Unified High-Probability Analysis of Stochastic Variance-Reduced Estimation

Zhankun Luo[‡*] **Antesh Upadhyay**[‡*] **M. Berk Sahin**[*] **Sang Bin Moon**[*]
Anuran Makur[⋆*] **Abolfazl Hashemi**[*]
[*]School of Electrical and Computer Engineering and
[⋆]Department of Computer Science
Purdue University
West Lafayette, IN 47907
{luo333, aantesh, sahinm, moon182, amakur, abolfazl}@purdue.edu

Abstract

Stochastic estimators are fundamental to large-scale optimization, where population quantities must be inferred from noisy oracle observations. Although influential methods such as momentum, SPIDER, STORM, and PAGE have been highly successful, their analyses are largely estimator-specific and expectation-based, obscuring the structural tradeoffs that determine reliability. In this paper, we develop a unified framework for stochastic variance-reduced estimation based on a recursion with three components: memory retention, reset probability, and a correction term for iterate movement. This framework recovers several classical estimators, motivates new second-order variants, and yields a bias-variance decomposition of estimation error. Our main result is a unified high-probability bound proved using a new dimension-free vector-valued Freedman inequality, valid for smooth normed spaces involving random sums of vector martingales. The result applies in both Euclidean and non-Euclidean settings, including the analysis of mirror-descent-based methods in Banach spaces. As applications, we obtain high-probability oracle complexities for unconstrained optimization with mirror descent, establishing the logarithmic dependence on the confidence level. We also derive the first $\tilde{\mathcal{O}}(\varepsilon^{-3})$ oracle-complexity bounds for stochastic optimization with expectation constraints, improving upon the existing $\tilde{\mathcal{O}}(\varepsilon^{-4})$ complexity by leveraging variance-reduced estimation for the first time in this setting.

1 Introduction

Stochastic optimization algorithms optimize a function over iterations by using noisy oracle information rather than exact population quantities. In the most familiar setting [116, 111, 104, 46], the quantity of interest is the population gradient, and standard first-order algorithms must repeatedly construct accurate gradient estimates from stochastic samples to make progress. From this perspective, the effectiveness of a first-order method depends not only on how it updates the iterate, but also on how well it maintains an estimator that is accurate, stable, and inexpensive to compute as the iterate evolves. Many widely used methods can be viewed in this way. Classical momentum stabilizes SGD by averaging past gradients [90], whereas variance-reduced methods, such as SVRG, STORM, and PAGE, reuse information across iterates through stochastic differences and occasional reset steps [76, 43, 33, 85, 130].

Although these methods are often presented as distinct algorithmic ideas, they address a common underlying problem: how to recursively estimate an evolving population quantity from noisy samples while controlling both bias and variance. Despite the empirical successes of these methods, however,

[*]The authors marked with [‡] contributed equally.

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

the theoretical picture remains fragmented. Existing analyses are typically developed separately for individual estimators, and their proofs are often tightly tied to the algebra of a particular recursion. As a result, it is difficult to isolate which design choices are actually responsible for performance: the reuse of past information, the frequency of resets, or the way past information is adjusted as the iterate moves. In addition, most existing guarantees are expectation-based. While such guarantees characterize average behavior, they can conceal rare but important estimation failures, which are especially relevant in large-scale learning and in constrained optimization.

In this paper, we develop a *unified framework for high-probability analysis of stochastic variance-reduced estimation in optimization*. We consider a general population quantity $g(w_t)$ observed through an unbiased stochastic oracle $G(w_t, \xi_t)$, where the iterate w_t changes over time t, and ξ_t denotes the oracle randomness. We introduce an estimator for $g(w_t)$ governed by *three simple design choices:* 1) how much past information must be retained, 2) how often resets using fresh estimates occur, and 3) how past information computed at w_{t-1} must be corrected so that it remains informative at the current iterate w_t. This viewpoint places several seemingly different recursive estimators within a single template and makes it possible to compare them through a common error decomposition. Beyond gradient estimation, the framework also applies to other optimization-relevant quantities.

$$\min_{w\in\mathcal{W}} f(w) \coloneqq \mathbb{E}_\xi[F(w,\xi)] \qquad (\dagger) \qquad\qquad \min_{w\in\mathcal{W}} \mathbb{E}_\xi[F(w,\xi)] \ \text{ s.t. } \ h(w) \coloneqq \mathbb{E}_\xi[H(w,\xi)] \le 0. \quad (\ddagger)$$

The two representative examples above illustrate the scope of our framework. In problem (†), the relevant quantity is the population gradient $g(w) = \nabla f(w)$, and the goal is to maintain an accurate stochastic estimate of the gradient along the optimization trajectory. In problem (‡), however, the quantity of interest is the constraint value $h(w)$, which must be estimated from noisy samples to assess feasibility [82]. In this setting, poor estimation can lead either to unreliable feasibility detection or to overly conservative large-batch procedures. This paper, therefore, studies a broader estimation problem in stochastic optimization, with stochastic gradient estimation and expectation-constraint estimation as two central instances.

Another limitation of the existing literature is that high-probability guarantees are predominantly restricted to Euclidean geometries. In more general first-order methods, such as mirror descent [14, 12, 22], the estimator naturally lives in the dual space rather than the primal space. The literature currently lacks a unified, high-probability concentration framework that accommodates arbitrary smooth Banach spaces without relying on estimator-specific algebra. Our analysis in this paper is developed at a level of generality that covers both standard Euclidean updates and generalized first-order methods such as mirror descent, while focusing on high-probability guarantees. Please see Appendix G for a detailed discussion of related work.

Contributions. Our main contributions are as follows:

- **Unified estimator framework.** We propose a single recursive stochastic estimator ($*$) that subsumes classical momentum, STORM, PAGE, and SPIDER as special cases, while also suggesting new second-order variants; see Table 1.
- **Unified high-probability analysis.** We prove a general high-probability bound for the estimation error of every estimator in the framework, together with an explicit bias-variance decomposition; see Theorem 4.2. The proof relies on a new vector-valued Freedman inequality; see Theorem B.2.
- **Unconstrained optimization.** We establish high-probability guarantees for finding first-order ε-stationary points in stochastic optimization with mirror descent, and we derive the resulting oracle complexities for different estimator choices; see Table 5 and Theorem E.1 in Appendix E
- **Constrained optimization.** We provide the first variance-reduced oracle complexity guarantees for expectation-constrained stochastic optimization, improving the rate from $\tilde{\mathcal{O}}(\varepsilon^{-4})$ in [82] to $\tilde{\mathcal{O}}(\varepsilon^{-3})$; see Table 3.

2 Preliminaries

We briefly introduce the notations and setup used throughout the paper. Let $(\mathcal{X}, \|\cdot\|)$ be a reflexive Banach space, and let $(\mathcal{X}^*, \|\cdot\|_*)$ denote its dual space of all continuous bounded linear functionals. For $u \in \mathcal{X}^*$ and $x \in \mathcal{X}$, we write $\langle u, x\rangle$ for the canonical duality pairing between $\mathcal{X}^*$ and $\mathcal{X}$ such

that $\langle \cdot, \cdot \rangle : \mathcal{X}^* \times \mathcal{X} \to \mathbb{R}$. The dual norm is defined by $\|u\|_* := \sup_{\|x\| \le 1, x \in \mathcal{X}} \langle u, x \rangle$ for any $u \in \mathcal{X}^*$. The optimization iterates lie in a nonempty closed, convex set $\mathcal{W} \subseteq \mathcal{X}$.

Definition 2.1 (Mirror Map)**.** The mirror map $\Phi : \mathcal{D} \to \mathbb{R} \cup \{+\infty\}$ with $\mathcal{D} \subseteq \mathcal{X}$ is proper, weakly lower semi-continuous (w.l.s.c), and 1-strongly convex on a nonempty closed convex set $\mathcal{W} \subseteq \text{int}(\mathcal{D})$ with respect to $\|\cdot\|$. Furthermore, Φ is Gâteaux differentiable on $\mathcal{W}$.

Definition 2.2 (Proximal Gradient Mapping)**.** For $w \in \mathcal{W}$, an input vector $u \in \mathcal{X}^*$, and a step size $\eta > 0$, the **proximal gradient mapping** is $P(w, u, \eta) = (w - w^+)/\eta$, where the updated point w^+ is the unique minimizer of the proximal subproblem: $w^+ = \arg\min_{w' \in \mathcal{W}} \langle \eta u, w' \rangle + D_\Phi(w', w)$ and $D_\Phi(w', w)$ denotes the Bregman divergence induced by Φ.

The existence and uniqueness are guaranteed by variational principles. The strong convexity of Φ renders the objective strictly convex and coercive, while its closed convexity ensures it is w.l.s.c. [42, Cor. 2.2, p. 11]. By the Mazur's Lemma and Eberlein-Šmulian theorem [21, Thm. 3.7, 3.18], coercivity within the reflexive space $\mathcal{X}$ forces bounded sequences to converge weakly to a feasible limit inside the closed set $\mathcal{W}$. Consequently, the objective reliably attains a unique global minimum [42, Prop. 1.2, p. 35], [11, Thm. 2.11, p. 72]. Based on variational inequality as supported by [42], we derive properties of $P(w, u, \eta)$ (see Lemma A.7 and [47, 12]), including $\|P(w, u, \eta)\| \le \|u\|_*$. For a more detailed discussion on the existence, uniqueness and properties of the proximal gradient mapping $P(w, u, \eta)$, please refer to Remark A.5 and Lemma A.7 in Appendix A.

We consider a stochastic oracle $G : \mathcal{W} \times \Omega \to \mathcal{G}$ and its population counterpart $g(w) := \mathbb{E}[G(w, \xi)]$, where ξ denotes the oracle randomness, $\mathcal{G}$ is a κ-smooth ($\kappa \ge 1$) normed space with its norm $\|\cdot\|_{\mathcal{G}}$. For high-probability analysis, we introduce $C(\delta, \kappa) := \sqrt{\kappa} + \sqrt{3\ln(1/\delta)}$ for $\delta \in (0, 1)$, arising from our proposed vector-valued Freedman inequality (see Theorem B.2). Our goal is to estimate the time-varying target $g(w_t)$ along a trajectory $\{w_t\}_{t \ge 0}$. Given a recursive estimator v_t, we define the estimation error $e_t := v_t - g(w_t)$. Our analysis is not restricted to Euclidean SGD and applies more broadly to generalized first-order methods, including mirror descent [14, 12, 22].

Mirror Descent Update. At each iteration t, we form a unified estimator v_t and use its corresponding update vector U_t, to perform a mirror descent update with step size $\eta_t > 0$:

$$w_{t+1} = \arg\min_{w \in \mathcal{W}} \{\eta_t \langle U_t, w \rangle + D_\Phi(w, w_t)\} = w_t - \eta_t P(w_t, U_t, \eta_t). \tag{1}$$

Assumption 2.3 (Boundedness of Generalized Updates)**.** Let U_t be the update vector computed at iteration t. We assume the algorithm employs a normalization or clipping mechanism such that the update vector is bounded almost surely, satisfying $\|U_t\|_* \le G$ for all $t \ge 0$.

We note that this boundedness assumption, while necessary, is not as strong since in many problems it can be ensured by applying normalization or clipping operations (e.g., Application 1 in Subsection 6.1), or holds due to certain Lipschitzness conditions (e.g., Application 2 in Subsection 6.2). Under the above assumption, we show that $\|w_{t+1} - w_t\| = \eta_t \|P(w_t, U_t, \eta_t)\| \le \eta_t \|U_t\|_* \le \eta_t G$.

3 A Unified Estimator

In stochastic optimization, estimating an evolving population quantity $g(w_t)$ from noisy observations $G(w_t, \xi_t)$ requires combining fresh information with historical memory. A classical example is momentum [90] that filters noise via $v_t = (1 - \beta)G(w_t, \xi_t) + \beta v_{t-1}$. If we define the estimation error by $e_t := v_t - g(w_t)$, then the corresponding error recursion by $e_t = \beta e_{t-1} + (1-\beta)(G(w_t, \xi_t) - g(w_t)) - \beta(g(w_t) - g(w_{t-1}))$. The final term captures the deterministic lag induced by reusing information from the previous iterate. Variance-reduced estimators, such as STORM and SPIDER, avoid this deterministic bias by incorporating stochastic differences. This suggests a common design principle for recursive estimators: once past information is carried over from iteration $t-1$, it should be corrected before being reused at the new iterate w_t. We make this principle explicit through an explicit *correction term* $\mathcal{T}_t$, which leads to a unified recursion governed by three design choices: memory retention β_t, the reset probability p_t, and the correction mechanism $\mathcal{T}_t$.

The Unified Recursion. Let $\{\mathcal{F}_t\}_{t \ge 0}$ be the filtration generated by oracle noise and reset events (see Appendix A). We assume parameter predictability and sub-Gaussian oracle noise stated below.

Assumption 3.1 (Parameter Predictability)**.** The parameters η_{t-1}, $p_t \in [0, 1]$, and $\beta_t \in (0, 1]$ are $\mathcal{F}_{t-1}$-adapted. The reset indicator $b_t \sim \text{Bernoulli}(p_t)$ is sampled at time t, ensuring $\mathbb{E}[b_t | \mathcal{F}_{t-1}] = p_t$.

Assumption 3.2 (Sub-Gaussianity)**.** The oracle is unbiased, $\mathbb{E}[G(w_t, \xi) \mid \mathcal{F}_{t-1}] = g(w_t)$. Furthermore, for any fresh sample ξ drawn at time t (whether a single sample ξ_t or a component $\xi_t^{(i)}$ of a batch of B independent samples $\xi_t^B \equiv (\xi_t^{(i)})_{i=1}^B$), the noise is conditionally sub-Gaussian with variance proxy $\sigma^2 > 0$, satisfying $\mathbb{E}[\exp(\|G(w_t, \xi) - g(w_t)\|_G^2/\sigma^2) \mid \mathcal{F}_{t-1}] \leq 2$.

