6 \cdot \eta \cdot (T+1)}} = O(T^{-1/2}).$$

Note that the constant here is c_6 (given in Lemma 4.6) and differs from the earlier constant c_3 in Remark 3.2; c_6 depends explicitly on the minibatch size $|\mathcal{B}|$, the step-size η, and the model constants (see Lemma 4.6).

Theorem 4.7. *Suppose F is locally strongly convex around Y^* with strong convexity constant $c_Y > 0$. If Assumptions A_1-A_6 hold and the step size satisfies*

$$\eta < \min\left\{\frac{1}{c_4}\cdot\frac{|\mathcal{B}|}{2N},\ \frac{\sqrt{2\mu(1-\mu)}|\mathcal{B}|}{8B_1^2B_3(2N-|\mathcal{B}|)},\ 1,\ \frac{|\mathcal{B}|}{c_5N}\lambda_{\min}\right\},$$

then while the iterates remain in a neighborhood of Y^* where the local assumptions hold, the expected optimality gap converges linearly to zero: there exist constants $C > 0$ and $\rho \in (0,1)$ such that

$$\mathbb{E}\big[F(Y(\theta_k))\big] - F(Y^*) \le C\rho^k,$$

where $\mathbb{E}\big[F(Y(\theta_k))\big]$ denotes $\mathbb{E}_{\mathcal{B}_0,\cdots,\mathcal{B}_{k-1}}\big[F(Y(\theta_k))\big]$ and Y^* is the minimum point of F.

Proof. Proof of Theorem 4.7 proceeds identically to that of Theorem 3.4, replacing c_3 with c_6 and appealing to Lemma 4.6 instead of Lemma 3.3. We omit the details. □

Corollary 4.8. *Under the assumptions of Theorem 4.7, while the iterates remain in the neighborhood of Y^*, the optimality gap $F(Y(\theta_k)) - F(Y^*)$ converges to zero almost surely at a linear rate. More precisely, for almost every sample path ω, there exist $\rho(\omega) \in (0,1)$ and $k_0(\omega) \in \mathbb{N}$ such that*

$$F(Y(\theta_k)) - F(Y^*) \le \rho(\omega)^k, \quad \forall k \ge k_0(\omega).$$

Proof. Proof of Corollary 4.8 proceeds identically to that of Theorem 3.5, appealing to Lemma 4.6 and Theorem 4.7 instead of Lemma 3.3 and Theorem 3.4. We omit the details. □

Theorem 4.9. *Suppose F is locally convex around Y^*. If Assumptions A_1-A_6 hold and the step size satisfies*

$$\eta < \min\left\{\frac{1}{c_4}\cdot\frac{|\mathcal{B}|}{2N},\ \frac{\sqrt{2\mu(1-\mu)}|\mathcal{B}|}{8B_1^2B_3(2N-|\mathcal{B}|)},\ 1,\ \frac{|\mathcal{B}|}{c_5N}\lambda_{\min}\right\},$$

then while the iterates remain in a neighborhood of Y^* where the local assumptions hold, the expected optimality gap converges sublinearly to zero at a rate of $O(1/k)$: there exists a constant $C > 0$,

$$\mathbb{E}\big[F(Y(\theta_k))\big] - F(Y^*) \le \frac{C}{k},$$

where $\mathbb{E}\big[F(Y(\theta_k))\big]$ denotes $\mathbb{E}_{\mathcal{B}_0,\cdots,\mathcal{B}_{k-1}}\big[F(Y(\theta_k))\big]$ and Y^* is a minimum point of F.

Proof. Proof of Theorem 4.9 proceeds identically to that of Theorem 3.6, replacing c_3 with c_6 and appealing to Lemma 4.6 instead of Lemma 3.3. We omit the details. □

Corollary 4.10. *Under the assumptions of Theorem 4.9, while the iterates remain in the neighborhood of Y^*, the following hold:*

- *$F(Y(\theta_k)) - F(Y^*)$ converges to zero almost surely.*
- *The sequence $F(Y(\theta_k))$ admits almost surely convergent random subsequences at rate of $O(1/k)$, i.e., for almost every sample path ω, there exist a subsequence $k_n(\omega)$ and a constant $C(\omega)$ such that*

$$F(Y(\theta_{k_n})) - F(Y^*) \leq \frac{C(\omega)}{k_n}, \quad \forall\ n \in \mathbb{N}.$$

Proof. Proof of Corollary 4.10 proceeds identically to that of Theorem 3.7, appealing to Lemma 4.6 and Theorem 4.9 instead of Lemma 3.3 and Theorem 3.6. We omit the details. □

5 Numerical experiment

We conduct two numerical experiments to validate the main theoretical claims of the paper:

- (i) The effect of local convexity on convergence rates: we demonstrate the stark contrast between linear (exponential) convergence under a locally strongly convex objective and sublinear (polynomial) convergence under a convex but non-strongly-convex objective.
- (ii) For sample-coupled objectives the proposed cached surrogate gradient ('H') effectively approximates the ideal per-batch gradient ('G') and significantly outperforms the naive batch-local estimator.

Below we summarize the settings and the numerical results, and then discuss the relation between the numerical results and our theoretical results.

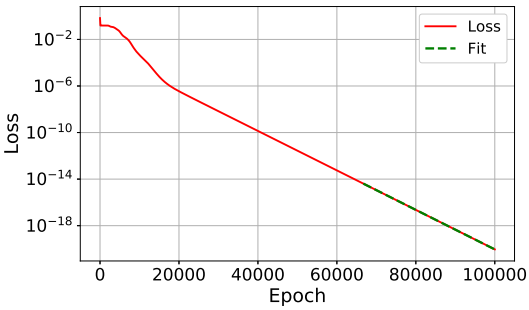
5.1 Experiment 1: Convergence Rates under Different Convexity Assumptions

We examine the convergence behavior under two convexity regimes using the same model architecture (a small three-layer MLP with input $4 \to 64 \to 64 \to$ output 2) and synthetic data ($N = 10$) generated by independent uniform sampling: each input x_i is drawn uniformly from $[0, 1]^4$ and each target y_i is drawn independently and uniformly from $[0, 1]^2$. Training is performed with sample-wise updates (mini-batch size 1), and each experiment is repeated 5 times with different random seeds; reported curves are averaged over runs. We compare two loss functions:

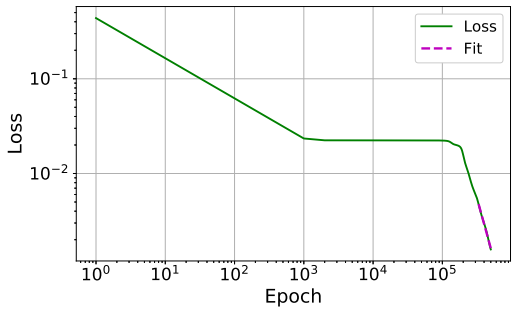
- *Strongly Convex Case:* mean squared error ($\sum(\hat{y} - y)^2$), trained with a fixed learning rate $\eta = 0.01$ for 10^5 epochs.
- *Non-Strongly Convex Case:* fourth-power loss ($\sum(\hat{y} - y)^4$), trained with a fixed learning rate $\eta = 0.001$ for 5×10^5 epochs.

The results are summarized in Fig. 5.1.

The results validate our theoretical analysis: the strongly convex MSE objective drives linear convergence, evidenced by the straight-line trend on a semilog plot. In contrast, the convex but non-strongly-convex fourth-power objective results in sublinear convergence, following a clear polynomial law on a log–log plot. The fitted exponents confirm the problem-dependent nature of the convergence rate, which is faster than the worst-case $O(1/k)$ bound but remains sublinear.



(a) MSE: exponential decay.



(b) Fourth-power: polynomial decay.

Figure 5.1: Convergence contrast under different convexity assumptions. (a) The MSE loss exhibits linear (exponential) convergence. A late-stage linear fit on log(loss) yields a slope of $\approx -2.96 \times 10^{-4} (R^2 \approx 0.999)$, confirming $F_k \propto 0.9996^k$. (b) The fourth-power loss exhibits sub-linear (polynomial) convergence. A late-stage log–log fit yields a slope of ≈ -2.65 ($R^2 \approx 0.995$), confirming $F_k \propto k^{-2.65}$.

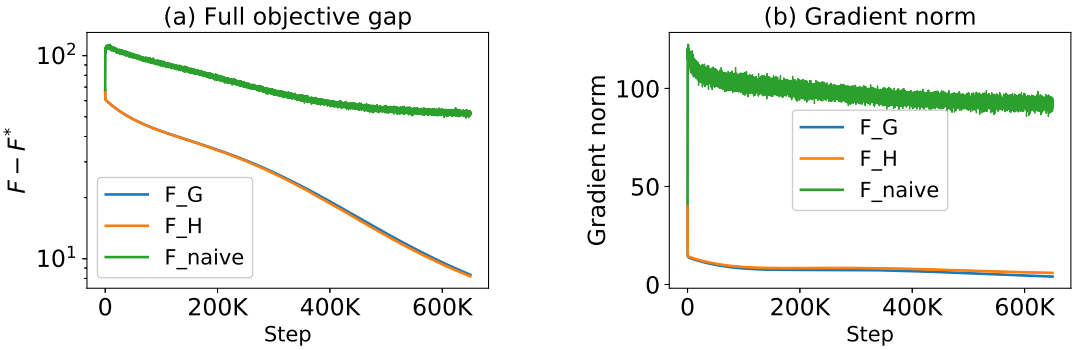


Figure 5.2: Full objective gap $F(Y) - F^*$ and gradient norm $\|\nabla_Y F\|_{\mathrm{F}}$ (log scale) for the three schemes $G/H/naive$. This composite figure shows the evolution of the full objective gap and the corresponding gradient magnitude over training.

5.2 Experiment 2: Gram objective

We consider the coupled-sample objective

$$F(Y) = \frac{1}{2}\|\frac{1}{d}YY^\top - S\|_{\mathrm{F}}^2,$$

with $d = 8$, $Y \in \mathbb{R}^{200\times 8}$, batch size 16. We parameterize each row $y_i^\top$ of Y by a single-hidden-layer MLP ($32 \to 128 \to 8$) with LeakyReLU activation, mapping fixed random inputs $x_i \in \mathbb{R}^{32}$. We compare three update schemes executed in parallel from the same randomized initialization: (i) 'G' – ideal per-batch rows of the full $\nabla_Y F$; (ii)'H' – cached surrogate using $\tilde{Y}$ and incremental Gram updates; (iii) *naive* – batch-local computation restricted to the batch. For each scheme we record the full objective $F(Y)$ and the Frobenius norm $\|\nabla_Y F\|_{\mathrm{F}}$ and perform 5 independent runs and report averaged curves. See Fig. 5.2 and Fig. 5.3.

The experiments show that the ideal method 'G' and the cached surrogate 'H' produce nearly indistinguishable reductions in the full objective F, and both substantially outperform the naive

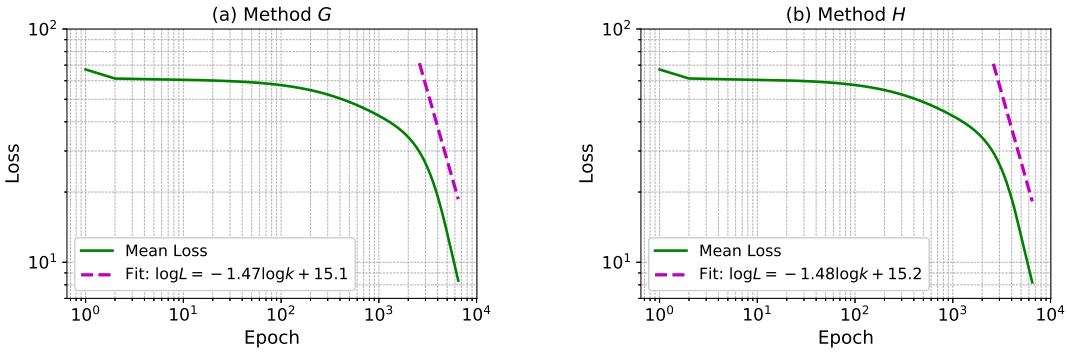


Figure 5.3: Diagnostics for methods G (left) and H (right). Log–log plot of training loss versus epochs. Late-stage log–log fits yield slopes ≈ -1.467 ($R^2 \approx 0.9854$) for G and ≈ -1.483 ($R^2 \approx 0.9871$) for H, indicating $F_k \propto k^{-1.467}$ and $F_k \propto k^{-1.483}$ respectively.

estimator; the gradient-norm plots give the same ordering. Furthermore, in the experiments, late-stage log–log diagnostics reveal sublinear decay of the gap $F - F^*$, consistent with the $O(1/k)$ convergence behavior established theoretically.

Collectively, the two experiments confirm the main theoretical messages: locally strongly convex losses admit exponential (linear) convergence under fixed-step updates; convex but non-strongly convex losses degrade to sublinear decay; and for sample-coupled objectives the cached-surrogate 'H' provides a practical, low-cost route to approximate the ideal per-batch gradient 'G' and achieve markedly improved convergence relative to naive batch-local estimators. While the experiments here are synthetic and intended to isolate theoretical phenomena, comparable empirical evaluations on real datasets are reported in SpecNet2 [41].

6 Conclusion

We propose an SGD-style framework for sample-coupled objectives that uses two batch-gradient constructs: the ideal per-batch gradient 'G', and a cached-surrogate gradient 'H' that approximates 'G' when full-data quantities are costly to compute. Our main contribution is a unified local convergence theory: under mild smoothness and Jacobian-boundedness assumptions (A_1-A_6) we develop a unified local convergence theory showing that both 'G'-driven and 'H'-driven updates enjoy the same qualitative regimes: linear convergence under local strong convexity, and sublinear convergence under mere local convexity ($O(1/k)$ in expectation). Controlled experiments corroborate the theoretical regimes and show that 'H' closely tracks 'G' in practice, while incurring only modest additional computation or storage compared with standard mini-batch SGD. Future work includes sharpening quantitative trade-offs (cache staleness, approximation error, batch size), extending the analysis beyond local neighborhoods toward global guarantees, and adapting the surrogate idea to momentum and adaptive optimizers.

A Proofs and Detailed Derivations

A.1 Proof of Lemma 3.1

Proof. By the Cauchy–Schwarz inequality,

$$\begin{aligned}
\|G\|_{\mathrm{F}}^2 &= \Big\|\sum_{i\in\mathcal{B}} z_i\Big\|_{\mathrm{F}}^2 = \Big\|\sum_{i\in\mathcal{B}} (\nabla_Y F)_i \cdot \nabla_\theta Y_i\Big\|_{\mathrm{F}}^2 \\
&\le \Big(\sum_{i\in\mathcal{B}} \|(\nabla_Y F)_i \cdot \nabla_\theta Y_i\|_{\mathrm{F}}\Big)^2 \\
&\le \Big(\sum_{i\in\mathcal{B}} \|(\nabla_Y F)_i\|_2 \, \|\nabla_\theta Y_i\|_{\mathrm{F}}\Big)^2 \\
&\le \Big(\sum_{i\in\mathcal{B}} \|(\nabla_Y F)_i\|_2^2\Big)\Big(\sum_{i\in\mathcal{B}} \|\nabla_\theta Y_i\|_{\mathrm{F}}^2\Big) \\
&\le B_1^2 \, \|\nabla_Y F\|_{\mathrm{F}}^2.
\end{aligned}$$

□

A.2 Proof of Lemma 3.2

Proof. By the Taylor expansion with integral remainder, we have

$$Y^+ - Y = \nabla_\theta Y(\theta)(\theta^+ - \theta) + R$$

where

$$R = \int_0^1 (1-t)\nabla_\theta^2 Y(\theta + t(\theta^+ - \theta))[\theta^+ - \theta, \theta^+ - \theta]\,\mathrm{d}t,$$

and $\nabla_\theta^2 Y$ is the second-order derivative tensor of Y with respect to θ. Then, by taking the Frobenius norm and Assumption A_4 on the line segment between θ and θ^+, we obtain

$$\|R\|_{\mathrm{F}} \le \frac{B_2}{2}\|\theta^+ - \theta\|_{\mathrm{F}}^2 = \frac{B_2}{2} \cdot \eta^2 \|G\|_{\mathrm{F}}^2.$$

Similarly, for the first-order difference, we have

$$\|Y^+ - Y\|_{\mathrm{F}} \le \sup_{t\in[0,1]} \|\nabla_\theta Y(\theta + t(\theta^+ - \theta))\|_{\mathrm{F}} \cdot \|\theta^+ - \theta\|_{\mathrm{F}} \le B_1 \cdot \eta \|G\|_{\mathrm{F}}.$$

Since $\nabla_Y F$ is L-Lipschitz by Assumption A_2 (i.e. F is L-smooth in Y), we have the standard quadratic upper bound

$$F(Y^+) \;\le\; F(Y) + \langle \nabla_Y F, Y^+ - Y\rangle + \frac{L}{2}\|Y^+ - Y\|_{\mathrm{F}}^2.$$

Substituting the per-row decomposition and summing over $i = 1, \cdots, N$ gives

$$
\begin{aligned}
F(Y^+) &\le F(Y) + \langle \nabla_Y F, Y^+ - Y \rangle + \frac{L}{2}\|Y^+ - Y\|_{\mathrm{F}}^2 \\
&= F(Y) + \operatorname{tr}\left((Y^+ - Y)^\top (\nabla_Y F)\right) + \frac{L}{2}\|Y^+ - Y\|_{\mathrm{F}}^2 \\
&= F(Y) + \operatorname{tr}\left((\nabla_\theta Y(\theta)(\theta^+ - \theta) + R)^\top (\nabla_Y F)\right) + \frac{L}{2}\|Y^+ - Y\|_{\mathrm{F}}^2 \\
&= F(Y) + \operatorname{tr}\left(\sum_{i=1}^N \nabla_\theta Y_i(\theta)(\theta^+ - \theta)(\nabla_Y F)_i\right) + \langle \nabla_Y F, R \rangle + \frac{L}{2}\|Y^+ - Y\|_{\mathrm{F}}^2 \\
&\le F(Y) - \eta \cdot \operatorname{tr}\Big(\sum_{i=1}^N \nabla_\theta Y_i(\theta) \cdot G^\top (\nabla_Y F)_i\Big) + \left(\|\nabla_Y F\|_2 \cdot \frac{B_2}{2} + \frac{L}{2} B_1^2\right) \cdot \eta^2 \|G\|_{\mathrm{F}}^2 \\
&= F(Y) - \eta \cdot \sum_{i=1}^N \operatorname{tr}\big(\left((\nabla_Y F)_i \cdot \nabla_\theta Y_i\right) \cdot G^\top\big) + \left(\|\nabla_Y F\|_{\mathrm{F}} \cdot \frac{B_2}{2} + \frac{L}{2} B_1^2\right) \cdot \eta^2 \|G\|_{\mathrm{F}}^2 \\
&= F(Y) - \eta \cdot \sum_{i=1}^N \langle z_i, G^\top \rangle + \frac{1}{2}\left(B_0 B_2 + L B_1^2\right) \cdot \eta^2 \|G\|_{\mathrm{F}}^2 \\
&\le F(Y) - \eta \cdot \sum_{i=1}^N \langle z_i, G^\top \rangle + \frac{1}{2}\left(B_0 B_2 + L B_1^2\right) c_1 \cdot \eta^2 \|\nabla_Y F\|_{\mathrm{F}}^2,
\end{aligned}
$$

where we denote $c_1 = B_1^2$ and then $\|G\|_{\mathrm{F}}^2 \le c_1 \|\nabla_Y F\|_{\mathrm{F}}^2$. Let $c_2 = \frac{1}{2}\left(B_0 B_2 + L B_1^2\right) c_1$. Then

$$
F(Y^+) \le F(Y) - \eta \cdot \sum_{i=1}^N \langle z_i, G^\top \rangle + c_2 \cdot \eta^2 \|\nabla_Y F\|_{\mathrm{F}}^2.
$$

□

A.3 Proof of Lemma 3.3

Proof. Starting from Lemma 3.2 we have the deterministic bound

$$
F(Y^+) \le F(Y) - \eta \cdot \sum_{i=1}^N \langle z_i, G^\top \rangle + c_2 \cdot \eta^2 \|\nabla_Y F\|_{\mathrm{F}}^2.
$$

Take conditional expectation with respect to the current mini-batch $\mathcal{B}$. Under the usual uniform sampling assumption where each index is included in $\mathcal{B}$ with probability $|\mathcal{B}|/N$ (this holds for both sampling with replacement and for sampling without replacement under the usual i.i.d. or exchangeable data assumptions), the conditional expectation of $G^\top$ satisfies

$$
\mathbb{E}_{\mathcal{B}}[G^\top] = \frac{|\mathcal{B}|}{N} \sum_i^N z_j.
$$

Hence

$$\begin{aligned}
\mathbb{E}_{\mathcal{B}}\big[F(Y^+)\big] &\le \mathbb{E}_{\mathcal{B}}\big[F(Y) - \eta\cdot\sum_{i=1}^{N}\langle z_i, G^\top\rangle + c_2\cdot\eta^2\|\nabla_Y F\|_{\mathrm{F}}^2\big] \\
&= F(Y) - \eta\cdot\frac{|\mathcal{B}|}{N}\cdot\sum_{i=1}^{N}\sum_{j}^{N}\langle z_i, z_j\rangle + c_2\cdot\eta^2\|\nabla_Y F\|_{\mathrm{F}}^2 \\
&= F(Y) - \eta\cdot\frac{|\mathcal{B}|}{N}\cdot\sum_{i=1}^{N}\sum_{j}^{N}(\nabla_Y F)_i\cdot\nabla_\theta Y_i\cdot\nabla_\theta Y_j^\top\cdot(\nabla_Y F)_j^\top + c_2\cdot\eta^2\|\nabla_Y F\|_{\mathrm{F}}^2 \\
&= F(Y) - \eta\cdot\frac{|\mathcal{B}|}{N}\cdot\hat{g}\big(\nabla_\theta Y\cdot(\nabla_\theta Y)^\top\big)\hat{g}^\top + c_2\cdot\eta^2\|\nabla_Y F\|_{\mathrm{F}}^2,
\end{aligned}$$

where $\hat{g} = \big((\nabla_Y F)_1, (\nabla_Y F)_2, \cdots, (\nabla_Y F)_N\big) \in \mathbb{R}^{1\times Nd}$ is the row vector formed by horizontally concatenating $(\nabla_Y F)_i$, then since Assumption A_5 holds, we have

$$\hat{g}\big(\nabla_\theta Y\cdot(\nabla_\theta Y)^\top\big)\hat{g}^\top \ge \lambda_{\min}\|\hat{g}\|_2^2.$$

Noting that $\|\hat{g}\|_2^2 = \sum_{i=1}^{N}\|(\nabla_Y F)_i\|_2^2 = \|\nabla_Y F\|_{\mathrm{F}}^2$, we obtain

$$\begin{aligned}
\mathbb{E}_{\mathcal{B}}\big[F(Y^+)\big] &\le F(Y) - \eta\cdot\frac{|\mathcal{B}|}{N}\cdot\hat{g}\big(\nabla_\theta Y\cdot(\nabla_\theta Y)^\top\big)\hat{g}^\top + c_2\cdot\eta^2\|\nabla_Y F\|_{\mathrm{F}}^2 \\
&\le F(Y) - \eta\cdot\frac{|\mathcal{B}|}{N}\cdot\lambda_{\min}\|\nabla_Y F\|_{\mathrm{F}}^2 + c_2\cdot\eta^2\|\nabla_Y F\|_{\mathrm{F}}^2 \\
&= F(Y) - \eta\cdot\big[\frac{|\mathcal{B}|}{N}\lambda_{\min} - c_2\eta\big]\|\nabla_Y F\|_{\mathrm{F}}^2.
\end{aligned}$$