At iteration t, given the reset indicator $b_t \sim \text{Bernoulli}(p_t)$, we define the unified estimator as:

$$v_t := \begin{cases} G(w_t, \xi_t^B), & b_t = 1, \\ G(w_t, \xi_t) + \beta_t \Delta_t + \mathcal{T}_t(w_t, w_{t-1}, \xi_t), & b_t = 0, \end{cases} \tag{$*$}$$

where ξ_t^B represents a fresh batch of samples, ξ_t is a single stochastic sample, and $\Delta_t := v_{t-1} - G(w_{t-1}, \xi_t)$, $G(w_t, \xi_t^B) := \frac{1}{B}\sum_{i=1}^B G(w_t, \xi_t^{(i)})$ is the average of evaluated function values $G(w_t, \xi_t^{(i)}), i = 1, \ldots, B$. Viewed through this lens, $\mathcal{T}_t$ is the mechanism that corrects stale information from the previous iterate. Different choices of $\mathcal{T}_t$ recover classical estimators and naturally suggest new ones.

3.1 Error Dynamics and Effective Innovation

We next derive the common error recursion induced by ($*$). When a reset occurs, i.e., $b_t = 1$, the estimator is reinitialized using a fresh stochastic estimate. The more interesting case is the recursive steps $b_t = 0$, where past estimation error is propagated and modified by the correction mechanism.

Expanding $e_t = v_t - g(w_t)$ and substituting $v_{t-1} = e_{t-1} + g(w_{t-1})$ yields the core error recursion $e_t = \beta_t e_{t-1} + \mathcal{I}_t(\mathcal{T})$. The estimation error thus consists of a contracted past error and a stochastic source term, which we define as the *Effective Innovation*.

Definition 3.3 (Effective Innovation)**.** The effective innovation $\mathcal{I}_t(\mathcal{T})$ at iteration t is defined as

$$\mathcal{I}_t(\mathcal{T}) := (1-\beta_t)\underbrace{(G(w_t, \xi_t) - g(w_t))}_{\text{Direct Noise}} + \beta_t \underbrace{\Delta(w_t, w_{t-1}, \xi_t)}_{\text{Difference Noise}} + \mathcal{T}_t(w_t, w_{t-1}, \xi_t), \tag{2}$$

where $\Delta(w_t, w_{t-1}, \xi_t) := (G(w_t, \xi_t) - G(w_{t-1}, \xi_t)) - (g(w_t) - g(w_{t-1}))$ is the centered difference.

The decomposition in (2) makes the bias-variance structure of the unified recursion. The conditional expectation of $\mathcal{I}_t(\mathcal{T})$ determines the predictable bias, while its variability drives the variance.

Specific choices of the correction term $\mathcal{T}_t$. By specifying $\mathcal{T}_t$, we recover existing variance-reduced methods and suggest new higher-order variants, summarized in Table 1.

• **Family 1: Zeroth-Order Correction (Standard Recursion and Resetting).** Setting $\mathcal{T}_t := \beta_t\big(G(w_{t-1}, \xi_t) - G(w_t, \xi_t)\big)$ in ($*$) recovers the momentum-style update $v_t = (1-\beta_t)G(w_t, \xi_t) + \beta_t v_{t-1}$. This attenuates direct oracle noise through averaging, but introduces a deterministic bias induced by iterate movement, since information formed at w_{t-1} is reused at w_t.

• **Family 2: First-Order Correction (Differential Recursion and Resetting).** By setting $\mathcal{T}_t := 0$, the recursion in ($*$) is driven entirely by stochastic differences. This eliminates the deterministic drift inherent to momentum updates, leaving an estimation error governed solely by centered difference noise that scales with $\|w_t - w_{t-1}\|$. This formulation recovers STORM ($\beta_t < 1$), PAGE ($p_t > 0, \beta_t = 1$), and SPIDER (periodic resets).

• **Family 3: Second-Order Correction (Second-Order Recursion and Resetting).** Letting $\mathcal{T}_t$ encode a Taylor correction, $\mathcal{T}_t := \beta_t\big(G(w_{t-1}, \xi_t) - [G(w_t, \xi_t) + \nabla G(w_t, \xi_t)(w_{t-1} - w_t)]\big)$, incorporates local curvature information in ($*$). This reduces the leading-order deterministic bias to a second-order residual, while the remaining stochastic error is governed by the Hessian-vector product.

4 Unified Estimation Error Bound

We now develop a high-probability bound for the unified estimator introduced in Section 3. The error recursion derived in Section 3.1 shows that the estimation error is driven by the accumulation of effective innovations over epochs delimited by reset times. Our goal in this section is to control this accumulation uniformly with high probability. We first impose a regularity condition on the effective innovation, separating bias from the centered stochasticity and enabling martingale concentration.

Table 1: Special cases of the unified estimator. Here, $\Delta G := \beta_t(G(w_{t-1},\xi_t) - G_t(w_t,\xi_t))$, $2^{\text{nd}} := \beta_t\big(G(w_{t-1},\xi_t) - [G(w_t,\xi_t) + \nabla G(w_t,\xi_t)(w_{t-1}-w_t)]\big)$, and $\mathbb{I}_{\mathsf{Sched}}$ refers to a deterministic schedule (e.g., $t \mod E = 0$). Bolded entries are **new** estimators. **SO** denotes Second-Order.

Case		Standard	Differential	Second-Order
β_t	p_t	$\mathcal{T}_t = \Delta G$	$\mathcal{T}_t = 0$	$\mathcal{T}_t = 2^{\text{nd}}$
1. β	0	Momentum	STORM	SO Momentum
2. 1	p	**Probabilistic Momentum**	PAGE/ Loopless SVRG	**SO PAGE**
3. 1	$\mathbb{I}_{\mathsf{Sched}}$	**Periodic Momentum**	SPIDER/SVRG	**SO SPIDER**

Assumption 4.1 (Conditional Regularity of Innovation)**.** We assume there exist non-negative, predictable processes $\{\mathcal{B}_t\}_{t\geq 1}$ and $\{\Sigma_t\}_{t\geq 1}$ adapted to $\mathcal{F}_{t-1}$ such that, almost surely

i) The conditional bias is bounded by $\mathcal{B}_t$, i.e., $\|\mathbb{E}[\mathcal{I}_t(\mathcal{T}) \mid \mathcal{F}_{t-1}]\|_{\mathcal{G}} \leq \mathcal{B}_t$.

ii) The centered innovation $\mathcal{Z}_t := \mathcal{I}_t(\mathcal{T}) - \mathbb{E}[\mathcal{I}_t(\mathcal{T}) \mid \mathcal{F}_{t-1}]$ is sub-Gaussian conditioned on $\mathcal{F}_{t-1}$ with a random, predictable variance proxy Σ_t^2. That is, $\mathbb{E}\left[\exp\left(\frac{\|\mathcal{Z}_t\|_{\mathcal{G}}^2}{\Sigma_t^2}\right)\middle|\mathcal{F}_{t-1}\right] \leq 2$. We also assume that on the $\mathcal{F}_{t-1}$-measurable event $\{\Sigma_t = 0\}$, $\mathcal{Z}_t = 0$ almost surely.

Assumption 4.1 controls the effective innovation by bounding its conditional bias $\mathcal{B}_t$ and its centered stochastic component via the variance proxy Σ_t^2. We now state the main theorem for estimation error.

Main theorem 4.2 (Unified Estimation Error Bound)**.** *Consider the unified estimator* $(*)$ *governed by predictable sequences* $\{\beta_t\}_{t\geq 1}, \{p_t\}_{t\geq 1}$ *and* $\{\eta_{t-1}\}_{t\geq 1}$ *satisfying Assumption 3.1. Let* τ_m *denote the reset times, defined recursively by* $\tau_0 = 0$, $\tau_m = \inf\{t > \tau_{m-1} : b_t = 1\}$ *for* $1 \leq m \leq T$,, *and let* N_t *denote the counting process for the number of resets up to time* t *such that* $N_t := \sum_{s=1}^t b_s$, *and define the cumulative momentum factors* $\Lambda_{t,j} := \prod_{i=j+1}^t \beta_i$, $A_t := \prod_{i=1}^t \beta_i^{-1}$, *with the convention* $\Lambda_{t,t} = 1$. *Under Assumptions 3.2, 4.1, for any* $\delta \in (0,1)$ *and any deterministic variance budget* $V_t \geq 0$, *with probability at least* $1-\delta$, *the condition* $\mathfrak{V}_t^2 := \sum_{j=\tau_{N_t}+1}^t A_j^2 \Sigma_j^2 \leq V_t$ *implies that the estimation error for any* $t \in [T], T \geq 2$ *satisfies*

$$\|e_t\|_{\mathcal{G}} \leq \Lambda_{t,\tau_{N_t}} C\left(\frac{\delta}{2T}, \kappa\right) \sigma/\sqrt{B} + \sum_{j=\tau_{N_t}+1}^t \Lambda_{t,j}\mathcal{B}_j + \sqrt{2}C\left(\frac{\delta}{2T}, \kappa\right) A_t^{-1}\sqrt{V_t}. \quad (3)$$

The bound in Theorem 4.2 decomposes the estimation error into three terms. The first term captures the effect of the most recent reset or batch initialization and decays according to the cumulative momentum since that reset. The second term accumulates the conditional bias of the effective innovations through the quantities $\mathcal{B}_j$. The third term is a high-probability concentration term controlled by the predictable variance accumulated over the current reset interval. This decomposition makes explicit how reset probability, momentum strength, and the choice of correction term jointly determine the quality of the estimator. The complete details of the proof are presented in Appendix B. The technical difficulty in proving this result requires establishing concentration bounds for sums of martingale difference sequences over random reset times. To handle this, we combine a masking argument for stopping-time intervals (see Lemma B.1) with the vector-valued Freedman inequality (see Theorem B.2), which is the main concentration tool underlying the proof.

5 Instantiations of the Unified Estimation Error Bound

Theorem 4.2 reduces the analysis of any estimator in our unified framework to controlling two quantities associated with the *Effective Innovation*: the conditional bias bound $\mathcal{B}_t$ and the predictable variance proxy Σ_t^2 in Assumption 4.1. In this section, we instantiate these quantities for the three families. The resulting bounds reveal how $\mathcal{T}_t$ dictates the bias-variance tradeoff. Our standing assumptions from previous sections remain in play throughout. In particular, we use Assumption 2.3 to control the displacement of consecutive iterates, Assumption 3.1 for the predictability of the control sequences, and Assumption 3.2 for the sub-Gaussianity of the oracle. The only additional ingredients

needed in this section are family-specific regularity conditions that allow us to identify $\mathcal{B}_t$ and Σ_t^2 in Assumption 4.1 concretely.

Beyond Assumptions 2.3, 3.1, and 3.2, we require the additional assumptions associated with the three estimator families. These enable us to satisfy Assumption 4.1 and apply Theorem 4.2 to bound the estimation error $\|e_t\|_{\mathcal{G}}$. All proofs are available in Appendix C.

5.1 Family 1: Zeroth-Order Correction

We begin with the zeroth-order family, which recovers the momentum-style update, where the correction term is $\mathcal{T}_t(w_t, w_{t-1}, \xi_t) = \beta_t\big(G(w_{t-1}, \xi_t) - G(w_t, \xi_t)\big)$, and its effective innovation is $\mathcal{I}_t(\mathcal{T}) = (1-\beta_t)(G(w_t, \xi_t) - g(w_t)) - \beta_t(g(w_t) - g(w_{t-1}))$.

Assumption 5.1 (Lipschitz Continuity)**.** The mean function g is L-Lipschitz continuous, satisfying $\|g(w) - g(w')\|_{\mathcal{G}} \le L\|w - w'\|$ for all $w, w' \in \mathcal{W}$.

Assumption 5.1 controls the drift caused by reusing past information at w_{t-1} while the target has moved to $g(w_t)$. Under Assumption 5.1, we show that the conditional regularity of the innovation (Assumption 4.1) is satisfied. This holds because the conditional bias $\mathcal{B}_t$ is bounded as follows:

$$\|\mathbb{E}[\mathcal{I}_t(\mathcal{T}) \mid \mathcal{F}_{t-1}]\|_{\mathcal{G}} = \beta_t\|\mathbb{E}[g(w_t) - g(w_{t-1}) \mid \mathcal{F}_{t-1}]\|_{\mathcal{G}} \le \beta_t\eta_{t-1}LG =: \mathcal{B}_t,$$

and the predictable variance proxy for $\mathcal{Z}_t = \mathcal{I}_t(\mathcal{T}) - \mathbb{E}[\mathcal{I}_t(\mathcal{T}) \mid \mathcal{F}_{t-1}] = (1-\beta_t)(G(w_t, \xi_t) - g(w_t))$ evaluates to $\Sigma_t^2 = (1-\beta_t)^2\sigma^2$ (see Proposition C.1 and its proof in Appendix C).

5.2 Family 2: First-Order Correction

We next consider the first-order differential family, corresponding to the choice $\mathcal{T}_t(w_t, w_{t-1}, \xi_t) = 0$ and its effective innovation is $\mathcal{I}_t(\mathcal{T}) = (1-\beta_t)(G(w_t, \xi_t) - g(w_t)) + \beta_t\Delta(w_t, w_{t-1}, \xi_t)$. This family includes STORM, PAGE, and SPIDER-type estimators as special cases. The key feature of this recursion is that the deterministic drift from the moving target is removed and replaced by the centered difference noise.

Assumption 5.2 (Sub-Gaussian Lipschitzness of the Centered Differences)**.** For the analysis of variance-reduced estimators that rely on estimate differences, such as STORM, we require a stronger tail condition on the continuity of the noise. Let $\Delta(w, w', \xi) := (G(w, \xi) - G(w', \xi)) - (g(w) - g(w'))$ denote the centered estimate difference noise. We assume the stochastic oracle is ℓ-sub-Gaussian Lipschitz, meaning that for all $w, w' \in \mathcal{W}$, $\mathbb{E}\left[\exp\left(\frac{\|\Delta(w,w',\xi)\|_{\mathcal{G}}^2}{\ell^2\|w-w'\|^2}\right)\right] \le 2$.

Assumption 5.2 is tailored to first-order differential recursions. It ensures that stochastic differences concentrate with a variance proxy proportional to the squared distance between iterates (see two concrete examples for Assumption 5.2 in Subsection C.5, Appendix C). Under Assumption 5.2, the innovation satisfies the conditional regularity condition in Assumption 4.1. Specifically, this holds with a conditional bias of $\mathcal{B}_t = \|\mathbb{E}[\mathcal{I}_t(\mathcal{T}) \mid \mathcal{F}_{t-1}]\|_{\mathcal{G}} = 0$ and a predictable variance proxy of $\Sigma_t^2 = 2(1-\beta_t)^2\sigma^2 + 2\beta_t^2\ell^2\eta_{t-1}^2G^2$ (see Proposition C.2 and its proof in Appendix C).

Specific choices of (p_t, β_t) smoothly recover standard algorithms: STORM when $p_t = 0$, $\beta_t < 1$, loopless SVRG/PAGE when $p_t \in (0,1)$, $\beta_t = 1$, and SPIDER/SVRG with periodic resets and $\beta_t = 1$. In the fully differential regime ($\beta_t = 1$), the direct-noise term vanishes, leaving only the difference-noise variance proxy $\Sigma_t^2 = 2\ell^2\eta_{t-1}^2G^2$.

5.3 Family 3: Second-order Correction

We finally consider the second-order family, where the correction term incorporates local curvature information $\mathcal{T}_t(w_t, w_{t-1}, \xi_t) := \beta_t\left(G(w_{t-1}, \xi_t) - [G(w_t, \xi_t) + \nabla G(w_t, \xi_t)(w_{t-1} - w_t)]\right)$, and its effective innovation is $\mathcal{I}_t(\mathcal{T}) = (1-\beta_t)(G(w_t, \xi_t) - g(w_t)) + \beta_t(g(w_{t-1}) - g(w_t) - \nabla G(w_t, \xi_t)(w_{t-1} - w_t))$. This correction is designed to remove the leading-order drift term and leave only a second-order Taylor residual, therefore we introduce $L(\mathcal{X}, \mathcal{G})$ as the set of all bounded linear operators from $\mathcal{X}$ to $\mathcal{G}$, and for any $T \in L(\mathcal{X}, \mathcal{G})$ and its norm is defined by $\|T\|_{L(\mathcal{X},\mathcal{G})} := \sup_{\|x\|\le 1, x\in\mathcal{X}} \|Tx\|_{\mathcal{G}}$.

Assumption 5.3 (Sub-Gaussian Hessian Noise)**.** Let $\nabla G(w, \xi)$ denote the stochastic operator returned by the oracle such that $\mathbb{E}_\xi[\nabla G(w, \xi)] = \nabla g(w)$. We assume the noise in Hessian-vector products is sub-Gaussian. Specifically, $\mathbb{E}\left[\exp\left(\|\nabla G(w, \xi) - \nabla g(w)\|_{L(\mathcal{X},\mathcal{G})}^2/\gamma^2\right)\right] \le 2$ for any $w \in \mathcal{W}$.

Table 2: Family-wise instantiations of Theorem 4.2. Each row corresponds to a representative choice of (β_t, p_t) within a family, assuming a constant step size $\eta_t = \eta$. For deterministic periodic resets (Case 3), the initialization term is union-bounded over the T/E scheduled epochs to yield $C(\delta/(2T/E), \kappa)$, though $C(\delta/(2T), \kappa)$ from Theorem 4.2 remains valid up to logarithmic factors.

Algorithm	β_t	p_t	**Upper Bound on** $\|e_t\|_{\mathcal{G}},\ \forall t \in [T]$
Family 1: Standard Recursion and Resetting (Zeroth-Order) **Assumptions:** 2.3,3.1,3.2,5.1			
Momentum	β	0	$\beta^t C\left(\frac{\delta}{2T}, \kappa\right) \sigma/\sqrt{B} + \frac{GL\eta}{1-\beta} + C\left(\frac{\delta}{2T}, \kappa\right) \sigma\sqrt{1-\beta}$
Probabilistic Momentum	1	p	$C\left(\frac{\delta}{4T}, \kappa\right) \sigma/\sqrt{B} + \frac{GL\eta}{p} \log \frac{4T}{\delta}$
Periodic Momentum	1	$\mathbb{I}_{\{t \bmod E=0\}}$	$C\left(\frac{\delta}{2T/E}, \kappa\right) \sigma/\sqrt{B} + GL\eta E$
Family 2: Differential Recursion and Resetting (First-Order) **Assumptions:** 2.3,3.1,3.2,5.2			
STORM	β	0	$\beta^t C\left(\frac{\delta}{2T}, \kappa\right) \sigma/\sqrt{B} + C\left(\frac{\delta}{2T}, \kappa\right) \sqrt{\frac{2(1-\beta)^2\sigma^2 + 2\beta^2\ell^2\eta^2G^2}{1-\beta^2}}$
PAGE / Loopless SVRG	1	p	$C\left(\frac{\delta}{4T}, \kappa\right) \sigma/\sqrt{B} + C\left(\frac{\delta}{4T}, \kappa\right) \ell\eta G\sqrt{\frac{2}{p} \log \frac{4T}{\delta}}$
SPIDER / SVRG	1	$\mathbb{I}_{\{t \bmod E=0\}}$	$C\left(\frac{\delta}{2T/E}, \kappa\right) \sigma/\sqrt{B} + C\left(\frac{\delta}{2T}, \kappa\right) \ell\eta G\sqrt{2E}$
Family 3: Second-Order Recursion and Resetting (Second-Order) **Assumptions:** 2.3,3.1,3.2,5.3,5.4			
Second-Order Momentum	β	0	$\beta^t C\left(\frac{\delta}{2T}, \kappa\right) \sigma/\sqrt{B} + \frac{\alpha\beta\eta^2G^2}{2(1-\beta)} + C\left(\frac{\delta}{2T}, \kappa\right) \sqrt{\frac{2(1-\beta)^2\sigma^2 + 2\beta^2\gamma^2\eta^2G^2}{1-\beta^2}}$
Second-Order PAGE	1	p	$C\left(\frac{\delta}{4T}, \kappa\right) \sigma/\sqrt{B} + \frac{\alpha\eta^2G^2}{2p} \log \frac{4T}{\delta} + C\left(\frac{\delta}{4T}, \kappa\right) \gamma\eta G\sqrt{\frac{2}{p} \log \frac{4T}{\delta}}$
Second-Order SPIDER	1	$\mathbb{I}_{\{t \bmod E=0\}}$	$C\left(\frac{\delta}{2T/E}, \kappa\right) \sigma/\sqrt{B} + \frac{\alpha\eta^2G^2E}{2} + C\left(\frac{\delta}{2T}, \kappa\right) \gamma\eta G\sqrt{2E}$

Assumption 5.4 (Smoothness)**.** g is α-smooth. That is, $\|\nabla g(w) - \nabla g(w')\|_{L(\mathcal{X},\mathcal{G})} \le \alpha\|w - w'\|$, where ∇g is the Jacobian of g, or the Hessian of the objective if $g = \nabla f$, bounded in spectral norm.

Assumptions 5.3 and 5.4 allow us to control the stochastic noise in the Hessian-vector products and bound the second-order Taylor remainder, such that $\|g(w_{t-1}) - g(w_t) - \nabla g(w_t)(w_{t-1} - w_t)\|_{\mathcal{G}} \le \frac{\alpha}{2}\|w_t - w_{t-1}\|^2$. Under these assumptions, the conditional regularity of the innovation (Assumption 4.1) holds with a conditional bias bounded by $\mathcal{B}_t$ since

$$\|\mathbb{E}[\mathcal{I}_t(\mathcal{T}) \mid \mathcal{F}_{t-1}]\|_{\mathcal{G}} = \beta_t\|g(w_{t-1}) - g(w_t) - \nabla g(w_t)(w_{t-1} - w_t)\|_{\mathcal{G}} \le \frac{\alpha}{2}\beta_t\eta_{t-1}^2G^2 =: \mathcal{B}_t,$$

and the predictable variance proxy for $\mathcal{Z}_t = \mathcal{I}_t(\mathcal{T}) - \mathbb{E}[\mathcal{I}_t(\mathcal{T}) \mid \mathcal{F}_{t-1}] = (1-\beta_t)(G(w_t, \xi_t) - g(w_t)) - \beta_t(\nabla G(w_t, \xi_t) - \nabla g(w_t))(w_{t-1} - w_t)$ is given by $\Sigma_t^2 = 2(1-\beta_t)^2\sigma^2 + 2\beta_t^2\gamma^2\eta_{t-1}^2G^2$ (see Proposition C.3 and its proof in Appendix C).

6 Optimization Guarantees from Unified Estimation

We now show how the unified estimation bound translates into optimization guarantees in two representative settings. The key point is that Theorem 4.2 reduces the analysis to obtaining a uniform high-probability bound on the estimator error. Once such a bound is available, the remaining argument becomes application-specific. In unconstrained stochastic optimization, this yields first-order stationarity guarantees for mirror descent. In expectation-constrained stochastic optimization, it yields feasibility and optimality guarantees for the optimization algorithm.

6.1 Application 1: Stochastic Optimization with Mirror Descent

We consider the stochastic optimization problem (†) of minimizing $f(w) = \mathbb{E}_\xi[F(w, \xi)]$ over a closed convex set $\mathcal{W}$, where we assume f is bounded below such that $\inf_{w\in\mathcal{W}} f(w) > -\infty$. We define the true objective gradient as $g(w) := \nabla f(w) = \nabla\mathbb{E}_\xi[F(w, \xi)]$ and its unbiased stochastic estimator as $G(w, \xi)$. To relate the unified estimator v_t of the objective gradient $g(w_t)$ to the update vector U_t, we select $(\mathcal{G}, \|\cdot\|_{\mathcal{G}}) := (\mathcal{X}^*, \|\cdot\|_*)$ and choose $U_t = v_t/\|v_t\|_*$ as the normalized direction of v_t. We then perform a mirror descent update $w_{t+1} = w_t - \eta_t P(w_t, U_t, \eta_t)$ using a constant step size $\eta_t = \eta > 0$; since the update vector $\|U_t\|_* = 1$, Assumption 2.3 is satisfied with $G = 1$.

Stationarity Criterion. For any choice of the correction term $\mathcal{T}_t$, we measure convergence using the criterion $\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(w_t)\|_*\|P_t\|^2 \le \varepsilon$ for mirror descent with a normalized update vector, where the *proximal stationarity witness* is defined as $P_t \coloneqq P\big(w_t, \nabla f(w_t)/\|\nabla f(w_t)\|_*, \frac{\eta}{2}\big)$. For an unconstrained Hilbert space (e.g., $\mathcal{X} = \mathbb{R}^d$ with $\mathcal{W} = \mathcal{X}$) under the ℓ_2-norm, selecting the mirror map $\Phi(w) = \frac{1}{2}\|w\|_2^2$ yields exactly $P(w,u,\eta) = u$. This implies $P_t = \nabla f(w_t)/\|\nabla f(w_t)\|_2$ and $\|P_t\| = 1$, meaning our results with above stationary criterion recover the classic theoretical guarantees for the ε-stationarity of $\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(w_t)\|_2$ [90, 33, 31, 32, 85, 43]. For comparison, STORM-type algorithm that uses the gradient estimator directly as the update vector [139] exhibits a comparable oracle complexity on the order of ε^{-3}, while relying on $\|P(w, \nabla f(w), \eta)\|$ as its criterion for first-order stationarity.

Oracle / Iteration Complexities. For brevity, the results of oracle / iteration complexities for all special cases of these three families are summarized in Table 5 with different selections of β_t, p_t and $\eta_t = \eta$ as listed in Table 6, Appendix E, where we define the initial suboptimality gap as $\Delta_f \coloneqq f(w_1) - \inf_{w\in\mathcal{W}} f(w)$, and introduce the following shorthand constants:

$$\sigma_L \coloneqq (\Delta_f L)^{\frac{1}{2}}, \quad \sigma_\ell \coloneqq (\Delta_f \ell)^{\frac{1}{2}}, \quad \sigma_\gamma \coloneqq (\Delta_f \gamma)^{\frac{1}{2}}, \quad \sigma_\alpha \coloneqq (\Delta_f^2 \alpha)^{\frac{1}{3}}. \tag{4}$$

The parameters L, ℓ, γ, and α refer to constants in Assumptions 5.1, 5.2, 5.3, and 5.4, respectively.

Theorem 6.1 (Complexities, Informal)**.** Under the parameter selections in Table 6, the oracle and iteration complexities N, T to achieve ε-stationarity with high probability are:
Family 1: **SGD-M:** $\tilde{\mathcal{O}}\big(\frac{\sigma_L^2\sigma^2}{\varepsilon^4} \vee \frac{\sigma^3}{\varepsilon^3}\big)$;
Probabilistic Momentum: $\tilde{\mathcal{O}}\big(\frac{\sigma_L^2\sigma^2}{\varepsilon^4}\big)$; **Periodic Momentum:** $\tilde{\mathcal{O}}\big(\frac{\sigma_L^2\sigma^2}{\varepsilon^4}\big)$.
Family 2: **STORM:** $\tilde{\mathcal{O}}\big(\frac{\sigma^3}{\varepsilon^3} \vee \frac{\sigma_\ell^2\sigma}{\varepsilon^3} \vee \frac{\sigma_L^2}{\varepsilon^2}\big)$;
PAGE: $\tilde{\mathcal{O}}\big(\frac{\sigma_\ell^2\sigma}{\varepsilon^3} \vee \frac{\sigma_L^2}{\varepsilon^2}\big)$; **SPIDER:** $\tilde{\mathcal{O}}\big(\frac{\sigma_\ell^2\sigma}{\varepsilon^3} \vee \frac{\sigma_L^2}{\varepsilon^2}\big)$.
Family 3: **SO Momentum:** $\tilde{\mathcal{O}}\big(\frac{\sigma^3}{\varepsilon^3} \vee \frac{\sigma_\gamma^2\sigma}{\varepsilon^3} \vee \frac{\sigma_\alpha^{3/2}\sigma}{\varepsilon^{5/2}} \vee \frac{\sigma_L^2}{\varepsilon^2}\big)$;
SO PAGE: $\tilde{\mathcal{O}}\big(\frac{\sigma_\gamma^2\sigma}{\varepsilon^3} \vee \frac{\sigma_\alpha^{3/2}\sigma}{\varepsilon^{5/2}} \vee \frac{\sigma_L^2}{\varepsilon^2}\big)$; **SO SPIDER:** $\tilde{\mathcal{O}}\big(\frac{\sigma_\gamma^2\sigma}{\varepsilon^3} \vee \frac{\sigma_\alpha^{3/2}\sigma}{\varepsilon^{5/2}} \vee \frac{\sigma_L^2}{\varepsilon^2}\big)$.