Let $c_3 = \frac{|\mathcal{B}|}{N}\lambda_{\min} - c_2\eta$, since $\eta < \frac{|\mathcal{B}|}{N}\cdot\frac{\lambda_{\min}}{c_2}$, we have $c_3 > 0$, i.e., there exists a constant $c_3 > 0$, such that

$$\mathbb{E}_{\mathcal{B}}[F(Y^+)] \le F(Y) - c_3\cdot\eta\|\nabla_Y F\|_{\mathrm{F}}^2.$$

□

A.4 Detailed Derivation for Remark 3.2

Summing the inequality from Lemma 3.3 over $k = 0, \cdots, T$ and taking total expectation gives

$$\begin{aligned}
\sum_{k=0}^{T}\mathbb{E}\big[\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2\big] &\le \frac{1}{c_3\cdot\eta}\cdot\sum_{k=0}^{T}\Big(\mathbb{E}\big[F(Y(\theta_k))\big] - \mathbb{E}\big[F(Y(\theta_{k+1}))\big]\Big) \\
&= \frac{1}{c_3\cdot\eta}\cdot\Big(F(Y(\theta_0)) - \mathbb{E}\big[F(Y(\theta_{T+1}))\big]\Big) \\
&\le \frac{1}{c_3\cdot\eta}\cdot F(Y(\theta_0)),
\end{aligned}$$

where for brevity we write $\mathbb{E}\big[\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2\big]$ as a shorthand for $\mathbb{E}\big[\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2|\mathcal{B}_0, \cdots, \mathcal{B}_{k-1}\big]$ and similarly $\mathbb{E}\big[F(Y(\theta_k))\big]$ denotes $\mathbb{E}_{\mathcal{B}_0,\cdots,\mathcal{B}_{k-1}}\big[F(Y(\theta_k))\big]$. Hence

$$\min_{k=0,\cdots,T}\mathbb{E}\big[\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2\big] \le \frac{F(Y(\theta_0))}{c_3\cdot\eta\cdot(T+1)},$$

It follows that the expected squared Frobenius norm of the gradient decays as $O(1/T)$, Consequently, the expected Frobenius norm of the gradient satisfies

$$\min_{k=0,\cdots,T} \mathbb{E}\big[\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}\big] \leq \sqrt{\frac{F(Y(\theta_0))}{c_3 \cdot \eta \cdot (T+1)}} = O(T^{-1/2}).$$

Therefore, for the nonconvex case we obtain the standard stochastic stationarity guarantee: the gradient norm vanishes in expectation at rate $O(T^{-1/2})$. This is a weak stationarity result and does not affect the stronger convergence claims derived under convexity.

A.5 Proof of Theorem 3.4

Proof. By local strong convexity around Y^* we have the standard inequality

$$\|\nabla_Y F(Y(\theta))\|_{\mathrm{F}}^2 \geq 2c_Y\big(F(Y(\theta)) - F(Y^*)\big),$$

where Y^* is the minimum point of F. Then, by Lemma 3.3 and by applying the strong-convexity inequality to the conditional expectation of the squared gradient, we get

$$\begin{aligned}
&\mathbb{E}_{\mathcal{B}_0,\cdots,\mathcal{B}_k}\big[F(Y(\theta_{k+1})) - F(Y^*)\big] \\
=&\mathbb{E}_{\mathcal{B}_0,\cdots,\mathcal{B}_{k-1}}\mathbb{E}_{\mathcal{B}_k}\big[F(Y(\theta_{k+1})) - F(Y^*) \mid \mathcal{B}_0,\cdots,\mathcal{B}_{k-1}\big] \\
\leq&\mathbb{E}_{\mathcal{B}_0,\cdots,\mathcal{B}_{k-1}}\big[F(Y(\theta_k)) - F(Y^*) - c_3\cdot\eta\|\nabla_Y F(\theta_k)\|_{\mathrm{F}}^2\big] \\
\leq&\mathbb{E}_{\mathcal{B}_0,\cdots,\mathcal{B}_{k-1}}\big[F(Y(\theta_k)) - F(Y^*) - c_3\cdot\eta\cdot 2c_Y\big(F(Y(\theta_k)) - F(Y^*)\big)\big] \\
\leq&(1-2c_3c_Y\eta)\mathbb{E}_{\mathcal{B}_0,\cdots,\mathcal{B}_{k-1}}\big[F(Y(\theta_k)) - F(Y^*)\big] \\
\leq&\cdots \\
\leq&(1-2c_3c_Y\eta)^{k+1}\cdot\big(F(Y(\theta_0)) - F(Y^*)\big).
\end{aligned}$$

Let $C = F(Y(\theta_0)) - F(Y^*)$ and $\rho = 1 - 2c_3c_Y\eta$, then

$$\mathbb{E}\big[F(Y(\theta_k))\big] - F(Y^*) \leq C\rho^k,$$

where $C > 0$, $\rho \in (0,1)$ and $\mathbb{E}\big[F(Y(\theta_k))\big]$ denotes $\mathbb{E}_{\mathcal{B}_0,\cdots,\mathcal{B}_{k-1}}\big[F(Y(\theta_k))\big]$, which means $\mathbb{E}\big[F(Y(\theta_k))\big] - F(Y^*)$ converges linearly to zero. □

A.6 Proof of Corollary 3.5

Proof. First note from Lemma 3.3 that, for the chosen η, there exists a deterministic constant $c_3 > 0$, such that

$$\mathbb{E}\big[F(Y(\theta_{k+1}))|\mathcal{F}_k\big] \leq F(Y(\theta_k)) - c_3\cdot\eta\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2,$$

where $\mathcal{F}_k$ is the filtration generated by $\mathcal{B}_0,\cdots,\mathcal{B}_k$. In particular,

$$\mathbb{E}\big[F(Y(\theta_{k+1}))|\mathcal{F}_k\big] \leq F(Y(\theta_k)).$$

Hence $F(Y(\theta_k))$ is a supermartingale with respect to $\mathcal{F}_k$. Since $F(Y(\theta_k)) \geq F(Y^*)$ and, under our assumptions, the random variables $F(Y(\theta_k))$ are integrable under our assumptions, Doob's supermartingale convergence theorem implies that $F(Y(\theta_k))$ converges almost surely to a finite random limit L_0. Note from Theorem 3.4 that $\mathbb{E}\big[F(Y(\theta_k))\big] \to F(Y^*)$ as $k \to \infty$. Since $F(Y(\theta_k)) \xrightarrow{\text{a.s.}} L_0$ and $F(Y(\theta_k)) \geq F(Y^*)$, Fatou's lemma gives

$$\mathbb{E}[L_0] \leq \liminf_{k\to\infty}\mathbb{E}\big[F(Y(\theta_k))\big] = F(Y^*).$$

But $L_0 \geq F(Y^*)$ almost surely, hence the only possibility is $L_0 = F(Y^*)$. Thus $F(Y(\theta_k))$ converges almost surely to $F(Y^*)$.

Next we upgrade the expected linear rate to an almost sure (pathwise) linear tail bound. From Theorem 3.4 there exist deterministic constants $C > 0$ and $\rho \in (0, 1)$ such that

$$\mathbb{E}\big[F(Y(\theta_k)) - F(Y^*)\big] \leq C\rho^k, \quad \forall k \geq 0,$$

Fix any $\epsilon > 0$ and define events:

$$B_k(\epsilon) \coloneqq \left\{F(Y(\theta_k)) - F(Y^*) > [\rho(1+\epsilon)]^k\right\}, \quad k = 1, 2, \ldots$$

By Markov's inequality and the expectation bound:

$$\begin{aligned}\mathbb{P}(B_k(\epsilon)) &= \mathbb{P}\bigg(F(Y(\theta_k)) - F(Y^*) > [\rho(1+\epsilon)]^k\bigg) \\ &\leq \frac{\mathbb{E}[F(Y(\theta_k)) - F(Y^*)]}{[\rho(1+\epsilon)]^k} \\ &\leq \frac{C\rho^k}{[\rho(1+\epsilon)]^k} = C(1+\epsilon)^{-k}.\end{aligned}$$

Let $\alpha \coloneqq (1+\epsilon)^{-1} \in (0, 1)$. The series $\sum_{k=1}^{\infty} \alpha^k$ converges, so:

$$\sum_{k=1}^{\infty} \mathbb{P}(B_k(\epsilon)) \leq C \sum_{k=1}^{\infty} \alpha^k < \infty.$$

By the Borel-Cantelli lemma, $\mathbb{P}\left(\limsup_{k\to\infty} B_k(\epsilon)\right) = 0$. Thus, a.s. there exists a random index $K_\epsilon \in \mathbb{N}$ (depending on the sample path) such that for all $k \geq K_\epsilon$:

$$F(Y(\theta_k)) - F(Y^*) \leq [\rho(1+\epsilon)]^k.$$

This implies that the convergence rate is linear with base $\rho(1+\epsilon)$ while $k \geq K_\epsilon$ for any $\epsilon > 0$. Equivalently, for almost every sample path ω, there exist $\rho(\omega) \in (\rho, 1)$ and $k_0(\omega) \in \mathbb{N}$, such that

$$F(Y(\theta_k)) - F(Y^*) \leq \rho(\omega)^k, \qquad \forall k \geq k_0(\omega).$$

□

A.7 Proof of Theorem 3.6

Proof. By assumption, for all $k \geq 0$, there exist $B > 0$ and a minimum point Y^* such that $Y(\theta_k \in B(Y^*, B)$. Then local convexity gives

$$0 \leq F(Y(\theta_k)) - F(Y^*) \leq \langle \nabla_Y F, Y(\theta_k) - Y^* \rangle \leq \|\nabla_Y F\|_{\mathrm{F}} \|Y(\theta_k) - Y^*\|_{\mathrm{F}} \leq B\|\nabla_Y F(\theta_k)\|_{\mathrm{F}}.$$

Taking expectations and setting $e_k = \mathbb{E}\big[F(Y(\theta_k)) - F(Y^*)\big]$ and $d_k = \mathbb{E}\|\nabla_Y F(\theta_k)\|_{\mathrm{F}}^2$, we get $e_k \leq B \cdot \mathbb{E}\|\nabla_Y F(\theta_k)\|_{\mathrm{F}} \leq B\sqrt{\mathbb{E}\|\nabla_Y F(\theta_k)\|_{\mathrm{F}}^2} = B\sqrt{d_k}$, hence $d_k \geq e_k^2/B^2$.

Next, Lemma 3.3 gives the single-step descent inequality

$$\mathbb{E}\big[F(Y(\theta_{k+1}))|\mathcal{F}_k\big] \leq F(Y(\theta_k)) - c_3 \cdot \eta \|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2,$$

where $\mathcal{F}_k$ is the filtration generated by $\mathcal{B}_0, \cdots, \mathcal{B}_k$. Taking total expectation and subtracting $F(Y^*)$ from both sides yields

$$e_{k+1} \leq e_k - c_3\eta d_k \leq e_k - \frac{c_3\eta}{B^2} e_k^2 \leq e_k - \frac{c_3\eta}{B^2} e_k e_{k+1}.$$

Hence

$$\frac{1}{e_{k+1}} - \frac{1}{e_k} \geq \frac{c_3\eta}{B^2}$$

and telescoping yields

$$e_k \leq \frac{B^2}{c_3\eta} \cdot \frac{1}{k}$$

for all $k \geq 1$. Then there exists a constant $C > 0$ such that

$$e_k \leq \frac{C}{k}, \quad \text{i.e.} \quad \mathbb{E}\big[F(Y(\theta_k))\big] - F(Y^*) \leq \frac{C}{k}.$$

□

A.8 Proof of Corollary 3.7

Proof. Let $\mathcal{F}_k = \sigma(\mathcal{B}_0, \ldots, \mathcal{B}_k)$. By Lemma 3.3 we have almost surely

$$\mathbb{E}\big[F(Y(\theta_{k+1}))|\mathcal{F}_k\big] \leq F(Y(\theta_k)) - c_3 \cdot \eta \|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2.$$

Dropping the nonpositive term on the right-hand side yields

$$\mathbb{E}\big[F(Y(\theta_{k+1}))|\mathcal{F}_k\big] \leq F(Y(\theta_k)) \quad \text{a.s.},$$

so $(F(Y(\theta_k)))k \geq 0$ is a supermartingale bounded below by $F(Y^*)$. Under Assumptions A_1-A_5 the random variables $F(Y(\theta_k))$ are integrable, hence Doob's supermartingale convergence theorem implies the existence of a finite random variable F_∞ such that

$$F(Y(\theta_k)) \to F_\infty \quad \text{a.s.}$$

By Theorem 3.6 we have $\mathbb{E}[F(Y(\theta_k))] \downarrow F(Y^*)$, therefore $F_\infty = F(Y^*)$ almost surely. This proves (i). Regarding (ii), set $e_k := \mathbb{E}[F(Y(\theta_k))] - F(Y^*) \geq 0$. By Theorem 3.6 there exists $C > 0$ such that $e_k \leq C/k$ for all k. Define the nonnegative random variables

$$X_k \coloneqq k\big(F(Y(\theta_k)) - F(Y^*)\big).$$

By Fatou's lemma and the bound on e_k we obtain

$$\mathbb{E}\big[\liminf_{k\to\infty} X_k\big] \leq \liminf_{k\to\infty} \mathbb{E}\big[X_k\big] = \liminf_{k\to\infty} k e_k \leq C.$$

Hence $\liminf_{k\to\infty} X_k < \infty$ almost surely. Consequently, for almost every ω there exists an increasing subsequence $k_n(\omega)$ along which $X_{k_n(\omega)}(\omega)$ is bounded; equivalently, there exists $C(\omega)$ such that

$$F(Y(\theta_{k_n})) - F(Y^*) \leq \frac{C(\omega)}{k_n(\omega)}, \quad \forall\, n \in \mathbb{N},$$

which means the sequence $F(Y(\theta_k))$ admits almost surely convergent random subsequences at rate of $O(1/k)$.

□

A.9 Detailed Derivation for Remark 3.3

Using Markov's inequality together with the $O(1/k)$ expectation bound from Theorem 3.6 and then applying the Borel-Cantelli lemma, one obtains deterministic (but very sparse) subsequences along which almost-sure polynomial rates arbitrarily close to $1/k$ hold. Concretely, for any $\epsilon > 0$ define the deterministic subsequence $k_n(\epsilon) := \lfloor n^{2/\epsilon} \rfloor + 1$. Then for almost every ω there exists a path-dependent constant $C(\omega, \epsilon) > 0$ such that

$$F(Y(\theta_{k_n(\epsilon)}))(\omega) - F(Y^*) \leq C(\omega, \epsilon) k_n(\epsilon)^{-1+\epsilon}, \quad \forall n \in \mathbb{N}.$$

This deterministic-subsequence statement is primarily of theoretical interest: the subsequence is sparse and the multiplicative constant $C(\omega, \epsilon)$ depends on the sample path. Nevertheless, it has a practical implication. Although the pathwise convergence rate for the full sequence remains unquantified, the guaranteed existence of almost-surely convergent subsequences ensures the asymptotic attainability of near-optimal solutions along predetermined iterates. In practice, one may exploit this fact by implementing checkpointing protocols that evaluate and store iterates at the prescribed deterministic indices $k_n(\epsilon)$; such checkpoints provide implementable verification mechanisms that, with probability one, eventually observe iterates achieving the stated polynomial rates. Thus, while the result is theoretically weaker than a uniform pathwise rate, it supplies an actionable strategy for empirical validation and for selecting iterates likely to be near-optimal in long runs.

A.10 Proof of Lemma 4.1

Proof. Recall $H = \sum_{i \in \mathcal{B}} p_i$ with $p_i = (\widetilde{\nabla_Y F})_i \cdot \nabla_\theta Y_i$. By the Cauchy–Schwarz inequality for sums,

$$\begin{aligned}
\|H\|_{\mathrm{F}}^2 &= \Big\| \sum_{i \in \mathcal{B}} p_i \Big\|_{\mathrm{F}}^2 \leq \Big(\sum_{i \in \mathcal{B}} \|p_i\|_{\mathrm{F}} \Big)^2 \\
&= \Big(\sum_{i \in \mathcal{B}} \|(\widetilde{\nabla_Y F})_i \cdot \nabla_\theta Y_i\|_{\mathrm{F}} \Big)^2 \\
&\leq \Big(\sum_{i \in \mathcal{B}} \|(\widetilde{\nabla_Y F})_i\|_2 \, \|\nabla_\theta Y_i\|_{\mathrm{F}} \Big)^2 \\
&\leq \Big(\sum_{i \in \mathcal{B}} \|(\widetilde{\nabla_Y F})_i\|_2^2 \Big) \Big(\sum_{i \in \mathcal{B}} \|\nabla_\theta Y_i\|_{\mathrm{F}}^2 \Big) \\
&\leq B_1^2 \, \|\widetilde{\nabla_Y F}\|_{\mathrm{F}}^2.
\end{aligned}$$

□

A.11 Proof of Lemma 4.2

Proof. By the Taylor expansion with integral remainder, we have

$$Y^+ - Y = \nabla_\theta Y(\theta)(\theta^+ - \theta) + R$$

where

$$R = \int_0^1 (1-t) \nabla_\theta^2 Y(\theta + t(\theta^+ - \theta))[\theta^+ - \theta, \theta^+ - \theta] \, \mathrm{d}t,$$

and $\nabla_\theta^2 Y$ is the second-order derivative tensor of Y with respect to θ. Then, by taking the Frobenius norm and Assumption A_4 on the line segment between θ and θ^+, we obtain

$$\|R\|_{\mathrm{F}} \leq \frac{B_2}{2} \|\theta^+ - \theta\|_{\mathrm{F}}^2 = \frac{B_2}{2} \cdot \eta^2 \|H\|_{\mathrm{F}}^2.$$

Similarly, for the first-order difference, we have

$$\|Y^+ - Y\|_{\mathrm{F}} \le \sup_{t\in[0,1]} \|\nabla_\theta Y(\theta + t(\theta^+ - \theta))\|_{\mathrm{F}} \cdot \|\theta^+ - \theta\|_{\mathrm{F}} \le B_1 \cdot \eta \|H\|_{\mathrm{F}}.$$

Since $\nabla_Y F$ is L-Lipschitz by Assumption A_2 (i.e. F is L-smooth in Y), we have the standard quadratic upper bound

$$F(Y^+) \;\le\; F(Y) + \langle \nabla_Y F, Y^+ - Y\rangle + \frac{L}{2}\|Y^+ - Y\|_{\mathrm{F}}^2.$$

Substituting the per-row decomposition and summing over $i = 1, \cdots, N$ gives

$$\begin{aligned}
F(Y^+) \;&\le F(Y) + \langle \nabla_Y F, Y^+ - Y\rangle + \frac{L}{2}\|Y^+ - Y\|_{\mathrm{F}}^2 \\
&= \; F(Y) + \operatorname{tr}\left((\nabla_\theta Y(\theta)(\theta^+ - \theta) + R)^\top (\nabla_Y F)\right) + \frac{L}{2}\|Y^+ - Y\|_{\mathrm{F}}^2 \\
&= F(Y) + \operatorname{tr}\left(\sum_{i=1}^N \nabla_\theta Y_i(\theta)(\theta^+ - \theta)(\nabla_Y F)_i\right) + \langle \nabla_Y F, R\rangle + \frac{L}{2}\|Y^+ - Y\|_{\mathrm{F}}^2 \\
&\le F(Y) - \eta \cdot \operatorname{tr}\Big(\sum_{i=1}^N \nabla_\theta Y_i(\theta) \cdot H^\top (\nabla_Y F)_i\Big) + \left(\|\nabla_Y F\|_{\mathrm{F}} \cdot \frac{B_2}{2} + \frac{L}{2}B_1^2\right) \cdot \eta^2 \|H\|_{\mathrm{F}}^2 \\
&= F(Y) - \eta \cdot \sum_{i=1}^N \operatorname{tr}\Big(\left((\nabla_Y F)_i \cdot \nabla_\theta Y_i\right) \cdot H^\top\Big) + \left(\|\nabla_Y F\|_{\mathrm{F}} \cdot \frac{B_2}{2} + \frac{L}{2}B_1^2\right) \cdot \eta^2 \|H\|_{\mathrm{F}}^2 \\
&= F(Y) - \eta \cdot \sum_{i=1}^N \langle z_i, H^\top\rangle + \frac{1}{2}\left(B_0 B_2 + L B_1^2\right) \cdot \eta^2 \|H\|_{\mathrm{F}}^2 \\
&\le F(Y) - \eta \cdot \sum_{i=1}^N \langle z_i, H^\top\rangle + c_2 \cdot \eta^2 \|\widetilde{\nabla_Y F}\|_{\mathrm{F}}^2,
\end{aligned}$$

where $z_i = \left((\nabla_Y F)_i \cdot \nabla_\theta Y_i\right)^\top$ is as defined in Section 3.1. □

A.12 Proof of Lemma 4.3

Proof. We sample a mini-batch $\mathcal{B}$ for the current update, which determines the batch gradient H and, in turn, $\tilde{Y}^+$ and Y^+, while $\tilde{Y}$ and Y remain fixed from the preceding step. Using the Frobenius-norm representation,

$$\|\tilde{Y}^+ - Y^+\|_{\mathrm{F}}^2 = \sum_{i\in\mathcal{B}^c} \|\tilde{Y}_i^+ - Y_i^+\|_{\mathrm{F}}^2.$$

For each $i \in \mathcal{B}^c$, we decompose

$$\tilde{Y}_i^+ - Y_i^+ = \tilde{Y}_i - Y_i^+ = (\tilde{Y}_i - Y_i) + (Y_i - Y_i^+).$$

Expanding the squared norm and applying the inequality $2\langle a, b\rangle \le s|a|^2 + \frac{1}{s}|b|^2$ (valid for any $s > 0$) yields

$$\|\tilde{Y}_i - Y_i^+\|_2^2 \le (1 + \frac{1}{s})\|\tilde{Y}_i - Y_i\|_2^2 + (1+s)\|Y_i - Y_i^+\|_2^2.$$

Let $\mathcal{F}$ denote the filtration generated by all randomness up to the current iterate. By the tower property, the total expectation decomposes as

$$\mathbb{E}[\cdot] = \mathbb{E}_{\mathcal{F}}\Big[\mathbb{E}_{\mathcal{B}}\big[\cdot \mid \mathcal{F}\big]\Big],$$

where the inner expectation is taken over the random mini-batch $\mathcal{B}$. For the single-step analysis, we focus on this inner expectation. Summing over $i \in \mathcal{B}^c$ and taking conditional expectation given $\mathcal{F}$ yields

$$\mathbb{E}_{\mathcal{B}}\Big[\|\tilde{Y}^+ - Y^+\|_{\mathrm{F}}^2 \;\Big|\; \mathcal{F}\Big] \le \left(1+\frac{1}{s}\right)\mathbb{E}_{\mathcal{B}}\left[\sum_{i\in\mathcal{B}^c}\|\tilde{Y}_i - Y_i\|_{\mathrm{F}}^2 \;\middle|\; \mathcal{F}\right] + (1+s)\,\mathbb{E}_{\mathcal{B}}\left[\sum_{i\in\mathcal{B}^c}\|Y_i - Y_i^+\|_{\mathrm{F}}^2 \;\middle|\; \mathcal{F}\right].$$