We use $\tilde{\mathcal{O}}$ to hide the dependency on $\log\frac{1}{\varepsilon}$ and $\log\frac{1}{\delta}$ in the informal Theorem 6.1 above. It is important to note that our comparisons to existing literature remain valid and fair despite differences in underlying sub-Gaussian assumptions versus bounded expected variance. For Family 1, our complexities for SGD-M match the ε^{-4} rate of existing work that also employs a normalized update vector (see Theorem 1 in [31]). It allows us to achieve a sharper subdominant σ^3/ε^3 term compared to the $(\sigma^4/\sigma_L^2)/\varepsilon^2 + (\sigma_L^2\sigma^2)/\varepsilon^4$ rate established in [31], since the latter is strictly lower-bounded by our rate via $2\sqrt{xy} \le x + y$. While previous works within Family 2 and Family 3 rely on unnormalized update vectors, our approach incorporates normalization; nevertheless, direct comparisons to their respective baselines of complexities remain mathematically fair. Specifically, the complexity rates for STORM align with the dominant σ^3/ε^3 term established in existing work (Theorem 1 of [33]), and the $(\sigma_\ell^2\sigma)/\varepsilon^3$ oracle complexity matches previous findings for SPIDER (Theorem 1 in [43]). For Family 3, which incorporates second-order correction, the oracle complexities of our new second-order variants align with the $(\sigma_\gamma^2\sigma)/\varepsilon^3 + (\sigma_\alpha^{3/2}\sigma)/\varepsilon^{5/2} + \sigma_L^2/\varepsilon^2$ rate in prior work on variance reduction via Hessian-vector products (Theorem 2, page 8 of [6]). For details and configurations of our theoretical results, please refer to Tables 5, 6, as well as the formal Theorem E.1 in Appendix E.

6.2 Application 2: Expectation-Constrained Stochastic Optimization

We now consider the expectation-constrained stochastic optimization problem (‡), in which feasibility must be determined from noisy observations of the constraint function. Unlike Application 1, where the estimator tracks a gradient used for the update direction, here the estimator tracks the scalar constraint value used for feasibility detection. This setting fits directly into the estimation framework developed in the previous sections by setting $g(w) := h(w)$, $G(w,\xi) := H(w,\xi)$, and hence $(\mathcal{G}, \|\cdot\|_{\mathcal{G}}) := (\mathbb{R}, |\cdot|)$. The unified estimator $(*)$ then produces an estimate of $h(w_t)$, whose error is controlled by Theorem 4.2. This perspective yields a general reduction from high-probability estimation guarantees to optimization guarantees for constrained stochastic problems. Throughout this subsection, we use a constant step size $\eta_t \equiv \eta > 0$.

Table 3: Dominant ε-dependence of the oracle complexity to achieve ε-stationarity (Application 1) or ε-optimality and feasibility tolerance (Application 2), with logarithmic factors in ε^{-1} and δ^{-1} suppressed. Application 1 refers to stochastic optimization with mirror descent (†); Application 2 refers to stochastic optimization with an expectation constraint (‡). Full expressions with explicit constants appear in Appendices E and F.

	Application 1 (†)			**Application 2** (‡)		
	Family 1	Family 2	Family 3	Family 1	Family 2	Family 3
Case 1: $\beta<1,\ p=0$ Case 2: $\beta=1,\ p>0$ Case 3: $\beta=1, p=\mathbb{I}_{\{t \bmod E=0\}}$	ε^{-4}	ε^{-3}	ε^{-3}	ε^{-4}	ε^{-3}	ε^{-3}

Assumption 6.2 (Constrained Setup)**.** The functions f, h are convex on $\mathcal{W}$. The domain $\mathcal{W}$ is bounded such that $\sup_{u,v\in\mathcal{W}} \|u-v\| \le D$ and $\sup_{u,v\in\mathcal{W}} D_\Phi(u,v) \le \frac{R^2}{2}$. Furthermore, the stochastic subgradients are conditionally unbiased given the filtration $\mathcal{F}_{t-1}$ and uniformly bounded:

$$\mathbb{E}[F'(w_t,\zeta_t) \mid \mathcal{F}_{t-1}] \in \partial f(w_t), \quad \mathbb{E}[H'(w_t,\zeta_t) \mid \mathcal{F}_{t-1}] \in \partial h(w_t),$$

with $\|F'(w,\zeta)\|_* \le G$ and $\|H'(w,\zeta)\|_* \le G$ almost surely for all $w \in \mathcal{W}$. Moreover, conditioned on the history and the switching decision, the stochastic subgradient used in the update remains unbiased, since ζ_t is independent of $\mathcal{F}_{t-1}, \xi_t, \xi_t^B$ and b_t. There exists an optimal solution $w^\star \in \arg\min_{w\in\mathcal{W}:\, h(w)\le 0} f(w)$ to (‡).

Switching Gradient Method (SGM). To solve (‡), we use the primal-only *Switching Gradient Method* (SGM) [111, 13, 82]. At iteration t, the estimator v_t is used to decide whether to update using the objective subgradient or the constraint subgradient. Given a switching threshold $\varepsilon > 0$, define

$$U_t := \mathbb{I}_{\{v_t\le\varepsilon\}}\, F'(w_t,\zeta_t) + \big(1-\mathbb{I}_{\{v_t\le\varepsilon\}}\big)\, H'(w_t,\zeta_t), \tag{5}$$

and perform mirror-descent $w_{t+1} = \arg\min_{w\in\mathcal{W}} \{\eta\langle U_t, w\rangle + D_\Phi(w,w_t)\}$. Thus, whenever the estimated constraint value is sufficiently small, the method prioritizes descent on the objective; otherwise, it prioritizes reducing constraint violation.

Theorem 6.3 (Constrained SGM via Unified Estimation)**.** Suppose Assumption 6.2 holds and the unified estimator (∗) satisfies the uniform tail bound $\mathbb{P}\big(\sup_{t\in[T]} |e_t| > \mathcal{E}\big) \le \delta/2$ for some deterministic $\mathcal{E} \ge 0$. If we set the switching threshold to

$$\varepsilon := \frac{R^2}{2\eta T} + \frac{\eta G^2}{2} + \frac{2DG}{\sqrt{T}}\sqrt{2\log\frac{4}{\delta}} + \mathcal{E},$$

then, with probability at least $1-\delta$, the selected set $\mathcal{A} := \{t\in[T] : v_t \le \varepsilon\}$ is non-empty, and the uniform average $\bar{w} = \frac{1}{|\mathcal{A}|}\sum_{t\in\mathcal{A}} w_t$ satisfies both $f(\bar{w}) - f(w^\star) \le \varepsilon$ and $h(\bar{w}) \le \varepsilon + \mathcal{E}$.

Theorem 6.3 provides a reusable reduction: once a family-specific estimator yields a high-probability bound $\mathcal{E}$ (see instantiations in Table 2), the optimization guarantee follows by combining this envelope with the SGM error terms from Table 2. Optimizing the parameters case by case yields the oracle complexities in Table 3; full statements are given in Corollaries F.4–F.8 in Appendix F. The rate is dictated by the estimator envelope $\mathcal{E}$. Family 1 retains a first-order estimation bias and matches the baseline $\mathcal{O}(\varepsilon^{-4}\log^4(1/\delta))$ rate of [82], whereas Families 2 and 3 remove this and improve the dominant rate to $\tilde{\mathcal{O}}(\varepsilon^{-3})$ with milder confidence dependence. In particular, for Families 2 and 3, the ε^{-3} oracle-complexity terms have at most $\log^{3/2}(1/(\varepsilon\delta))$ logarithmic confidence dependence in the fully recursive and loopless-reset settings (Cases 1 and 2), whereas the periodic-reset setting (Case 3) exhibits an even tighter logarithmic dependence. Although Family 1 does not improve the dominant rate, its fully recursive variant ($\beta_t = \beta < 1, p_t = 0$) uses batch size 1 throughout, avoiding repeated large-batch constraint evaluations.

For primal-only switching methods with a convex expectation constraint and high-probability guarantees of simultaneous optimality and feasibility, [82] obtain $\widetilde{\mathcal{O}}(\varepsilon^{-4})$ total oracle complexity using $\widetilde{\mathcal{O}}(\varepsilon^{-2})$ samples per iteration for constraint estimation. Under comparable light-tail assumptions, our

Family 1 matches this rate while using $\mathcal{O}(1)$ constraint samples per iteration. Under the additional sub-Gaussian centered-difference assumption required by Family 2, the total oracle complexity improves to $\widetilde{\mathcal{O}}(\varepsilon^{-3})$. Family 3 achieves the same rate when the population constraint is smooth and the stochastic derivative noise is sub-Gaussian. Accordingly, we qualify our claim as the first, to our knowledge, high-probability $\widetilde{\mathcal{O}}(\varepsilon^{-3})$ stochastic-oracle guarantee within this primal-only switching and convex expectation-constraint setting. This comparison does not include primal-dual, augmented-Lagrangian, or compositional methods that employ different oracle models, assumptions, or solution criteria.

7 Conclusion

In this paper, we introduced a unified framework for the design and analysis of recursive stochastic estimators. By decomposing estimator dynamics into memory retention, reset probability, and a dedicated correction term for iterate movement, we provided a lens through which various variance-reduced estimators can be analyzed with high probability. A main contribution of our work is the development of a unified high-probability estimation bound for general smooth normed spaces. By moving beyond expectation-based and Euclidean-centric analyses, this bound provides robust estimation guarantees for generalized first-order methods, including mirror descent. We demonstrated the power of this approach by recovering high-probability guarantees for unconstrained stochastic optimization and establishing the first variance-reduced oracle complexities for expectation-constrained problems, improving the dominant rate from $\tilde{\mathcal{O}}(\varepsilon^{-4})$ to $\tilde{\mathcal{O}}(\varepsilon^{-3})$.

Acknowledgements

This work was supported in part by the National Science Foundation (NSF) under Grants DMS-2502560 and CNS-2313109, and in part by an NSF CAREER Award under Grant CCF-2337808.

References

[1] Sai Aparna Aketi, Abolfazl Hashemi, and Kaushik Roy. Global update tracking: A decentralized learning algorithm for heterogeneous data. *Advances in neural information processing systems*, 36:48939–48961, 2023.

[2] Ahmet Alacaoglu, Yura Malitsky, and Stephen J Wright. Towards weaker variance assumptions for stochastic optimization. *arXiv preprint arXiv:2504.09951*, 2025.

[3] Mohammad S Alkousa. On modification of an adaptive stochastic mirror descent algorithm for convex optimization problems with functional constraints. In *Computational Mathematics and Applications*, pages 47–63. Springer, 2020.

[4] Zeyuan Allen-Zhu and Yuanzhi Li. Neon2: Finding local minima via first-order oracles. In *Advances in Neural Information Processing Systems*, pages 3716–3726, 2018.

[5] Luigi Ambrosio, Elia Brué, and Daniele Semola. *Lectures on optimal transport*. Springer, Cham, Switzerland, second edition, 2021.

[6] Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Ayush Sekhari, and Karthik Sridharan. Second-order information in non-convex stochastic optimization: Power and limitations. In *Conference on Learning Theory*, pages 242–299. PMLR, 2020.

[7] Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, pages 1–50, 2022.

[8] Hilal Asi and John C Duchi. Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM Journal on Optimization*, 29(3):2257–2290, 2019.

[9] Francis Bach and Eric Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

[10] Syreen Banabilah, Moayad Aloqaily, Eitaa Alsayed, Nida Malik, and Yaser Jararweh. Federated learning review: Fundamentals, enabling technologies, and future applications. *Information processing & management*, 59(6):103061, 2022.

[11] Viorel Barbu and Teodor Precupanu. *Convexity and optimization in Banach spaces*. Springer Science & Business Media, 2012.

[12] Heinz H Bauschke, Jonathan M Borwein, and Patrick L Combettes. Bregman monotone optimization algorithms. *SIAM Journal on control and optimization*, 42(2):596–636, 2003.

[13] Anastasia Bayandina, Pavel Dvurechensky, Alexander Gasnikov, Fedor Stonyakin, and Alexander Titov. Mirror descent and convex optimization problems with non-smooth inequality constraints. In *Large-scale and distributed optimization*, pages 181–213. Springer, 2018.

[14] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[15] Dimitri P Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.

[16] Julius R Blum. Approximation methods which converge with probability one. *The Annals of Mathematical Statistics*, pages 382–386, 1954.

[17] Digvijay Boob and Mohammad Khalafi. Optimal primal-dual algorithm with last iterate convergence guarantees for stochastic convex optimization problems. *arXiv preprint arXiv:2410.18513*, 2024.

[18] Zdravko Botev and Ad Ridder. Variance reduction. *Wiley statsRef: Statistics reference online*, pages 1–6, 2017.

[19] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale learning. *SIAM Review*, 60(2):223–311, 2018.

[20] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, UK, 2013.

[21] Haim Brézis. *Functional analysis, Sobolev spaces and partial differential equations*, volume 2. Springer, 2011.

[22] Regina Sandra Burachik and Susana Scheimberg. A proximal point method for the variational inequality problem in banach spaces. *SIAM Journal on Control and Optimization*, 39(5):1633–1649, 2000.

[23] Yair Carmon and Oliver Hinder. Making sgd parameter-free. In *Conference on Learning Theory*, pages 2360–2389. PMLR, 2022.

[24] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40:120–145, 2011.

[25] Pushpita Chatterjee, Debashis Das, and Danda B Rawat. Federated learning empowered recommendation model for financial consumer services. *IEEE Transactions on Consumer Electronics*, 70(1):2508–2516, 2023.

[26] Vishnu Pandi Chellapandi, Antesh Upadhyay, Abolfazl Hashemi, and Stanislaw H Żak. Fednmut–federated noisy model update tracking convergence analysis. *arXiv preprint arXiv:2403.13247*, 2024.