In the remainder of this section, we abbreviate $\mathbb{E}_{\mathcal{B}}[\cdot \mid \mathcal{F}]$ as $\mathbb{E}_{\mathcal{B}}[\cdot]$. Because $\tilde{Y}_i - Y_i$ is independent of the random batch $\mathcal{B}$, every index $i \in \{1,\dots,N\}$ is excluded from $\mathcal{B}$ with the same probability $1 - \frac{|\mathcal{B}|}{N}$. By linearity of expectation,

$$\mathbb{E}_{\mathcal{B}}\big[\sum_{i\in\mathcal{B}^c}\|\tilde{Y}_i - Y_i\|_{\mathrm{F}}^2\big] = (1-\frac{|\mathcal{B}|}{N})\sum_{i=1}^{N}\|\tilde{Y}_i - Y_i\|_{\mathrm{F}}^2 = (1-\frac{|\mathcal{B}|}{N})\|\tilde{Y} - Y\|_{\mathrm{F}}^2.$$

For the second term note that summing over $\mathcal{B}^c$ is bounded by summing over all samples:

$$\mathbb{E}_{\mathcal{B}}\big[\sum_{i\in\mathcal{B}^c}\|Y_i - Y_i^+\|_{\mathrm{F}}^2\big] \le \mathbb{E}_{\mathcal{B}}\big[\sum_{i=1}^{N}\|Y_i - Y_i^+\|_{\mathrm{F}}^2\big] = \mathbb{E}_{\mathcal{B}}\big[\|Y - Y^+\|_{\mathrm{F}}^2\big].$$

Using the update $\theta^+ - \theta = -\eta H^T$ and $\|\nabla_\theta Y\|_{\mathrm{F}} \le B_1$ (Assumption A_3), we have

$$\|Y - Y^+\|_{\mathrm{F}} \le B_1\|\theta - \theta^+\|_{\mathrm{F}} = B_1\eta\|H\|_{\mathrm{F}}.$$

Applying Lemma 4.1 ($\|H\|_{\mathrm{F}} \le B_1\|\widetilde{\nabla_Y F}\|_{\mathrm{F}}$) yields

$$\mathbb{E}_{\mathcal{B}}\big[\|Y - Y^+\|_{\mathrm{F}}^2\big] \le \eta^2 B_1^2\mathbb{E}_{\mathcal{B}}\big[\|H\|_{\mathrm{F}}^2\big] \le \eta^2 B_1^4\|\widetilde{\nabla_Y F}\|_{\mathrm{F}}^2.$$

Combining the preceding bounds gives the stated inequality

$$\mathbb{E}_{\mathcal{B}}\big[\|\tilde{Y}^+ - Y^+\|_{\mathrm{F}}^2\big] \le (1+\frac{1}{s})\left(1-\frac{|\mathcal{B}|}{N}\right)\|\tilde{Y} - Y\|_{\mathrm{F}}^2 + (1+s)\cdot\eta^2 B_1^4\|\widetilde{\nabla_Y F}\|_{\mathrm{F}}^2.$$

□

A.13 Proof of Lemma 4.4

Proof. Starting from Lemma 4.3, for each k we have (for any $s_k > 0$)

$$\begin{aligned}
&\mathbb{E}\|\tilde{Y}(\theta_k) - Y(\theta_k)\|_{\mathrm{F}}^2 \\
=&\mathbb{E}_{\mathcal{B}_0,\cdots,\mathcal{B}_{k-1}}\|\tilde{Y}(\theta_k) - Y(\theta_k)\|_{\mathrm{F}}^2 \\
=&\mathbb{E}_{\mathcal{B}_0,\cdots,\mathcal{B}_{k-2}}\mathbb{E}_{\mathcal{B}_{k-1}}\big[\|\tilde{Y}(\theta_k) - Y(\theta_k)\|_{\mathrm{F}}^2|\mathcal{B}_0,\cdots,\mathcal{B}_{k-2}\big] \\
\le&(1+\frac{1}{s_k})\left(1-\frac{|\mathcal{B}|}{N}\right)\mathbb{E}\|\tilde{Y}(\theta_{k-1}) - Y(\theta_{k-1})\|_{\mathrm{F}}^2 \\
&+\eta^2(1+s_k)B_1^4\mathbb{E}\|\widetilde{\nabla_Y F}\left(\tilde{Y}(\theta_{k-1}), Y(\theta_{k-1})\right)\|_{\mathrm{F}}^2.
\end{aligned}$$

Then

$$
\begin{aligned}
&\mathbb{E}\|\tilde{Y}(\theta_k) - Y(\theta_k)\|_{\mathrm{F}}^2 \\
\leq&(1+\frac{1}{s_k})\left(1-\frac{|\mathcal{B}|}{N}\right)\mathbb{E}\|\tilde{Y}(\theta_{k-1}) - Y(\theta_{k-1})\|_{\mathrm{F}}^2 + \eta^2(1+s_k)B_1^4\mathbb{E}\|\widetilde{\nabla_Y F}\left(\tilde{Y}(\theta_{k-1}), Y(\theta_{k-1})\right)\|_{\mathrm{F}}^2 \\
\leq&(1+\frac{1}{s_{k-1}})(1+\frac{1}{s_k})\left(1-\frac{|\mathcal{B}|}{N}\right)^2\mathbb{E}\|\tilde{Y}(\theta_{k-2}) - Y(\theta_{k-2})\|_{\mathrm{F}}^2 \\
&+\eta^2 B_1^4(1+s_{k-1})(1+\frac{1}{s_k})\left(1-\frac{|\mathcal{B}|}{N}\right)\mathbb{E}\|\widetilde{\nabla_Y F}\left(\tilde{Y}(\theta_{k-2}), Y(\theta_{k-2})\right)\|_{\mathrm{F}}^2 \\
&+\eta^2 B_1^4(1+s_k)\mathbb{E}\|\widetilde{\nabla_Y F}\left(\tilde{Y}(\theta_{k-1}), Y(\theta_{k-1})\right)\|_{\mathrm{F}}^2 \\
\leq&\cdots \\
\leq&\left(\prod_{j=1}^{k}(1+\frac{1}{s_j})\right)\cdot\left(1-\frac{|\mathcal{B}|}{N}\right)^k\mathbb{E}\|\tilde{Y}(\theta_0) - Y(\theta_0)\|_{\mathrm{F}}^2 \\
&+\sum_{i=1}^{k}\eta^2 B_1^4(1+s_{k+1-i})\cdot\left(\prod_{j=0}^{i-2}(1+\frac{1}{s_{k-j}})\right)\cdot\left(1-\frac{|\mathcal{B}|}{N}\right)^{i-1}\mathbb{E}\|\widetilde{\nabla_Y F}\left(\tilde{Y}(\theta_{k-i}), Y(\theta_{k-i})\right)\|_{\mathrm{F}}^2 \\
=&\sum_{i=1}^{k}\eta^2 B_1^4(1+s_{k+1-i})\cdot\left(\prod_{j=0}^{i-2}(1+\frac{1}{s_{k-j}})\right)\cdot\left(1-\frac{|\mathcal{B}|}{N}\right)^{i-1}\mathbb{E}\|\widetilde{\nabla_Y F}\left(\tilde{Y}(\theta_{k-i}), Y(\theta_{k-i})\right)\|_{\mathrm{F}}^2,
\end{aligned}
\tag{A.1}
$$

where the last equality holds because the cached value coincides with the true value at initialization, i.e., $\tilde{Y}(\theta_0) = Y(\theta_0)$. Set for $j \geq 1$

$$
s_j := \frac{\mu(1-\mu^{j-1})}{(1-\mu)},
$$

One checks that $s_j > 0$ for $\mu \in (0,1)$. With this choice,

$$
\prod_{j=0}^{i-2}(1+\frac{1}{s_{k-j}}) = \prod_{j=0}^{i-2}\left(1+\frac{1-\mu}{\mu(1-\mu^{k-j-1})}\right) = \prod_{j=0}^{i-2}\left(\frac{1-\mu^{k-j}}{\mu(1-\mu^{k-j-1})}\right) = \frac{1-\mu^k}{\mu^{i-1}(1-\mu^{k-i+1})}. \tag{A.2}
$$

Therefore, subsitute s_{k+1-i} and equality A.2 into inequality A.1, we have

$$
\begin{aligned}
&\mathbb{E}\|\tilde{Y}(\theta_k) - Y(\theta_k)\|_{\mathrm{F}}^2 \\
\leq&\sum_{i=1}^{k}\eta^2 B_1^4(1+\frac{\mu-\mu^{k-i+1}}{1-\mu})\cdot\frac{1-\mu^k}{\mu^{i-1}(1-\mu^{k-i+1})}\cdot\mu^{2i-2}\mathbb{E}\|\widetilde{\nabla_Y F}\left(\tilde{Y}(\theta_{k-i}), Y(\theta_{k-i})\right)\|_{\mathrm{F}}^2 \\
=&\sum_{i=1}^{k}\eta^2 B_1^4\frac{1-\mu^{k-i+1}}{1-\mu}\cdot\frac{1-\mu^k}{1-\mu^{k-i+1}}\cdot\mu^{i-1}\mathbb{E}\|\widetilde{\nabla_Y F}\left(\tilde{Y}(\theta_{k-i}), Y(\theta_{k-i})\right)\|_{\mathrm{F}}^2 \\
=&\sum_{i=1}^{k}\eta^2 B_1^4\frac{1-\mu^k}{\mu(1-\mu)}\mu^i\mathbb{E}\|\widetilde{\nabla_Y F}\left(\tilde{Y}(\theta_{k-i}), Y(\theta_{k-i})\right)\|_{\mathrm{F}}^2 \\
=&B_1^4\cdot\frac{\eta^2\cdot(1-\mu^k)}{\mu(1-\mu)}\sum_{i=1}^{k}\mu^i\mathbb{E}\|\widetilde{\nabla_Y F}\left(\tilde{Y}(\theta_{k-i}), Y(\theta_{k-i})\right)\|_{\mathrm{F}}^2 \\
\leq&B_1^4\cdot\frac{\eta^2}{\mu(1-\mu)}\sum_{i=1}^{k}\mu^i\mathbb{E}\|\widetilde{\nabla_Y F}\left(\tilde{Y}(\theta_{k-i}), Y(\theta_{k-i})\right)\|_{\mathrm{F}}^2.
\end{aligned}
$$

□

A.14 Proof of Lemma 4.5

Proof. We prove both assertions by induction on k. The base case $k = 0$ is immediate since $\tilde{Y}(\theta_0) = Y(\theta_0)$, so $\widetilde{\nabla_Y F} = \nabla_Y F$ and the inequality 4.1 holds trivially. We first establish inequality 4.2 for $k = 0$.

By Taylor expansion of $\nabla_Y F$ around $Y(\theta_0)$, we have the following integral form:

$$\nabla_Y F(Y(\theta_1)) = \nabla_Y F(Y(\theta_0)) + R_0,$$

where

$$R_0 = \int_0^1 \nabla_Y^2 F\big(Y(\theta_0) + t\,(Y(\theta_1) - Y(\theta_0))\,\big)\,(Y(\theta_1) - Y(\theta_0))\,\mathrm{d}t.$$

Using the parameter update $\theta_1 - \theta_0 = -\eta H(\theta_0)^\top$ and the chain rule,

$$Y(\theta_1) - Y(\theta_0) = \int_0^1 \nabla_\theta Y(\theta_0 + t(\theta_1 - \theta_0))(\theta_1 - \theta_0)\mathrm{d}t,$$

here, rather than considering $\nabla_\theta Y$ as a matrix, we retain its tensor form. So, with the bound $\|\nabla_Y^2 F\|_{\mathrm{F}} \le L$ (Assumption A_2) and $\|\nabla_\theta Y\|_{\mathrm{F}} \le B_1$ (Assumption A_3),

$$\|R_0\|_{\mathrm{F}} \le L\|Y(\theta_1) - Y(\theta_0)\|_{\mathrm{F}} \le LB_1\eta\|H(\theta_0)\|_{\mathrm{F}}.$$

Taking expectation, applying Lemma 4.1 ($\|H\|_{\mathrm{F}} \le B_1\|\widetilde{\nabla_Y F}\|_{\mathrm{F}}$) and using $\widetilde{\nabla_Y F}(\tilde{Y}, Y) = \nabla_Y F(Y)$ at θ_0, we obtain

$$\mathbb{E}\|R_0\|_{\mathrm{F}}^2 \le (B_1^2 L)^2\eta^2\mathbb{E}\|\nabla_Y F(Y(\theta_0))\|_{\mathrm{F}}^2.$$

Then by Minkowski inequality one gets

$$\sqrt{\mathbb{E}\|\nabla_Y F(Y(\theta_0)) + R_0\|_{\mathrm{F}}^2} \le \sqrt{\mathbb{E}\|\nabla_Y F(Y(\theta_0))\|_{\mathrm{F}}^2} + \sqrt{\mathbb{E}\|R_0\|_{\mathrm{F}}^2},$$

hence

$$\begin{aligned}
\mathbb{E}\|\nabla_Y F(Y(\theta_1))\|_{\mathrm{F}}^2 =& \mathbb{E}\|\nabla_Y F(Y(\theta_0)) + R_0\|_{\mathrm{F}}^2 \\
\le& \mathbb{E}\|\nabla_Y F(Y(\theta_0))\|_{\mathrm{F}}^2 + 2\sqrt{\mathbb{E}\|\nabla_Y F(Y(\theta_0))\|_{\mathrm{F}}^2 \cdot \mathbb{E}\|R_0\|_{\mathrm{F}}^2} + \mathbb{E}\|R_0\|_{\mathrm{F}}^2 \\
\le& \|\nabla_Y F(Y(\theta_0))\|_{\mathrm{F}}^2 + 2B_1^2 L \cdot \eta \cdot \sqrt{\|\nabla_Y F(Y(\theta_0))\|_{\mathrm{F}}^2 \cdot \|\nabla_Y F(Y(\theta_0))\|_{\mathrm{F}}^2} \\
&+ (B_1^2 L)^2 \cdot \eta^2 \cdot \|\nabla_Y F(Y(\theta_0))\|_{\mathrm{F}}^2 \\
=& \left(1 + 2B_1^2 L\eta + (B_1^2 L)^2 \cdot \eta^2\right) \|\nabla_Y F(Y(\theta_0))\|_{\mathrm{F}}^2 \\
\le& \left(1 + 2B_1^2 L\eta + (B_1^2 L)^2 \cdot \eta\right) \|\nabla_Y F(Y(\theta_0))\|_{\mathrm{F}}^2 \\
\le& \left(1 + c_4 \cdot \eta\right) \|\nabla_Y F(Y(\theta_0))\|_{\mathrm{F}}^2,
\end{aligned}$$

where the penultimate inequality follows from $\eta < 1$. Similarly the matching lower bound yields

$$\begin{aligned}
\mathbb{E}\|\nabla_Y F(Y(\theta_1))\|_{\mathrm{F}}^2 =& \mathbb{E}\|\nabla_Y F(Y(\theta_0)) + R_0\|_{\mathrm{F}}^2 \\
\ge& \mathbb{E}\left(\|\nabla_Y F(Y(\theta_0))\|_{\mathrm{F}} - \|R_0\|_{\mathrm{F}}\right)^2 \\
\ge& \mathbb{E}\|\nabla_Y F(Y(\theta_0))\|_{\mathrm{F}}^2 - 2\sqrt{\mathbb{E}\|\nabla_Y F(Y(\theta_0))\|_{\mathrm{F}}^2 \cdot \mathbb{E}\|R_0\|_{\mathrm{F}}^2} + \mathbb{E}\|R_0\|_{\mathrm{F}}^2 \\
\ge& \mathbb{E}\|\nabla_Y F(Y(\theta_0))\|_{\mathrm{F}}^2 - 2\sqrt{\mathbb{E}\|\nabla_Y F(Y(\theta_0))\|_{\mathrm{F}}^2 \cdot \mathbb{E}\|R_0\|_{\mathrm{F}}^2} - \mathbb{E}\|R_0\|_{\mathrm{F}}^2 \\
\ge& \left(1 - 2B_1^2 L\eta - (B_1^2 L)^2 \cdot \eta^2\right) \|\nabla_Y F(Y(\theta_0))\|_{\mathrm{F}}^2 \\
\ge& (1 - c_4 \cdot \eta)\|\nabla_Y F(Y(\theta_0))\|_{\mathrm{F}}^2.
\end{aligned}$$

Thus inequality 4.2 holds at $k = 0$. Assume that inequality 4.1 and 4.2 hold for all indices $i < k$. Then we prove them for k.

Under Assumption A_6 the approximation operator satisfies the linearization bound

$$\widetilde{\nabla_Y F}(\tilde{Y}, Y) = \nabla_Y F(Y) + \int_0^1 \mathcal{T}(Y + t(\tilde{Y} - Y))[\tilde{Y} - Y]\mathrm{d}t,$$

with $\|\mathcal{T}(\Theta)[\Delta]\|_{\mathrm{F}} \le B_3\|\Delta\|_{\mathrm{F}}$ for all relevant Θ. Hence

$$\begin{aligned}\|\widetilde{\nabla_Y F}(\tilde{Y}, Y)\|_{\mathrm{F}}^2 \le& 2\|\nabla_Y F(Y)\|_{\mathrm{F}}^2 + 2\|\int_0^1 \mathcal{T}(Y + t(\tilde{Y} - Y))[\tilde{Y} - Y]\mathrm{d}t\|_{\mathrm{F}}^2 \\ \le& 2\|\nabla_Y F(Y)\|_{\mathrm{F}}^2 + 2\int_0^1 B_3^2\|\tilde{Y} - Y\|_{\mathrm{F}}^2\mathrm{d}t \\ \le& 2\|\nabla_Y F(Y)\|_{\mathrm{F}}^2 + 2B_3^2\|\tilde{Y} - Y\|_{\mathrm{F}}^2.\end{aligned} \tag{A.3}$$

Combining Lemma 4.4 and the induction hypothesis,

$$\begin{aligned}\mathbb{E}\|\tilde{Y}(\theta_k) - Y(\theta_k)\|_{\mathrm{F}}^2 \le& B_1^4 \cdot \frac{\eta^2}{\mu(1-\mu)} \sum_{i=1}^{k} \mu^i \mathbb{E}\|\widetilde{\nabla_Y F}\left(\tilde{Y}(\theta_{k-i}), Y(\theta_{k-i})\right)\|_{\mathrm{F}}^2 \\ \le& B_1^4 \cdot \frac{4\eta^2}{\mu(1-\mu)} \sum_{i=1}^{k} \mu^i \mathbb{E}\|\nabla_Y F(Y(\theta_{k-i}))\|_{\mathrm{F}}^2 \\ \le& B_1^4 \cdot \frac{4\eta^2}{\mu(1-\mu)} \sum_{i=1}^{k} \mu^i (1 - c_4 \cdot \eta)^{-i} \mathbb{E}\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2.\end{aligned}$$

Set $\nu := \mu(1 - c_4\eta)^{-1}$. Under the first constraint on η from the lemma statement one checks that $\nu < 1$ ($\nu \le (1 - |\mathcal{B}|/2N)^{-1} < 1$). Hence summing the geometric series yields

$$\sum_{i=1}^{k} \mu^i (1 - c_4 \cdot \eta)^{-i} \le \sum_{i=1}^{\infty} \nu^i = \frac{\nu}{1-\nu} \le \frac{1}{1-\nu}.$$

Therefore

$$\mathbb{E}\|\tilde{Y}(\theta_k) - Y(\theta_k)\|_{\mathrm{F}}^2 \le B_1^4 \cdot \frac{4\eta^2}{\mu(1-\mu)} \cdot \frac{1}{1-\nu} \mathbb{E}\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2.$$

Plugging this bound into inequality A.3 gives

$$\begin{aligned}\mathbb{E}\|\widetilde{\nabla_Y F}(\tilde{Y}(\theta_k), Y(\theta_k))\|_{\mathrm{F}}^2 \le& 2\mathbb{E}\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2 + 2B_3^2\mathbb{E}\|\tilde{Y}(\theta_k) - Y(\theta_k)\|_{\mathrm{F}}^2 \\ \le& 2\mathbb{E}\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2 + 2B_3^2 B_1^4 \cdot \frac{4\eta^2}{\mu(1-\mu)} \cdot \frac{1}{1-\nu} \mathbb{E}\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2 \\ \le& 2\mathbb{E}\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2 + \frac{8B_3^2B_1^4}{\mu(1-\mu)} \cdot \frac{\mu(1-\mu)|\mathcal{B}|^2}{32B_1^4B_3^2(2N-|\mathcal{B}|)^2} \cdot \frac{1}{1-\nu} \mathbb{E}\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2 \\ =& 2\mathbb{E}\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2 + \frac{1}{2} \cdot \frac{|\mathcal{B}|^2}{2(2N-|\mathcal{B}|)^2} \cdot \frac{1}{1-\nu} \mathbb{E}\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2.\end{aligned}$$

Since

$$\begin{aligned}1 - \nu =& 1 - \frac{\mu}{1 - c_4 \cdot \eta} > 1 - \frac{\mu}{1 - |\mathcal{B}|/(2N)} \\ =& 1 - \sqrt{\frac{1 - |\mathcal{B}|/N}{[1 - |\mathcal{B}|/(2N)]^2}} = 1 - \sqrt{1 - \frac{|\mathcal{B}|^2}{(2N - |\mathcal{B}|)^2}} \\ >& 1 - \left(1 - \frac{|\mathcal{B}|^2}{2(2N - |\mathcal{B}|)^2}\right) = \frac{|\mathcal{B}|^2}{2(2N - |\mathcal{B}|)^2},\end{aligned}$$

where the last inequality follows from $\sqrt{1-x} \leq 1 - x/2$, then

$$\begin{aligned}\mathbb{E}\|\widetilde{\nabla_Y F}(\tilde{Y}(\theta_k), Y(\theta_k))\|_{\mathrm{F}}^2 \leq & 2\mathbb{E}\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2 + \frac{1}{2} \cdot \frac{|\mathcal{B}|^2}{2(2N-|\mathcal{B}|)^2} \cdot \frac{1}{1-\nu}\mathbb{E}\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2 \\ \leq & 4\mathbb{E}\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2.\end{aligned} \tag{A.4}$$

Similarly,

$$\begin{aligned}\mathbb{E}\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2 =& \mathbb{E}\|\widetilde{\nabla_Y F}(\tilde{Y}(\theta_k), Y(\theta_k)) - \int_0^1 \mathcal{T}\left(Y(\theta_k) + t(\tilde{Y}(\theta_k) - Y(\theta_k))\right)[\tilde{Y}(\theta_k) - Y(\theta_k)]\mathrm{d}t\|_{\mathrm{F}}^2 \\ \leq & 2\mathbb{E}\|\widetilde{\nabla_Y F}(\tilde{Y}(\theta_k), Y(\theta_k))\|_{\mathrm{F}}^2 + 2B_3^2\mathbb{E}\|\tilde{Y}(\theta_k) - Y(\theta_k)\|_{\mathrm{F}}^2 \\ \leq & 2\mathbb{E}\|\widetilde{\nabla_Y F}(\tilde{Y}(\theta_k), Y(\theta_k))\|_{\mathrm{F}}^2 + \frac{1}{2}\mathbb{E}\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2,\end{aligned}$$

which means $\mathbb{E}\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2 \leq 4\mathbb{E}\|\widetilde{\nabla_Y F}(\tilde{Y}(\theta_k), Y(\theta_k))\|_{\mathrm{F}}^2$. At last, according to Taylor expansion as at the beginning of the proof:

$$\nabla_Y F(Y(\theta_{k+1})) = \nabla_Y F(Y(\theta_k)) + R_k,$$

where

$$R_k = \int_0^1 \nabla_Y^2 F\left(Y(\theta_k) + t(Y(\theta_{k+1}) - Y(\theta_k))\right)(Y(\theta_{k+1}) - Y(\theta_k))\mathrm{d}t.$$