[27] Yiyue Chen, Abolfazl Hashemi, and Haris Vikalo. Communication-efficient variance-reduced decentralized stochastic optimization over time-varying directed graphs. *IEEE Transactions on Automatic Control*, 2021.

[28] Romain Cosson, Ali Jadbabaie, Anuran Makur, Amirhossein Reisizadeh, and Devavrat Shah. Gradient descent with low-rank objective functions. In *Proceedings of the 62nd IEEE Conference on Decision and Control (CDC)*, pages 3309–3314, Singapore, December 13–15 2023.

[29] Romain Cosson, Ali Jadbabaie, Anuran Makur, Amirhossein Reisizadeh, and Devavrat Shah. Low-rank gradient descent. *IEEE Open Journal of Control Systems*, 2:380–395, October 2023.

[30] Shisheng Cui and Uday V Shanbhag. On the analysis of variance-reduced and randomized projection variants of single projection schemes for monotone stochastic variational inequality problems. *Set-Valued and Variational Analysis*, 29:453–499, 2021.

[31] Ashok Cutkosky and Harsh Mehta. Momentum improves normalized sgd. In *International conference on machine learning*, pages 2260–2268. PMLR, 2020.

[32] Ashok Cutkosky and Harsh Mehta. High-probability bounds for non-convex stochastic optimization with heavy tails. *Advances in Neural Information Processing Systems*, 34:4883–4895, 2021.

[33] Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex sgd. *Advances in neural information processing systems*, 32, 2019.

[34] Rudrajit Das, Anish Acharya, Abolfazl Hashemi, Sujay Sanghavi, Inderjit S Dhillon, and Ufuk Topcu. Faster non-convex federated learning via global and local momentum. In *Uncertainty in Artificial Intelligence*, pages 496–506. PMLR, 2022.

[35] Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[36] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27, 2014.

[37] Aaron Defazio and Konstantin Mishchenko. Learning-rate-free learning by d-adaptation. In *International Conference on Machine Learning*, pages 7449–7479. PMLR, 2023.

[38] Justin Domke, Robert Gower, and Guillaume Garrigos. Provable convergence guarantees for black-box variational inference. In *Advances in Neural Information Processing Systems*, volume 36, pages 66289–66327, 2023.

[39] Yoel Drori and Ohad Shamir. The complexity of finding stationary points with stochastic gradient descent. *arXiv preprint arXiv:1910.01845*, 2019.

[40] Dmitriy Drusvyatskiy and Courtney Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178(1):503–558, 2019.

[41] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[42] Ivar Ekeland and Roger Temam. *Convex analysis and variational problems*. SIAM, 1999.

[43] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in neural information processing systems*, 31, 2018.

[44] Cong Fang, Zhouchen Lin, and Tong Zhang. Sharp analysis for nonconvex SGD escaping from saddle points. In *Conference on Learning Theory*, pages 1192–1234. PMLR, 2019.

[45] Arda Fazla, Ege C. Kaya, Antesh Upadhyay, and Abolfazl Hashemi. Lower bounds and proximally anchored SGD for non-convex minimization under unbounded variance. *arXiv preprint arXiv:2604.16620*, 2026.

[46] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM journal on optimization*, 23(4):2341–2368, 2013.

[47] Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1):267–305, 2016.

[48] EG Gladyshev. On stochastic approximation. *Theory of Probability & Its Applications*, 10(2):275–278, 1965.

[49] Eduard Gorbunov, Marina Danilova, Innokentiy Shibaev, Pavel Dvurechensky, and Alexander Gasnikov. High-probability complexity bounds for non-smooth stochastic convex optimization with heavy-tailed noise. *Journal of Optimization Theory and Applications*, 203(3):2679–2738, 2024.

[50] Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. A unified theory of SGD: Variance reduction, sampling, quantization and coordinate descent. In *International Conference on Artificial Intelligence and Statistics*, pages 680–690. PMLR, 2020.

[51] Eduard Gorbunov, Abdurakhmon Sadiev, Marina Danilova, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability convergence for composite and distributed stochastic minimization and variational inequalities with heavy-tailed noise. *arXiv preprint arXiv:2310.01860*, 2023.

[52] Benjamin Grimmer. Convergence rates for deterministic and stochastic subgradient methods without Lipschitz continuity. *SIAM Journal on Optimization*, 29(2):1350–1365, 2019.

[53] Luyao Guo, Xinli Shi, Shaofu Yang, and Jinde Cao. Disa: A dual inexact splitting algorithm for distributed convex composite optimization. *IEEE Transactions on Automatic Control*, 69(5):2995–3010, 2023.

[54] Mert Gurbuzbalaban, Umut Simsekli, and Lingjiong Zhu. The heavy-tail phenomenon in SGD. In *International Conference on Machine Learning*, pages 3964–3975. PMLR, 2021.

[55] Saad Hikmat Haji and Adnan Mohsin Abdulazeez. Comparison of optimization techniques based on gradient descent algorithm: A review. *PalArch's Journal of Archaeology of Egypt/Egyptology*, 18(4):2715–2743, 2021.

[56] Erfan Yazdandoost Hamedani and Necdet Serhat Aybat. A primal-dual algorithm with line search for general convex-concave saddle point problems. *SIAM Journal on Optimization*, 31(2):1299–1329, 2021.

[57] J. M. Hammersley and J. G. Mauldon. General principles of antithetic variates. *Mathematical Proceedings of the Cambridge Philosophical Society*, 52(3):476–481, July 1956.

[58] J. M. Hammersley and K. W. Morton. A new Monte Carlo technique: Antithetic variates. *Mathematical Proceedings of the Cambridge Philosophical Society*, 52(3):449–475, July 1956.

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

[60] Abolfazl Hashemi. A unified model for large-scale inexact fixed-point iteration: A stochastic optimization perspective. *IEEE Transactions on Automatic Control*, 70(4):2435–2449, 2024.

[61] Abolfazl Hashemi. Rampage: Randomized mid-point for debiased gradient extrapolation. *arXiv preprint arXiv:2603.22155*, 2026.

[62] Abolfazl Hashemi, Dongmin Lee, and Anuran Makur. Strong antithetic variance reduction inequalities. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, pages 1–6, Ann Arbor, MI, USA, June 22–27 2025.

[63] Abolfazl Hashemi, Dongmin Lee, and Anuran Makur. Strong antithetic variates: Theory and applications. Preprint, pp. 1–13, 2025.

[64] Ignacio Hounie, Alejandro Ribeiro, and Luiz FO Chamon. Resilient constrained learning. *Advances in Neural Information Processing Systems*, 36:71767–71798, 2023.

[65] Yankun Huang and Qihang Lin. Oracle complexity of single-loop switching subgradient methods for non-smooth weakly convex functional constrained optimization. *Advances in Neural Information Processing Systems*, 36:61327–61340, 2023.

[66] Yi Huang, Shisheng Cui, Jian Sun, and Ziyang Meng. Distributed proximal algorithms for nonsmooth optimization: Unified convergence analysis. *IEEE Transactions on Automatic Control*, 2025.

[67] Yi Huang, Ziyang Meng, Jian Sun, and Wei Ren. Distributed multiproximal algorithm for nonsmooth convex optimization with coupled inequality constraints. *IEEE Transactions on Automatic Control*, 68(12):8126–8133, 2023.

[68] Sasila Ilandarideva, Anatoli Juditsky, Guanghui Lan, and Tianjiao Li. Accelerated stochastic approximation with state-dependent noise. *arXiv preprint arXiv:2307.01497*, 2023.

[69] Rustem Islamov, Yarden As, and Ilyas Fatkhullin. Safe-ef: Error feedback for non-smooth constrained optimization. In *Forty-second International Conference on Machine Learning*, 2025.

[70] Andrew Jacobsen and Ashok Cutkosky. Unconstrained online learning with unbounded losses. In *International Conference on Machine Learning*, pages 14590–14630. PMLR, 2023.

[71] Ali Jadbabaie, Anuran Makur, and Amirhossein Reisizadeh. Adaptive low-rank gradient descent. In *Proceedings of the 62nd IEEE Conference on Decision and Control (CDC)*, pages 3315–3320, Singapore, December 13–15 2023.

[72] Ali Jadbabaie, Anuran Makur, and Devavrat Shah. Federated optimization of smooth loss functions. *IEEE Transactions on Information Theory*, 69(12):7836–7866, December 2023.

[73] Ali Jadbabaie, Anuran Makur, and Devavrat Shah. Gradient-based empirical risk minimization using local polynomial regression. *Stochastic Systems, INFORMS*, 14(4):363–402, December 2024.

[74] Prateek Jain, Sham M Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging, and model misspecification. *Journal of machine learning research*, 18(223):1–42, 2018.

[75] Zhichao Jia and Benjamin Grimmer. First-order methods for nonsmooth nonconvex functional constrained optimization with or without slater points. *SIAM Journal on Optimization*, 35(2):1300–1329, 2025.

[76] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26, 2013.

[77] Anatoli Juditsky and Arkadii S Nemirovski. Large deviations of vector-valued martingales in 2-smooth normed spaces. *arXiv preprint arXiv:0809.0813*, 2008.

[78] H. Kahn and A. W. Marshall. Methods of reducing sample size in Monte Carlo computations. *Journal of the Operations Research Society of America*, 1(5):263–278, November 1953.

[79] Ahmed Khaled and Peter Richtárik. Better theory for SGD in the nonconvex world. *Transactions on Machine Learning Research*, 2023.

[80] Ahmed Khaled, Othmane Sebbouh, Nicolas Loizou, Robert M Gower, and Peter Richtárik. Unified analysis of stochastic gradient methods for composite convex and smooth optimization. *Journal of Optimization Theory and Applications*, 199(2):499–540, 2023.

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

[82] Guanghui Lan and Zhiqiang Zhou. Algorithms for stochastic optimization with function or expectation constraints. *Computational Optimization and Applications*, 76(2):461–498, 2020.

[83] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[84] Dongmin Lee, William Lu, and Anuran Makur. On the oracle complexity of interpolation-based gradient descent. *IEEE Transactions on Automatic Control*, 71(9):5936–5951, September 2026.

[85] Zhize Li, Hongyan Bao, Xiangliang Zhang, and Peter Richtárik. Page: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *International conference on machine learning*, pages 6286–6295. PMLR, 2021.

[86] Zichong Li, Pin-Yu Chen, Sijia Liu, Songtao Lu, and Yangyang Xu. Stochastic inexact augmented lagrangian method for nonconvex expectation constrained optimization. *Computational Optimization and Applications*, 87(1):117–147, 2024.

[87] Deyi Liu, Lam M Nguyen, and Quoc Tran-Dinh. An optimal hybrid variance-reduced algorithm for stochastic composite nonconvex optimization. *arXiv preprint arXiv:2008.09055*, 2020.

[88] Wei Liu and Yangyang Xu. A single-loop spider-type stochastic subgradient method for expectation-constrained nonconvex nonsmooth optimization. *arXiv preprint arXiv:2501.19214*, 2025.

[89] Yan Liu, Mingjie Chen, Chaojie Ji, Hao Zhang, and Ruxin Wang. Serena: A unified stochastic recursive variance reduced gradient framework for riemannian non-convex optimization. In *Forty-second International Conference on Machine Learning*, 2025.

[90] Yanli Liu, Yuan Gao, and Wotao Yin. An improved analysis of stochastic gradient descent with momentum. *Advances in Neural Information Processing Systems*, 33:18261–18271, 2020.

[91] Zijian Liu, Ta Duy Nguyen, Thien Hang Nguyen, Alina Ene, and Huy Nguyen. High probability convergence of stochastic gradient methods. In *International Conference on Machine Learning*, pages 21884–21914. PMLR, 2023.

[92] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[93] Zhankun Luo and Abolfazl Hashemi. Unveiling the cycloid trajectory of em iterations in mixed linear regression. In *International Conference on Machine Learning*, pages 33329–33382. PMLR, 2024.

[94] Zhankun Luo and Abolfazl Hashemi. Structural properties, cycloid trajectories and non-asymptotic guarantees of em algorithm for mixed linear regression. *arXiv preprint arXiv:2511.04937*, 2025.

[95] Zhankun Luo and Abolfazl Hashemi. Characterizing evolution in expectation-maximization estimates for overspecified mixed linear regression. *Transactions on Machine Learning Research*, 2026.

[96] Zhankun Luo, Antesh Upadhyay, Sang Bin Moon, and Abolfazl Hashemi. First-order softmax weighted switching gradient method for distributed stochastic minimax optimization with stochastic constraints. *arXiv preprint arXiv:2603.05774*, 2026.

[97] Runchao Ma, Qihang Lin, and Tianbao Yang. Quadratically regularized subgradient methods for weakly convex optimization with weakly convex constraints. In *International Conference on Machine Learning*, pages 6554–6564. PMLR, 2020.

[98] Vien Mai and Mikael Johansson. Convergence of a stochastic gradient method with momentum for non-smooth non-convex optimization. In *International conference on machine learning*, pages 6630–6639. PMLR, 2020.

[99] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[100] Konstantin Mishchenko and Aaron Defazio. Prodigy: An expeditiously adaptive parameter-free learner. *arXiv preprint arXiv:2306.06101*, 2023.

[101] Sang Bin Moon and Abolfazl Hashemi. Optimistic regret bounds for online learning in adversarial markov decision processes. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2024.

[102] Arkadi Nemirovski. Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

[103] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[104] AS Nemirovsky, DB Yudin, and E Dawson. Wiley-interscience series in discrete mathematics, 1983.

[105] Gergely Neu and Nneka Okolo. Dealing with unbounded gradients in stochastic saddle-point optimization. In *Proceedings of the 41st International Conference on Machine Learning*, pages 37508–37530, 2024.

[106] Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *International conference on machine learning*, pages 2613–2621. PMLR, 2017.

[107] Ta Duy Nguyen, Alina Ene, and Huy L Nguyen. Improved convergence in high probability of clipped gradient methods with heavy tails. *arXiv preprint arXiv:2304.01119*, 2023.

[108] Victor M Panaretos and Yoav Zemel. Statistical aspects of Wasserstein distances. *Annual review of statistics and its application*, 6(1):405–431, March 2019.

[109] Le Peng, Gaoxiang Luo, Sicheng Zhou, Jiandong Chen, Ziyue Xu, Ju Sun, and Rui Zhang. An in-depth evaluation of federated learning on biomedical natural language processing for information extraction. *NPJ Digital Medicine*, 7(1):127, 2024.