And here, likewise, rather than considering $\nabla_\theta Y$ as a matrix, we retain its tensor form. Then according to Lemma 4.1 and inequality A.4

$$\begin{aligned}\mathbb{E}\|R_k\|_{\mathrm{F}}^2 \leq & \mathbb{E}\big[\int_0^1 L^2\|Y(\theta_{k+1}) - Y(\theta_k)\|_{\mathrm{F}}^2\big] \\ \leq & L^2 \cdot B_1^2 \cdot \eta^2 \cdot \mathbb{E}\|H(\theta_k)\|_{\mathrm{F}}^2 \\ \leq & (B_1^2 L)^2 \cdot \eta^2 \cdot \mathbb{E}\|\widetilde{\nabla_Y F}\left(\tilde{Y}(\theta_k), Y(\theta_k)\right)\|_{\mathrm{F}}^2 \\ \leq & (B_1^2 L)^2 \cdot \eta^2 \cdot 4\mathbb{E}\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2.\end{aligned}$$

Hence, similar to the proof for $k = 0$, we have

$$\begin{aligned}\mathbb{E}\|\nabla_Y F(Y(\theta_{k+1}))\|_{\mathrm{F}}^2 =& \mathbb{E}\|\nabla_Y F(Y(\theta_k)) + R_k\|_{\mathrm{F}}^2 \\ \leq & \mathbb{E}\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2 + 2\sqrt{\mathbb{E}\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2 \cdot \mathbb{E}\|R_k\|_{\mathrm{F}}^2} + \mathbb{E}\|R_k\|_{\mathrm{F}}^2 \\ \leq & \|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2 + 4B_1^2 L \cdot \eta \cdot \sqrt{\mathbb{E}\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2 \cdot \mathbb{E}\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2} \\ & + (2B_1^2 L)^2 \cdot \eta^2 \cdot \mathbb{E}\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2 \\ = & \left(1 + 4B_1^2 L\eta + 4(B_1^2 L)^2 \cdot \eta^2\right)\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2 \\ \leq & \left(1 + 4B_1^2 L\eta + 4(B_1^2 L)^2 \cdot \eta\right)\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2 \\ = & (1 + c_4 \cdot \eta)\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2,\end{aligned}$$

where the penultimate inequality follows from $\eta < 1$. In the same way, we can get

$$\mathbb{E}\|\nabla_Y F(Y(\theta_{k+1}))\|_{\mathrm{F}}^2 \geq \left(1 - 4B_1^2 L\eta - 4(B_1^2 L)^2 \cdot \eta^2\right)\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2 = (1 - c_4 \cdot \eta)\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2.$$

Thus, we have completed the proof the induction step for k. Therefore, by mathematical induction, the lemma holds. □

A.15 Proof of Lemma 4.6

Proof. According to Lemma 4.2,

$$F(Y^+) \le F(Y) - \eta \cdot \sum_{i=1}^{N} \langle z_i, H^\top \rangle + c_2 \cdot \eta^2 \|\widetilde{\nabla_Y F}\|_{\mathrm{F}}^2.$$

Let $\mathcal{F}$ denote the filtration generated by all randomness up to the current iterate. For the single-step analysis of this theorem, we take conditional expectation over the random mini-batch $\mathcal{B}$ given $\mathcal{F}$, denoted by $\mathbb{E}_{\mathcal{B}}[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}]$. This yields

$$\begin{aligned}
\mathbb{E}_{\mathcal{B}}\big[F(Y^+)\big] &\le \mathbb{E}_{\mathcal{B}}\big[F(Y) - \eta \cdot \sum_{i=1}^{N} \langle z_i, H^\top \rangle + c_2 \cdot \eta^2 \|\widetilde{\nabla_Y F}\|_{\mathrm{F}}^2\big] \\
&= F(Y) - \eta \cdot \sum_{i=1}^{N} \langle z_i, \mathbb{E}_{\mathcal{B}}[H^\top] \rangle + c_2 \cdot \eta^2 \|\widetilde{\nabla_Y F}\|_{\mathrm{F}}^2 \\
&= F(Y) - \eta \cdot \sum_{i=1}^{N} \langle z_i, \frac{|\mathcal{B}|}{N} \sum_{j=1}^{N} p_j \rangle + c_2 \cdot \eta^2 \|\widetilde{\nabla_Y F}\|_{\mathrm{F}}^2,
\end{aligned}$$

where p_j denotes the per-sample contribution and we used $\mathbb{E}_{\mathcal{B}}[H] = \frac{|\mathcal{B}|}{N} \sum_j p_j$ since p_j is $\mathcal{F}$-measurable (determined by the history up to the current iterate) and is thus independent of the current mini-batch $\mathcal{B}$. Rewriting the double sum in matrix form and using the identification (row-wise concatenation) of the block gradient vectors, the main descent term can be written as

$$\begin{aligned}
\mathbb{E}_{\mathcal{B}}\big[F(Y^+)\big] &\le F(Y) - \eta \cdot \frac{|\mathcal{B}|}{N} \cdot \sum_{i=1}^{N} \sum_{j=1}^{N} \langle z_i, p_j \rangle + c_2 \cdot \eta^2 \|\widetilde{\nabla_Y F}\|_{\mathrm{F}}^2 \\
&= F(Y) - \eta \cdot \frac{|\mathcal{B}|}{N} \cdot \sum_{i=1}^{N} \sum_{j=1}^{N} (\nabla_Y F)_i \cdot \nabla_\theta Y_i \cdot \nabla_\theta Y_j^\top \cdot (\widetilde{\nabla_Y F})_j^\top + c_2 \cdot \eta^2 \|\widetilde{\nabla_Y F}\|_{\mathrm{F}}^2 \\
&= F(Y) - \eta \cdot \frac{|\mathcal{B}|}{N} \cdot \hat{g}\big(\nabla_\theta Y \cdot (\nabla_\theta Y)^\top\big)\hat{h}^\top + c_2 \cdot \eta^2 \|\widetilde{\nabla_Y F}\|_{\mathrm{F}}^2,
\end{aligned}$$

where $\hat{g}$ corresponds to the row-vectorized $\nabla_Y F$ and $\hat{h}$ corresponds to the row-vectorized $\widetilde{\nabla_Y F}$. Since

$$\begin{aligned}
\widetilde{\nabla_Y F} = \widetilde{\nabla_Y F}(\tilde{Y}, Y) &= \widetilde{\nabla_Y F}(Y, Y) + \int_0^1 \mathcal{T}(Y + t(\tilde{Y} - Y))[\tilde{Y} - Y] \mathrm{d}t \\
&= \nabla_Y F + \int_0^1 \mathcal{T}(Y + t(\tilde{Y} - Y))[\tilde{Y} - Y] \mathrm{d}t.
\end{aligned}$$

Considering the row-vectorized form, let $\hat{\mathcal{T}}$ denote the corresponding derivative matrix (Jacobian) induced by row-wise vectorization, satisfying $\mathrm{vec}_r(\mathcal{T}(\Theta)[\Delta]) = \hat{\mathcal{T}}(\Theta)[\mathrm{vec}_r(\Delta)]$ for any matrix Δ. Denote

$$R_r = \int_0^1 \hat{\mathcal{T}}(Y + t(\tilde{Y} - Y))[\mathrm{vec}_r(\tilde{Y} - Y)] \mathrm{d}t,$$

by Assumption A_6, we have $\|R_r\|_{\mathrm{F}} \le B_3\|\tilde{Y} - Y\|_{\mathrm{F}}$. Then $\hat{h} = \hat{g} + R_r$, and furthermore

$$\begin{aligned}
\mathbb{E}_{\mathcal{B}}\big[F(Y^+)\big] \le & F(Y) - \eta \cdot \frac{|\mathcal{B}|}{N} \cdot \hat{g}\big(\nabla_\theta Y \cdot (\nabla_\theta Y)^\top\big)\hat{h}^\top + c_2 \cdot \eta^2 \|\widetilde{\nabla_Y F}\|_{\mathrm{F}}^2 \\
= & F(Y) - \eta \cdot \frac{|\mathcal{B}|}{N} \cdot \hat{g}\big(\nabla_\theta Y \cdot (\nabla_\theta Y)^\top\big)\hat{g}^\top + c_2 \cdot \eta^2 \|\widetilde{\nabla_Y F}\|_{\mathrm{F}}^2 \\
& - \eta^2 \cdot \frac{|\mathcal{B}|}{N} \cdot \hat{g}\big(\nabla_\theta Y \cdot (\nabla_\theta Y)^\top\big) \cdot \frac{1}{\eta} R_r^\top \\
\le & F(Y) - \eta \cdot \frac{|\mathcal{B}|}{N} \cdot \lambda_{\min} \|\hat{g}\|_{\mathrm{F}}^2 + c_2 \cdot \eta^2 \|\widetilde{\nabla_Y F}\|_{\mathrm{F}}^2 \\
& + \eta^2 \cdot \frac{|\mathcal{B}|}{N} \cdot \frac{1}{2} \left(\|\hat{g}\big(\nabla_\theta Y \cdot (\nabla_\theta Y)^\top\big)\|_{\mathrm{F}}^2 + \frac{1}{\eta^2}\|R_r\|_{\mathrm{F}}^2 \right) \\
\le & F(Y) - \eta \cdot \frac{|\mathcal{B}|}{N} \cdot \lambda_{\min} \|\hat{g}\|_{\mathrm{F}}^2 + c_2 \cdot \eta^2 \|\widetilde{\nabla_Y F}\|_{\mathrm{F}}^2 \\
& + \eta^2 \cdot \frac{|\mathcal{B}|}{2N} \cdot B_1^4 \|\hat{g}\|_{\mathrm{F}}^2 + \eta^2 \cdot \frac{|\mathcal{B}|}{2N} \cdot \frac{B_3^2}{\eta^2} \|\tilde{Y} - Y\|_{\mathrm{F}}^2 \\
= & F(Y) - \eta \cdot \frac{|\mathcal{B}|}{N} \left(\lambda_{\min} - \frac{B_1^4 \eta}{2} \right) \|\nabla_Y F\|_{\mathrm{F}}^2 + c_2 \cdot \eta^2 \|\widetilde{\nabla_Y F}\|_{\mathrm{F}}^2 + \frac{|\mathcal{B}| \cdot B_3^2}{2N} \|\tilde{Y} - Y\|_{\mathrm{F}}^2.
\end{aligned}$$

According to Lemma 4.4 and 4.5 we have

$$\begin{aligned}
\mathbb{E}\|\tilde{Y}(\theta_k) - Y(\theta_k)\|_{\mathrm{F}}^2 \le & B_1^4 \cdot \frac{\eta^2}{\mu(1-\mu)} \sum_{i=1}^{k} \mu^i \mathbb{E}\|\widetilde{\nabla_Y F}\left(\tilde{Y}(\theta_{k-i}), Y(\theta_{k-i})\right)\|_{\mathrm{F}}^2 \\
\le & B_1^4 \cdot \frac{4\eta^2}{\mu(1-\mu)} \sum_{i=1}^{k} \mu^i \mathbb{E}\|\nabla_Y F(Y(\theta_{k-i}))\|_{\mathrm{F}}^2 \\
\le & B_1^4 \cdot \frac{4\eta^2}{\mu(1-\mu)} \sum_{i=1}^{k} \mu^i (1 - c_4 \cdot \eta)^{-i} \mathbb{E}\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2 \\
\le & B_1^4 \cdot \frac{4\eta^2}{\mu(1-\mu)} \cdot \frac{1}{1-\nu} \mathbb{E}\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2.
\end{aligned}$$

Further applying Lemma 4.5 yields

$$\begin{aligned}
\mathbb{E}\big[F(Y(\theta_{k+1}))\big] \le & \mathbb{E}\big[F(Y(\theta_k))\big] - \eta \cdot \frac{|\mathcal{B}|}{N} \left(\lambda_{\min} - \frac{B_1^4 \eta}{2} \right) \mathbb{E}\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2 \\
& + c_2 \cdot \eta^2 \cdot 4\mathbb{E}\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2 + \frac{2|\mathcal{B}| \cdot B_1^4 B_3^2}{N\mu(1-\mu)(1-\nu)} \cdot \eta^2 \cdot \mathbb{E}\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2 \\
= & \mathbb{E}\big[F(Y(\theta_k))\big] - \eta \cdot \left(\frac{|\mathcal{B}|}{N} \lambda_{\min} - c_5 \cdot \eta \right) \mathbb{E}\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2.
\end{aligned}$$

Take $c_6 = \frac{|\mathcal{B}|}{N}\lambda_{\min} - c_5 \cdot \eta$, then

$$\mathbb{E}\big[F(Y(\theta_{k+1}))\big] \le \mathbb{E}\big[F(Y(\theta_k))\big] - c_6 \cdot \eta \mathbb{E}\|\nabla_Y F(Y(\theta_k))\|_{\mathrm{F}}^2.$$

□

Reference

[1] John Duchi, Elad Hazan, and Yoram Singer. "Adaptive subgradient methods for online learning and stochastic optimization". In: *Journal of Machine Learning Research* 12.7 (2011), pp. 2121–2159. URL: http://jmlr.org/papers/v12/duchi11a.html.