[110] Iosif Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, 22(4):1679–1706, 1994.

[111] Boris T Polyak. A general method for solving extremal problems. In *Soviet Mathematics Doklady*, volume 8, pages 593–597, 1967.

[112] Boris T Polyak. Introduction to optimization. 1987.

[113] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1571–1578, 2012.

[114] Liang Ran, Huaqing Li, Jinhui Hu, Qingguo Lü, Zheng Wang, Zhe Li, and Guo Chen. Distributed primal-dual proximal algorithms for convex optimization involving three composite functions. *IEEE Transactions on Control of Network Systems*, 11(1):389–401, 2023.

[115] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):119, 2020.

[116] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[117] Nicolas Roux, Mark Schmidt, and Francis Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. *Advances in neural information processing systems*, 25, 2012.

[118] Reüven Y. Rubinstein and Dirk P. Kroese. *Simulation and the Monte Carlo Method*. John Wiley & Sons, Inc., Hoboken, NJ, USA, third edition, 2017.

[119] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[120] Abdurakhmon Sadiev, Marina Danilova, Eduard Gorbunov, Samuel Horváth, Gauthier Gidel, Pavel Dvurechensky, Alexander Gasnikov, and Peter Richtárik. High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. In *International conference on machine learning*, pages 29563–29648. PMLR, 2023.

[121] Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International Conference on Machine Learning*, pages 71–79. PMLR, 2013.

[122] Aleksandr Shestakov, Valery Parfenov, and Aleksandr Beznosikov. Unified theory of adaptive variance reduction. *arXiv preprint arXiv:2511.04569*, 2025.

[123] Fedor Stonyakin, Alexey Stepanov, Alexander Gasnikov, and Alexander Titov. Mirror descent for constrained optimization problems with large subgradient values. *arXiv preprint arXiv:1908.00218*, 2019.

[124] Fedor S Stonyakin, M Alkousa, Alexey N Stepanov, and Aleksandr A Titov. Adaptive mirror descent algorithms for convex and strongly convex optimization problems with functional constraints. *Journal of Applied and Industrial Mathematics*, 13(3):557–574, 2019.

[125] Fedor S Stonyakin, Mohammad S Alkousa, Alexander A Titov, and Victoria V Piskunova. On some methods for strongly convex optimization problems with one functional constraint. In *Mathematical Optimization Theory and Operations Research: 18th International Conference, MOTOR 2019, Ekaterinburg, Russia, July 8-12, 2019, Proceedings 18*, pages 82–96. Springer, 2019.

[126] FS Stonyakin, AN Stepanov, AV Gasnikov, AA Titov, et al. Mirror descent for constrained optimization problems with large subgradient values of functional constraints. *Computer research and modeling*, 12(2):301–317, 2020.

[127] Matus Telgarsky. Stochastic linear optimization never overfits with quadratically-bounded losses on general data. In *Conference on Learning Theory*, pages 5453–5488. PMLR, 2022.

[128] Alexander A Titov, Fedor S Stonyakin, Mohammad S Alkousa, Seydamet S Ablaev, and Alexander V Gasnikov. Analogues of switching subgradient schemes for relatively lipschitz-continuous convex programming problems. In *International Conference on Mathematical Optimization Theory and Operations Research*, pages 133–149. Springer, 2020.

[129] Alexander A Titov, Fedor S Stonyakin, Alexander V Gasnikov, and Mohammad S Alkousa. Mirror descent and constrained online optimization problems. In *International Conference on Optimization and Applications*, pages 64–78. Springer, 2018.

[130] Hoang Tran and Ashok Cutkosky. Better sgd using second-order momentum. *Advances in Neural Information Processing Systems*, 35:3530–3541, 2022.

[131] Quoc Tran-Dinh, Nhan H Pham, Dzung T Phan, and Lam M Nguyen. Hybrid stochastic gradient descent algorithms for stochastic nonconvex optimization. *arXiv preprint arXiv:1905.05920*, 2019.

[132] Nilesh Tripuraneni, Mitchell Stern, Chi Jin, Jeffrey Regier, and Michael I Jordan. Stochastic cubic regularization for fast nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 2899–2908, 2018.

[133] Antesh Upadhyay, Sang Bin Moon, and Abolfazl Hashemi. Fedsgm: A unified framework for constraint aware, bidirectionally compressed, multi-step federated optimization. *arXiv preprint arXiv:2601.16897*, 2026.

[134] Jianghui Wang, Yang Chen, Xingyu Xie, Cong Fang, and Zhouchen Lin. Task-robust pre-training for worst-case downstream adaptation. *Advances in Neural Information Processing Systems*, 36:9458–9478, 2023.

[135] Mengdi Wang and Dimitri P Bertsekas. Stochastic first-order methods with random constraint projection. *SIAM Journal on Optimization*, 26(1):681–717, 2016.

[136] Tianjing Wang, Zhao Yang Dong, and Houbo Xiong. Adaptive multi-personalized federated learning for state of health estimation of multiple batteries. *IEEE Internet of Things Journal*, 2024.

[137] Jie Wen, Zhixia Zhang, Yang Lan, Zhihua Cui, Jianghui Cai, and Wensheng Zhang. A survey on federated learning: challenges and applications. *International journal of machine learning and cybernetics*, 14(2):513–535, 2023.

[138] Ward Whitt. Bivariate distributions with given marginals. *The Annals of Statistics*, 4(6):1280–1289, November 1976.

[139] Yangyang Xu and Yibo Xu. Momentum-based variance-reduced proximal stochastic gradient method for composite nonconvex stochastic optimization. *Journal of Optimization Theory and Applications*, 196(1):266–297, 2023.

[140] Yi Xu, Jin Rong, and Tianbao Yang. First-order stochastic algorithms for escaping from saddle points in almost linear time. In *Advances in Neural Information Processing Systems*, pages 5530–5540, 2018.

[141] Xingyi Yang. Stochastic gradient variance reduction by solving a filtering problem. *arXiv preprint arXiv:2012.12418*, 2020.

[142] Huizhuo Yuan, Yifeng Liu, Shuang Wu, Zhou Xun, and Quanquan Gu. MARS: Unleashing the power of variance reduction for training large models. In *International Conference on Machine Learning*, pages 73553–73587. PMLR, 2025.

[143] Zhe Zhang and Guanghui Lan. Solving convex smooth function constrained optimization is almost as easy as unconstrained optimization. *arXiv preprint arXiv:2210.05807*, 2022.

Part

A Notations, Symbols and Preliminaries 20

B Proofs for Results of Unified Estimation Error Bound 26

C Proofs for Instantiations of the Unified Estimation Error Bound 36

D Lemmas in Proofs for Two Applications 44

E Application 1: Stochastic Optimization with Mirror Descent 49

F Application 2: Expectation-Constrained Stochastic Optimization 67

G Related Work 81

Appendices

We organize the Appendices as follows:

- In Appendix A, we provide the notations, symbols, and preliminaries in Section 2.
- In Appendix B, we provide proofs for the results of the unified estimation error bound in Section 4.
- In Appendix C, we present proofs for instantiations of the unified estimation error bound in Section 5.
- In Appendix D, we prepare lemmas used in the proofs for the two applications: (†) and (‡).
- In Appendix E, we detail the first application (†) regarding stochastic optimization with mirror descent in Subsection 6.1.
- In Appendix F, we detail the second application (‡) regarding expectation-constrained stochastic optimization in Subsection 6.2.
- In Appendix G, we discuss related work.

A Notations, Symbols and Preliminaries

Symbol	Meaning
$\mathcal{X}, \|\cdot\|$	Reflexive Banach space with primal norm.
$\mathcal{X}^*, \|\cdot\|_*$	Dual space and dual norm of $\mathcal{X}$.
$\Phi, \mathcal{D}$	Mirror domain for the mirror map $\Phi : \mathcal{D} \to \mathbb{R}$, with $\mathcal{D} \subseteq \mathcal{X}$.
$\mathcal{W}$	Closed convex feasible iterate set with $\mathcal{W} \subseteq \text{int}(\mathcal{D})$.
$D_\Phi(x, y)$	Bregman divergence induced by Φ.
$P(w, u, \eta)$	Proximal gradient mapping: $P(w, u, \eta) := (w - w^+)/\eta$ where w^+ minimizes $\langle \eta u, w' \rangle + D_\Phi(w', w)$ over $w' \in \mathcal{W}$ (Definition 2.2).
w_t	Optimization iterate at time t, with $w_t \in \mathcal{W}$.
U_t	Update vector used by the (generalized) update rule at time t.
η_t	Learning rate / step size at time t.
$\mathcal{G}, \|\cdot\|_\mathcal{G}$	κ-smooth space equipped with $\|\cdot\|_\mathcal{G}$, and $G : \mathcal{X} \times \Omega \to \mathcal{G}$
κ	Smoothness parameter (≥ 1) for the geometry regularity of $(\mathcal{G}, \|\cdot\|_\mathcal{G})$.
$G(w, \xi), g(w)$	Stochastic oracle with randomness ξ, and $g(w) := \mathbb{E}_\xi[G(w, \xi)]$.
ξ_t / $\xi_t^B := (\xi_t^{(i)})_{i=1}^B$	Fresh oracle randomness at time t / "batch" oracle randomness at resets.
v_t, e_t	Unified estimator tracking $g(w_t)$, estimation error: $e_t := v_t - g(w_t)$.
$L(\mathcal{X}, \mathcal{G}), \|\cdot\|_{L(\mathcal{X},\mathcal{G})}$	$L(\mathcal{X}, \mathcal{G})$ is the set of all bounded linear operators from $\mathcal{X}$ to $\mathcal{G}$, and for any $T \in L(\mathcal{X}, \mathcal{G})$, its norm is given by $\|T\|_{L(\mathcal{X},\mathcal{G})} := \sup_{\|x\|\leq 1, x\in\mathcal{X}} \|Tx\|_\mathcal{G}$
$\mathcal{F}_t$	Filtration generated by oracle noise and reset events up to time t.
b_t	Reset indicator: $b_t \sim \text{Bernoulli}(p_t)$.
p_t	Reset probability (predictable, taking values in $[0, 1]$).
β_t	Momentum parameter (in $(0, 1]$), controlling exponential forgetting.
$\mathcal{T}_t$	Correction term used by the unified estimator.
$\mathcal{I}_t(\mathcal{T})$	Effective innovation at time t induced by correction term $\mathcal{T}_t$.
Δ_t	Difference term in the unified recursion: $\Delta_t := v_{t-1} - G(w_{t-1}, \xi_t)$.
$\Delta(w_t, w_{t-1}, \xi_t)$	Centered estimate-difference noise: $\Delta(w_t, w_{t-1}, \xi_t) := (G(w_t, \xi_t) - G(w_{t-1}, \xi_t)) - (g(w_t) - g(w_{t-1}))$.
τ_m	m-th reset time, with $\tau_0 = 0$ and $\tau_m = \inf\{t > \tau_{m-1} : b_t = 1\}$.
N_t	Counting process for the number of resets up to time t: $N_t := \sum_{s=1}^t b_s$ with the exact equivalence $\{\tau_m \leq t\} = \{N_t \geq m\}$.
$m(t)$	Epoch index such that $t \in (\tau_{m(t)-1}, \tau_{m(t)}]$, equivalently $m(t) := 1 + N_{t-1} = 1 + \sum_{s=1}^{t-1} b_s$.
$\Lambda_{t,j}$	Cumulative momentum multiplier: $\Lambda_{t,j} := \prod_{i=j+1}^t \beta_i$, $\Lambda_{t,t} = 1$.
A_t	Auxiliary multiplier: $A_t := \prod_{i=1}^t \beta_i^{-1}$.
$\mathcal{B}_t$	Conditional bias magnitude of the effective innovation (Assumption 4.1).
Σ_t^2	Conditional variance proxy for the centered effective innovation (Assumption 4.1).
$\mathfrak{V}_t^2$	Accumulated predictable variance proxy over the current epoch.
V_t	Deterministic variance budget such that $\mathfrak{V}_t^2 \leq V_t$.
σ^2	Oracle-noise variance proxy parameter (Assumption 4.1).
B	Batch/reset parameter used in the reset/initialization concentration term.
T	Deterministic time horizon (w.l.o.g. with $b_T = 1$).
E	Epoch length for periodic reset schedules (e.g., $p_t = \mathbb{I}_{\{t \bmod E=0\}}$).
	Application 1: Stochastic Optimization with Mirror Descent
P_t	Proximal stationarity witness at time t: $P_t := P\left(w_t, \frac{\nabla f(w_t)}{\|\nabla f(w_t)\|_*}, \frac{\eta}{2}\right)$ with $f(w) = \mathbb{E}_\xi[F(w, \xi)]$, $g(w) = \nabla f(w) = \mathbb{E}_\xi[G(w, \xi)]$.
Δ_f	Initial objective gap: $\Delta_f := f(w_1) - \inf_{w\in\mathcal{W}} f(w) < +\infty$.
σ_L	Shorthand scale $\sigma_L := (\Delta_f L)^{1/2}$, where L is from Assumption 5.1.