[2] Matthew D. Zeiler. "ADADELTA: An adaptive learning rate method". In: *arXiv preprint* (2012). arXiv: 1212.5701 [cs.LG].

[3] Diederik P. Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint* (2014). arXiv: 1412.6980 [cs.LG].

[4] Mark Schmidt, Nicolas Le Roux, and Francis Bach. "Minimizing finite sums with the stochastic average gradient". In: *Mathematical Programming* 162.1 (2017), pp. 83–112. DOI: 10.1007/s10107-

[5] Rie Johnson and Tong Zhang. "Accelerating stochastic gradient descent using predictive variance reduction". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 26. 2013, pp. 315–323. URL: https://proceedings.neurips.cc/paper/2013/hash/ac1dd209cbcc5e5d1

[6] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. "SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 27. 2014, pp. 1646–1654. URL: https://proceedings.neurips.cc/paper/2014/hash/f7cade80b7cc92b991cf4d2806d6bd78-Abs

[7] Purushottam Kar, Harikrishna Narasimhan, and Prateek Jain. "Online and Stochastic Gradient Methods for Non-decomposable Loss Functions". In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani et al. Vol. 27. Curran Associates, Inc., 2014. URL: https://proceedings.neurips.cc/paper_files/paper/2014/file/9638ddfc7e3d56a611292c15

[8] Ching-Yao Chuang et al. "Debiased Contrastive Learning". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 8765–8775. URL: https://neurips.cc/virtual/2020/public/poster_63c3ddcc7b23daa1e42dc41f9a44a873.html.

[9] Soham Dan and Dushyant Sahoo. "Variance Reduced Stochastic Proximal Algorithm for AUC Maximization". In: *Machine Learning and Knowledge Discovery in Databases. Research Track*. Springer International Publishing, 2021, pp. 184–199. ISBN: 9783030865238. DOI: 10.1007/978-3-030-86523-8_12. URL: http://dx.doi.org/10.1007/978-3-030-86523-8_12.

[10] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. "Optimization methods for large-scale machine learning". In: *SIAM Review* 60.2 (2018), pp. 223–311. DOI: 10.1137/16M1080173.

[11] Arkadi Nemirovski et al. "Robust stochastic approximation approach to stochastic programming". In: *SIAM Journal on Optimization* 19.4 (2009), pp. 1574–1609. DOI: 10.1137/070704277.

[12] Yurii Nesterov. "Gradient methods for minimizing composite functions". In: *Mathematical Programming* 140.1 (2013), pp. 125–161. DOI: 10.1007/s10107-012-0629-5.

[13] Ilya Sutskever et al. "On the importance of initialization and momentum in deep learning". In: *Proceedings of the 30th International Conference on Machine Learning (ICML)*. Vol. 28. Proceedings of Machine Learning Research. PMLR, 2013, pp. 1139–1147. URL: https://proceedings.mlr.press/v28/sutskever13.html.

[14] Michael Cohen, Jelena Diakonikolas, and Lorenzo Orecchia. "On Acceleration with Noise-Corrupted Gradients". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. 10–15 July. PMLR, July 2018, pp. 1019–1028. URL: https://proceedings.mlr.press/v80

[15] Zeyuan Allen-Zhu. "Katyusha: The first direct acceleration of stochastic gradient methods". In: *Journal of Machine Learning Research* 18.221 (2018), pp. 1–51. URL: http://jmlr.org/papers/v1

[16] Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. "A Universal Catalyst for First-Order Optimization". In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes et al. Vol. 28. Curran Associates, Inc., 2015. URL: https://proceedings.neurips.cc/paper_files/paper

[17] Roy Frostig et al. "Un-regularizing: Approximate proximal point and faster stochastic algorithms for empirical risk minimization". In: *International Conference on Machine Learning*. Vol. 37. Proceedings of Machine Learning Research. PMLR, 2015, pp. 2540–2548. URL: https://proceedings.mlr.press/v37/frostig15.html.

[18] Yunwen Lei et al. "Stochastic gradient descent for nonconvex learning without bounded gradient assumptions". In: *IEEE Transactions on Neural Networks and Learning Systems* 31.10 (2019), pp. 4394–4400. DOI: 10.1109/TNNLS.2019.2952217.

[19] Xiaoyu Li and Francesco Orabona. "On the convergence of stochastic gradient descent with adaptive stepsizes". In: *International Conference on Artificial Intelligence and Statistics*. Vol. 89. Proceedings of Machine Learning Research. PMLR, 2019, pp. 983–992. URL: https://proceedings.mlr.press/v89/li19c.html.

[20] Chiyuan Zhang et al. "Understanding deep learning (still) requires rethinking generalization". In: *Communications of the ACM* 64.3 (2021), pp. 107–115. DOI: 10.1145/3446776.

[21] Mingxing Tan and Quoc V. Le. "Efficientnet: Rethinking model scaling for convolutional neural networks". In: *Proceedings of the 36th International Conference on Machine Learning (ICML)*. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 6105–6114. URL: https://proceedings.mlr.press/v97/tan19a.html.

[22] Yanping Huang et al. "Gpipe: Efficient training of giant neural networks using pipeline parallelism". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 32. 2019, pp. 103–112. URL: https://proceedings.neurips.cc/paper/2019/hash/093f65e080a295f8076b1

[23] Alexander Kolesnikov et al. "Big transfer (bit): General visual representation learning". In: *Computer Vision – ECCV 2020: 16th European Conference*. Vol. 12350. Lecture Notes in Computer Science. Springer, 2020, pp. 491–507. DOI: 10.1007/978-3-030-58558-7_29.

[24] Mikhail Belkin et al. "Reconciling modern machine-learning practice and the classical bias–variance trade-off". In: *Proceedings of the National Academy of Sciences* 116.32 (2019), pp. 15849–15854. DOI: 10.1073/pnas.1903070116.

[25] Gao Huang et al. "Densely connected convolutional networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 4700–4708. DOI: 10.1109/CVPR.2017.243.

[26] Levent Sagun et al. "Empirical analysis of the Hessian of over-parametrized neural networks". In: *arXiv preprint* (2017). arXiv: 1706.04454 [cs.LG].

[27] Raef Bassily, Mikhail Belkin, and Siyuan Ma. *On exponential convergence of SGD in non-convex over-parametrized learning*. Tech. rep. arXiv, 2018. arXiv: 1802.06904 [cs.LG].

[28] Sharan Vaswani, Francis Bach, and Mark Schmidt. "Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron". In: *International Conference on Artificial Intelligence and Statistics*. Vol. 89. Proceedings of Machine Learning Research. PMLR, 2019, pp. 1195–1204. URL: https://proceedings.mlr.press/v89/vaswani19a.html.

[29] Chaoyue Liu et al. "Aiming towards the minimizers: fast convergence of SGD for over-parametrized problems". In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh et al. Vol. 36. Curran Associates, Inc., 2023, pp. 60748–60767. URL: https://proceedings.neurip

[30] Ahmed Khaled and Peter Richtárik. "Better theory for SGD in the nonconvex world". In: *arXiv preprint* (2020). arXiv: 2002.03329 [cs.LG].

[31] Robert M. Gower, Othmane Sebbouh, and Nicolas Loizou. "SGD for structured nonconvex functions: Learning rates, minibatching and interpolation". In: *International Conference on Artificial Intelligence and Statistics*. Vol. 130. Proceedings of Machine Learning Research. PMLR, 2021, pp. 1311–1319. URL: https://proceedings.mlr.press/v130/gower21a.html.

[32] Othmane Sebbouh, Robert M. Gower, and Aaron Defazio. "Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball". In: *Conference on Learning Theory*. Vol. 134. Proceedings of Machine Learning Research. PMLR, 2021, pp. 3932–3972. URL: https://proceedings.mlr.press/v134/sebbouh21a.html.

[33] Qiang Fu, Dongchu Xu, and Ashia Camage Wilson. "Accelerated stochastic optimization methods under quasar-convexity". In: *International Conference on Machine Learning*. Vol. 202. Proceedings of Machine Learning Research. PMLR, 2023, pp. 10535–10562. URL: https://proceedin

[34] Jikai Jin. "On the convergence of first order methods for quasar-convex optimization". In: *arXiv preprint* (2020). arXiv: 2010.04937 [cs.LG].

[35] Herbert Robbins and Sutton Monro. "A Stochastic Approximation Method". In: *The Annals of Mathematical Statistics* 22.3 (Sept. 1951), pp. 400–407. ISSN: 0003-4851. DOI: 10.1214/aoms/11777 URL: http://dx.doi.org/10.1214/aoms/1177729586.

[36] Léon Bottou. "Large-Scale Machine Learning with Stochastic Gradient Descent". In: *Proceedings of COMPSTAT'2010*. Ed. by Yves Lechevallier and Gilbert Saporta. Heidelberg: Physica-Verlag HD, 2010, pp. 177–186. ISBN: 978-3-7908-2604-3.

[37] Jeffrey Dean et al. "Large Scale Distributed Deep Networks". In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira et al. Vol. 25. Curran Associates, Inc., 2012. URL: https://proceedings.neurips.cc/paper_files/paper/2012/file/6aca97005c68f1206823

[38] Manzil Zaheer et al. "Deep Sets". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc

[39] Nitish Shirish Keskar et al. "On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima". In: *International Conference on Learning Representations*. 2017. URL: https://openreview.net/forum?id=H1oyRlYgg.

[40] Sebastian Ruder. "An overview of gradient descent optimization algorithms". In: *CoRR* abs/1609.04747 (2016). arXiv: 1609.04747. URL: http://arxiv.org/abs/1609.04747.

[41] Ziyu Chen, Yingzhou Li, and Xiuyuan Cheng. "SpecNet2: Orthogonalization-free spectral embedding by neural networks". In: *Proceedings of Mathematical and Scientific Machine Learning*. Vol. 190. Proceedings of Machine Learning Research. PMLR, 2022.

[42] Arthur Jacot, Franck Gabriel, and Clément Hongler. "Neural tangent kernel: Convergence and generalization in neural networks". In: *Advances in Neural Information Processing Systems*. Vol. 31. 2018.

[43] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. "A convergence theory for deep learning via over-parameterization". In: *International Conference on Machine Learning*. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 242–252.