Symbol	Meaning
σ_ℓ	Shorthand scale $\sigma_\ell := (\Delta_f\, \ell)^{1/2}$, where ℓ is from Assumption 5.2.
σ_γ	Shorthand scale $\sigma_\gamma := (\Delta_f\, \gamma)^{1/2}$, where γ is from Assumption 5.3.
σ_α	Shorthand scale $\sigma_\alpha := (\Delta_f^2\, \alpha)^{1/3}$, where α is from Assumption 5.4.
Application 2: Expectation-Constrained Stochastic Optimization	
$F(w,\xi),\ H(w,\xi)$	Sample Objective and Constraint functions in the constrained optimization application.
$f(w),\ h(w)$	Population objective and population constraint: $f(w) := \mathbb{E}[F(w,\xi)]$, $h(w) := \mathbb{E}[H(w,\xi)]$.
ζ_t	Fresh stochastic randomness for the stochastic sub-gradients at iteration t, which is independent of $\mathcal{F}_{t-1}, \xi_t, \xi_t^B$ and b_t.
$F'(w_t,\zeta_t),\ H'(w_t,\zeta_t)$	Stochastic sub-gradients at iteration t, unbiased gradient estimators.
$w^\star$	Any optimal feasible solution of the constrained problem.
ε	Switching threshold used to decide whether the update prioritizes the objective or the constraint.
$\mathcal{A}$	Active-phase feasible index set: $\mathcal{A} := \{t \in [T] : v_t \le \varepsilon\}$.
$\mathcal{B}$	Active-phase non-feasible index set: $\mathcal{B} := [T] \setminus \mathcal{A}$.
$\bar{w}$	Active-phase selected average: $\bar{w} := \frac{1}{\lvert\mathcal{A}\rvert} \sum_{t\in\mathcal{A}} w_t$.
$\mathcal{E}$	Deterministic tail-envelope bound satisfying $\lvert e_t\rvert \le \mathcal{E}$ uniformly for all $t \in [T]$ on a high-probability event.
$\mathsf{OptError}(\eta, T, \delta)$	Optimization Error in Theorem F.1: $\frac{R^2}{2\eta T} + \frac{\eta G^2}{2} + \frac{2DG}{\sqrt{T}}\sqrt{2\log\frac{4}{\delta}}$.

Banach space (primal norm, dual norm). The norm $\|\cdot\|$ on the reflexive Banach space $\mathcal{X}$ is the primal norm used throughout. Let $\mathcal{X}^*$ be the dual space, which is also a Banach space. For $u \in \mathcal{X}^*$, the dual norm is defined by $\|u\|_* := \sup_{\|x\|\le 1} |u(x)|$. In particular, with the definition of the canonical duality pairing $\langle u, x\rangle := u(x)$ for $u \in \mathcal{X}^*$ and $x \in \mathcal{X}$, we have

$$\|u\|_* = \sup_{\|x\|\le 1, x\in\mathcal{X}} |\langle u, x\rangle| = \sup_{\substack{x\in\mathcal{X}\\ x\neq 0}} \frac{|\langle u, x\rangle|}{\|x\|},$$

and hence $\|u\|_*\|x\| \ge \langle u, x\rangle$.

Definition A.1 (κ-Smoothness for Geometric Regularity)**.** (see [77, Def. 2.1, p. 2]) Let $\kappa \ge 1$ and $(\mathcal{G}, \|\cdot\|_\mathcal{G})$ is a normed space, then $(\mathcal{G}, \|\cdot\|_\mathcal{G})$ is κ-smooth if $p : \mathcal{G} \to \mathbb{R}, p(z) \equiv \|z\|_\mathcal{G}^2, \forall z \in \mathcal{G}$ is continuously differentiable, and for all $z, \Delta \in \mathcal{G}$,

$$p(z+\Delta) \le p(z) + \langle \nabla p(z), \Delta\rangle + \kappa p(\Delta).$$

In particular, for Hilbert spaces we have $\kappa = 1$. For Lebesgue spaces $\mathcal{G} = L_q$:

- If $q \ge 2$, then $\kappa = q - 1$.
- If $q \in (1, 2)$, then the dual exponent is $q' = \frac{q}{q-1} \in (2, \infty)$, and one typically applies κ-smoothness to the dual space $\mathcal{G}^* = (L_q)^* = L_{q'}$, giving $\kappa = q' - 1 = \frac{1}{q-1}$.
- If $q = 1$, then $\mathcal{G}^* = L_\infty$ is not κ-smooth for any finite κ in this framework.

Remark A.2 (Geometric Regularity)**.** For a κ-smooth normed space $(\mathcal{G}, \|\cdot\|)$, the κ-regularity requires the existence of a proxy function $V(\cdot)$ with a κ-Lipschitz gradient [77]. For the Lebesgue space L_q with $q \ge 2$, we have $\kappa = q - 1$. Furthermore, for Hilbert spaces (e.g., d-dimensional Euclidean space), we have $\kappa = 1$. To extend our analysis to the inherently non-smooth ℓ_∞ geometry, [2] we utilize a smooth ℓ_q approximation. Given a primal space $(\mathbb{R}^d, \|\cdot\|_p)$ with $p = 1 + \varepsilon$ and its conjugate $q = 1 + 1/\varepsilon$, the regularity constant scales as $\kappa = q - 1 = 1/\varepsilon$. Notably let $\varepsilon = \frac{1}{\ln(\max\{e,d\})}$ which yields $\kappa = \ln(\max\{e, d\})$, providing a dimension-insensitive deviation radius of order $\mathcal{O}(\sqrt{\ln d})$, while enforcing $\kappa = 1$ when $d \le 2$.

[2] This is desired as this is the natural space the gradients belong to in the most interesting application of mirror descent methods, i.e., solving problems on the probability simplex, on which the entropy is strongly convex with respect to the ℓ_1 norm.

Mirror map and Bregman divergence.

Definition A.3 (Mirror Map)**.** The mirror map $\Phi : \mathcal{D} \to \mathbb{R} \cup \{+\infty\}$ with $\mathcal{D} \subseteq \mathcal{X}$ is proper, weakly lower semi-continuous (w.l.s.c), and 1-strongly convex on a nonempty closed convex set $\mathcal{W} \subseteq \text{int}(\mathcal{D})$ with respect to $\|\cdot\|$. Furthermore, Φ is Gâteaux differentiable on $\mathcal{W}$.

Its Bregman divergence $D_\Phi(x, y)$ which is defined below satisfies the inequality for all $x, y \in \mathcal{W}$:

$$D_\Phi(x, y) \coloneqq \Phi(x) - \Phi(y) - \langle \nabla\Phi(y), x - y\rangle \geq \frac{1}{2}\|x - y\|^2.$$

Then we have the following inequality for all $x, y \in \mathcal{W}$:

$$D_\Phi(x, y) + D_\Phi(y, x) = \langle \nabla\Phi(x) - \nabla\Phi(y), x - y\rangle \geq \|x - y\|^2.$$

Definition A.4 (Proximal Gradient Mapping)**.** For $w \in \mathcal{W}$, an input vector $u \in \mathcal{X}^*$, and a step size $\eta > 0$, the **proximal gradient mapping** is $P(w, u, \eta) = (w - w^+)/\eta$, where the updated point w^+ is the unique minimizer of the proximal subproblem: $w^+ = \arg\min_{w' \in \mathcal{W}} \langle \eta u, w'\rangle + D_\Phi(w', w)$ and $D_\Phi(w', w)$ denotes the Bregman divergence induced by Φ.

Remark A.5 (Existence and Uniqueness of Proximal Gradient Mapping)**.** We guarantee the existence and uniqueness of the proximal update $P(w, u, \eta)$ by considering the minimization problem $\min_{w' \in \mathcal{W}} F(w')$. Specifically, defining the proximal objective functional as $F(w') := \langle \eta u, w'\rangle + D_\Phi(w', w)$, we proceed as follows:

1. *Coercivity and Boundedness:* The strong convexity of the mirror map Φ renders $F(w')$ coercive, ensuring that F is bounded from below and a finite infimum $\inf_{w' \in \mathcal{W}} F(w') > -\infty$ exists. Therefore, a minimizing sequence $\{w'_n\} \subset \mathcal{W}$ exists such that $\lim_{n\to\infty} F(w'_n) = \inf_{w' \in \mathcal{W}} F(w')$. Consequently, the coercivity of F indicates that the sequence $\{w'_n\} \subset \mathcal{W}$ must be bounded.

2. *Weak Subsequence (Eberlein-Šmulian):* By the Eberlein-Šmulian theorem, this bounded sequence within the reflexive Banach space $\mathcal{X}$ is guaranteed to contain a weakly convergent subsequence $w'_{n_k} \rightharpoonup w^+$.

3. *Feasibility (Mazur's Lemma):* Because the feasible set $\mathcal{W}$ is strongly closed and convex, Mazur's Lemma dictates it is also weakly closed. This ensures the weak limit point remains feasible, meaning $w^+ \in \mathcal{W}$.

4. *Existence:* Since the objective F is weakly lower semi-continuous (w.l.s.c.), it follows that $F(w^+) \leq \liminf_{k\to\infty} F(w'_{n_k}) = \lim_{n\to\infty} F(w'_n) = \inf_{w' \in \mathcal{W}} F(w')$. This proves w^+ achieves the infimum and is a global minimizer.

5. *Uniqueness:* Finally, the strict convexity of F guarantees that this global minimizer w^+ is unique.

Therefore, the proximal update $P(w, u, \eta) := (w - w^+)/\eta$ exists and is unique for any $w \in \mathcal{W} \subseteq \mathcal{X}, u \in \mathcal{X}^*$ and $\eta > 0$.

Remark A.6 (First-Order Sufficient/Necessary Optimality Condition)**.** Because the proximal objective $\mathcal{L}(w'; w, u, \eta) = \langle \eta u, w'\rangle + D_\Phi(w', w)$ is strictly convex and Gâteaux differentiable, and the domain $\mathcal{W}$ is convex, the variational inequality below is both a necessary and sufficient condition for global optimality. The necessity follows from the non-negativity of the directional derivative along feasible segments, while the sufficiency follows directly from the first-order characterization of convexity, which guarantees $\mathcal{L}(w'; w, u, \eta) \geq \mathcal{L}(w^+; w, u, \eta) + \langle \nabla\mathcal{L}(w^+; w, u, \eta), w' - w^+\rangle \geq \mathcal{L}(w^+; w, u, \eta)$ for all $w' \in \mathcal{W}$.

$$\langle \nabla\mathcal{L}(w^+; w, u, \eta),\ w' - w^+\rangle = \big\langle \eta u + \big(\nabla\Phi(w^+) - \nabla\Phi(w)\big),\ w' - w^+\big\rangle \geq 0.$$

By applying the Bregman Three-Point Identity, we have:

$$\langle \nabla\Phi(w^+) - \nabla\Phi(w), w' - w^+\rangle = D_\Phi(w', w) - D_\Phi(w', w^+) - D_\Phi(w^+, w)$$

Substituting the above into the first-order optimality condition, we obtain the equivalent condition:

$$\eta\langle u, w' - w^+\rangle + D_\Phi(w', w) - D_\Phi(w', w^+) - D_\Phi(w^+, w) \geq 0.$$

For Mirror Descent, the update rule with some update vector U_t and step size η_t is given by

$$w_{t+1} = \arg\min_{w \in \mathcal{W}} \left\{ \langle U_t, w \rangle + \frac{1}{\eta_t} D_\Phi(w, w_t) \right\} = w_t - \eta_t P(w_t, U_t, \eta_t)$$

Lemma A.7 (Properties of Proximal Gradient Mapping, see Lemmas 1-3 in [47] for first two properties)**.** For any $w \in \mathcal{W}$, input vector $u \in \mathcal{X}^*$, and step size $\eta > 0$, the proximal gradient mapping $P(w, u, \eta)$ has the following properties:

- (Inner Product Bounds) For any $u_1, u_2 \in \mathcal{X}^*$,

$$\begin{aligned} \|P(w, u_1, \eta) - P(w, u_2, \eta)\|^2 &\leq \langle u_1 - u_2, P(w, u_1, \eta) - P(w, u_2, \eta) \rangle \\ &\leq \|u_1 - u_2\|_* \|P(w, u_1, \eta) - P(w, u_2, \eta)\|. \end{aligned}$$

  In particular, $P(w, 0, \eta) = 0$ from the definition of proximal gradient mapping, therefore

$$\|P(w, u, \eta)\|^2 \leq \langle u, P(w, u, \eta) \rangle \leq \|u\|_* \|P(w, u, \eta)\|.$$

- (Non-expansiveness) For any $u_1, u_2 \in \mathcal{X}^*$,

$$\|P(w, u_1, \eta) - P(w, u_2, \eta)\| \leq \|u_1 - u_2\|_*.$$

  In particular, $P(w, 0, \eta) = 0$ from the definition of proximal gradient mapping, therefore

$$\|P(w, u, \eta)\| = \|P(w, u, \eta)\| \leq \|u\|_*.$$

- (Scaling Property) For any scalar $\alpha > 0$,

$$P(w, \alpha u, \eta) = \alpha P(w, u, \alpha \eta).$$

- (Monotonicity) For any $\eta' > \eta > 0$,

$$\langle u, \eta' P(w, u, \eta') \rangle \geq \langle u, \eta P(w, u, \eta) \rangle + \frac{\|\eta' P(w, u, \eta') - \eta P(w, u, \eta)\|^2}{\eta' - \eta}.$$

  Therefore, $\langle u, \eta' P(w, u, \eta') \rangle \geq \langle u, \eta P(w, u, \eta) \rangle$ and:

$$\|\eta' P(w, u, \eta') - \eta P(w, u, \eta)\| \leq (\eta' - \eta) \|u\|_*.$$

Proof. **Proof of the inner product bounds:**

We establish the lower bound of $\langle u_1 - u_2, P(w, u_1, \eta) - P(w, u_2, \eta) \rangle$ by applying the first-order optimality condition discussed in Remark on Definition 2.2 and using the fact $D_\Phi(x, y) + D_\Phi(y, x) \geq \|x - y\|^2$ for the 1-strongly convex mirror map Φ.

$$\eta \langle u, w' - w^+ \rangle + D_\Phi(w', w) - D_\Phi(w', w^+) - D_\Phi(w^+, w) \geq 0.$$

By introducing $w_1^+ := \arg\min_{w' \in \mathcal{W}} \langle \eta u_1, w' \rangle + D_\Phi(w', w)$, $w_2^+ := \arg\min_{w' \in \mathcal{W}} \langle \eta u_2, w' \rangle + D_\Phi(w', w)$, therefore $w_1^+ = w - \eta P(w, u_1, \eta), w_2^+ = w - \eta P(w, u_2, \eta)$, and letting $u \leftarrow u_1, w^+ \leftarrow w_1^+, w' \leftarrow w_2^+$ and $u \leftarrow u_2, w^+ \leftarrow w_2^+, w' \leftarrow w_1^+$ in the first-order optimality condition above:

$$\eta^2 \langle u_1, P(w, u_1, \eta) - P(w, u_2, \eta) \rangle \geq -D_\Phi(w_2^+, w) + D_\Phi(w_2^+, w_1^+) + D_\Phi(w_1^+, w)$$

$$\eta^2 \langle u_2, P(w, u_2, \eta) - P(w, u_1, \eta) \rangle \geq -D_\Phi(w_1^+, w) + D_\Phi(w_1^+, w_2^+) + D_\Phi(w_2^+, w)$$

Adding the above two inequalities, and applying $D_\Phi(x, y) + D_\Phi(y, x) \geq \|x - y\|^2$ for all $x, y \in \mathcal{W}$ and 1-strongly convex mirror map Φ, we obtain:

$$\eta^2 \langle u_1 - u_2, P(w, u_1, \eta) - P(w, u_2, \eta) \rangle \geq \|w_1^+ - w_2^+\|^2 = \eta^2 \|P(w, u_1, \eta) - P(w, u_2, \eta)\|^2.$$

Therefore, since $\eta > 0$, we establish the lower bound of $\langle u_1 - u_2, P(w, u_1, \eta) - P(w, u_2, \eta) \rangle$:

$$\langle u_1 - u_2, P(w, u_1, \eta) - P(w, u_2, \eta) \rangle \geq \|P(w, u_1, \eta) - P(w, u_2, \eta)\|^2.$$

From the definition of $\|\cdot\|_*$, we have $\langle u, x \rangle \leq \|u\|_* \|x\|$ for all $x \in \mathcal{X}, u \in \mathcal{X}^*$. Therefore, we establish the upper bound of $\langle u_1 - u_2, P(w, u_1, \eta) - P(w, u_2, \eta) \rangle$:

$$\langle u_1 - u_2, P(w, u_1, \eta) - P(w, u_2, \eta) \rangle \leq \|u_1 - u_2\|_* \|P(w, u_1, \eta) - P(w, u_2, \eta)\|.$$

Proof of the non-expansiveness:

Combining the lower and upper bounds, and noting that $x^2 \leq xy$ for some $x, y \geq 0$ implies $x \leq y$, we obtain the desired non-expansiveness of the proximal gradient mapping:

$$\|P(w, u_1, \eta) - P(w, u_2, \eta)\| \leq \|u_1 - u_2\|_*.$$

Proof of the scaling property:

The scaling property of the proximal gradient mapping follows from its definition, and noting that $\eta \cdot \alpha u = \alpha\eta \cdot u$.

Proof of the monotonicity:

For any $\eta' > \eta \geq 0$, $u \in \mathcal{X}^*, w \in \mathcal{W} \subseteq \mathcal{X}$, we introduce $w^+ := \arg\min_{w' \in \mathcal{W}} \langle \eta u, w' \rangle + D_\Phi(w', w)$, $w'^+ := \arg\min_{w' \in \mathcal{W}} \langle \eta' u, w' \rangle + D_\Phi(w', w)$, therefore $w^+ = w - \eta P(w, u, \eta)$, $w'^+ = w - \eta' P(w, u, \eta')$, and letting $w' \leftarrow w'^+$ and $\eta \leftarrow \eta', w^+ \leftarrow w'^+, w' \leftarrow w^+$ in the first-order optimality condition:

$$\eta \langle u, \eta P(w, u, \eta) - \eta' P(w, u, \eta') \rangle \geq -D_\Phi(w'^+, w) + D_\Phi(w'^+, w^+) + D_\Phi(w^+, w)$$

$$\eta' \langle u, \eta' P(w, u, \eta') - \eta P(w, u, \eta) \rangle \geq -D_\Phi(w^+, w) + D_\Phi(w^+, w'^+) + D_\Phi(w'^+, w)$$

Adding the above two inequalities and dividing by $\eta' - \eta > 0$, then applying $D_\Phi(x, y) + D_\Phi(y, x) \geq \|x - y\|^2$ for all $x, y \in \mathcal{W}$ and 1-strongly convex mirror map Φ, we obtain:

$$\langle u, \eta' P(w, u, \eta') - \eta P(w, u, \eta) \rangle \geq \frac{D_\Phi(w'^+, w^+) + D_\Phi(w^+, w'^+)}{\eta' - \eta} \geq \frac{\|\eta' P(w, u, \eta') - \eta P(w, u, \eta)\|^2}{\eta' - \eta}$$

By $\langle u, \eta' P(w, u, \eta') - \eta P(w, u, \eta) \rangle \leq \|u\|_* \|\eta' P(w, u, \eta') - \eta P(w, u, \eta)\|$ from the definition of the dual norm $\|\cdot\|_*$, we have:

$$\|\eta' P(w, u, \eta') - \eta P(w, u, \eta)\| \leq (\eta' - \eta)\|u\|_*.$$

□

Unified estimator v_t and estimation error e_t. For a target $g(w)$ and a stochastic oracle $G(w, \xi)$ with

$$\mathbb{E}_\xi[G(w, \xi)] = g(w)$$

which is an unbiased estimator of the target function $g(w)$. We introduce a unified estimator v_t to track the target function value $g(w_t)$ at the t-th iteration, by using $G(w, \xi)$ and the unbiased estimator of gradient $\nabla G(w, \xi)$ such that $\mathbb{E}[\nabla G(w, \xi)] = \nabla g(w)$ to reduce the variance of estimation error.

At each iteration t, let the reset indicator be $b_t \sim \text{Bernoulli}(p_t)$, where $p_t \in [0, 1]$ is the reset probability (possibly adaptive and predictable; see Assumption 3.1). When $b_t = 1$, the estimator uses an independent "batch" oracle randomness $\xi_t^B := (\xi_t^{(i)})_{i=1}^B$ from the oracle, then reset the estimator v_t to the average of evaluated function values $G(w_t, \xi_t^{(i)}), i = 1, \ldots, B$, with B samples $\xi_t^B := (\xi_t^{(i)})_{i=1}^B$. When $b_t = 0$, the estimator uses a single independent sample ξ_t drawn from the oracle, to evaluate function value $G(w, \xi)$ or gradient value $\nabla G(w, \xi)$ at w_t, w_{t-1} with the same sample ξ_t, and then update the estimator v_t by using the evaluated function/gradient value and the previous estimator v_{t-1} to track the target function value $g(w_t)$.

The unified estimator is specified by momentum parameters $\{\beta_t\}_{t \geq 1}$ with $\beta_t \in (0, 1]$ and by a correction term Δ_t and a correction term $\mathcal{T}_t$ as follows:

$$v_t := \begin{cases} G(w_t, \xi_t^B), & b_t = 1, \\ G(w_t, \xi_t) + \beta_t \Delta_t + \mathcal{T}_t, & b_t = 0, \end{cases} \tag{6}$$

where $\Delta_t := v_{t-1} - G(w_{t-1}, \xi_t)$, $\mathcal{T}_t := \mathcal{T}_t(w_t, w_{t-1}, \xi_t)$ depends on the evaluated function value $G(w, \xi)$ and estimated gradient value $\nabla G(w, \xi)$ at w_t, w_{t-1} with the same sample ξ_t, and $G(w_t, \xi_t^B) := \frac{1}{B} \sum_{i=1}^B G(w_t, \xi_t^{(i)})$ is the average of evaluated function values $G(w_t, \xi_t^{(i)}), i = 1, \ldots, B$. At the first iteration $t = 1$, we set $\Delta_1 = 0, \mathcal{T}_1 = 0$ by definition.

The tracking error e_t between the unified estimator v_t and the target function value $g(w_t)$ is given by

$$e_t := v_t - g(w_t).$$

Filtration, resets/epochs, and cumulative momentum multipliers. For high-probability analysis, let $\{\mathcal{F}_t\}_{t\geq 0}$ be the filtration generated by all oracle noise and reset events up to time t.

$$\mathcal{F}_0 = \Omega, \quad \mathcal{F}_t = \begin{cases} \sigma(\mathcal{F}_{t-1}, b_t, \xi_t\big|_{b_t=0}, \xi_t^B\big|_{b_t=1}) & \text{for Application 1 (†)}, \\ \sigma(\mathcal{F}_{t-1}, b_t, \xi_t\big|_{b_t=0}, \xi_t^B\big|_{b_t=1}, \zeta_t) & \text{for Application 2 (‡)}. \end{cases}.$$

For the Application 2, ζ_t is independent of $\mathcal{F}_{t-1}, b_t, \xi_t^B\ |_{b_t=1}, \xi_t\ |_{b_t=0}$.

Let the random sequence of epoch boundaries $\{\tau_m\}_{m\geq 0}$ for m-th reset time be defined by

$$\tau_0 = 0, \qquad \tau_m = \inf\{t > \tau_{m-1} : b_t = 1\} \quad (1 \leq m \leq T),$$

where T is a deterministic horizon. For each $t \in \{1, \ldots, T\}$, let $m(t)$ be the unique index such that $t \in (\tau_{m(t)-1}, \tau_{m(t)}]$.

For the momentum parameters $\{\beta_t\}_{t\geq 1}$ with $\beta_t \in (0, 1]$, we define the cumulative momentum multipliers for $j = 0, \ldots, t-1$ as follows:

$$\Lambda_{t,j} := \prod_{i=j+1}^{t} \beta_i, \qquad \Lambda_{t,t} = 1.$$

We also introduce $A_t := \prod_{i=1}^{t} \beta_i^{-1}$ and its inverse $A_t^{-1} = \prod_{i=1}^{t} \beta_i$ to simplify the notations.

Stopping time τ_m

To verify that the recursively defined reset sequence τ_m is a valid **stopping time**, we must show that $\{\tau_m \leq t\} \in \mathcal{F}_t$ for all t. We can prove this by introducing a counting process $N_t := \sum_{s=1}^{t} b_s$, which tracks the total number of resets that have occurred up to time t. Because the filtration is increasing ($\mathcal{F}_s \subseteq \mathcal{F}_t$ for $s \leq t$) and each b_s is $\mathcal{F}_s$-measurable, the sum $N_t \equiv \sum_{s=1}^{t} b_s$ is $\mathcal{F}_t$-measurable.

By our definition of τ_m, the m-th reset occurs at or before time t **if and only if** there have been at least m resets by time t. This yields the exact equivalence:

$$\{\tau_m \leq t\} = \{N_t \geq m\}$$

Because N_t is $\mathcal{F}_t$-measurable, the event $\{N_t \geq m\}$ belongs to the σ-algebra $\mathcal{F}_t$. Therefore, $\{\tau_m \leq t\} \in \mathcal{F}_t$, confirming that τ_m is a valid stopping time.

Epoch index $m(t)$

Furthermore, this counting process allows us to explicitly define the epoch index $m(t)$ for any time $t \geq 1$ purely in terms of the sequence b_t. Because N_{t-1} represents the exact number of resets that have occurred strictly prior to time t, we define:

$$m(t) := N_{t-1} + 1 = 1 + \sum_{s=1}^{t-1} b_s$$

Under this definition, the most recent reset prior to t occurred at $\tau_{m(t)-1}$, meaning $\tau_{m(t)-1} < t$. The current epoch's reset will occur at $\tau_{m(t)}$, meaning $t \leq \tau_{m(t)}$. This explicitly satisfies the necessary interval condition $t \in (\tau_{m(t)-1}, \tau_{m(t)}]$. Because $m(t)$ depends only on the reset variables b_s up to time $t-1$, it is $\mathcal{F}_{t-1}$-measurable, rendering $m(t)$ a predictable process.

Oracle noise variance proxy. We also use the oracle noise variance proxy $\sigma^2 > 0$ and a (positive) reset/batch parameter B in ξ_t^B and ξ_t, see also Assumption 4.1. We also introduce the conditional bias magnitude $\mathcal{B}_t$ and variance proxy Σ_t^2 for the effective innovation $\mathcal{I}_t(\mathcal{T})$ (see Definition 3.3) as shown in Assumption 4.1 to simplify the notations:

$$\mathcal{B}_t \geq \left\|\mathbb{E}[\mathcal{I}_t(\mathcal{T}) \mid \mathcal{F}_{t-1}]\right\|_{\mathcal{G}}, \quad \mathbb{E}\left[\exp\left(\frac{\|\mathcal{I}_t(\mathcal{T}) - \mathbb{E}[\mathcal{I}_t(\mathcal{T}) \mid \mathcal{F}_{t-1}]\|_{\mathcal{G}}^2}{\Sigma_t^2}\right)\middle| \mathcal{F}_{t-1}\right] \leq 2.$$

For the accumulated predictable variance proxy on the current epoch, we introduce the notation:

$$\mathfrak{V}_t^2 := \sum_{j=\tau_{m(t)-1}+1}^{t} A_j^2 \Sigma_j^2,$$

which is bounded by a deterministic variance budget $V_t \geq 0$ such that $\mathfrak{V}_t^2 \leq V_t$ as shown in Theorem B.2 and Main Theorem 4.2.

B Proofs for Results of Unified Estimation Error Bound

B.1 Lemma of Masked Martingale Difference Sequence

Lemma B.1 (Masked Martingale Difference Sequence)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space equipped with a filtration $(\mathcal{F}_n)_{n\geq 1}$. Let $(X_n)_{n\geq 1}$ be a martingale difference sequence adapted to $(\mathcal{F}_n)$, satisfying $\mathbb{E}[|X_n|] < \infty$ and $\mathbb{E}[X_n \mid \mathcal{F}_{n-1}] = 0$ almost surely. Let S and T be two stopping times with respect to $(\mathcal{F}_n)$ such that $S \leq T$ almost surely. Define the masked sequence $(Y_n)_{n\geq 1}$ as

$$Y_n \coloneqq X_n \cdot \mathbb{I}_{\{S<n\leq T\}}. \tag{7}$$

Then, $(Y_n)_{n\geq 1}$ is a martingale difference sequence with respect to $(\mathcal{F}_n)$. Consequently, for any $N \geq 1$, the sum over the random interval corresponds to the sum of the masked sequence

$$\sum_{n=S+1}^{T} X_n = \sum_{n=1}^{\infty} Y_n. \tag{8}$$

Proof. To establish that $(Y_n)_{n\geq 1}$ constitutes a martingale difference sequence, we must demonstrate that Y_n is integrable, adapted to $\mathcal{F}_n$, and satisfies the zero conditional mean property. We begin by examining the indicator process $H_n \coloneqq \mathbb{I}_{\{S<n\leq T\}}$. The event $\{S < n \leq T\}$ can be decomposed as the intersection $\{S \leq n-1\} \cap \{T \leq n-1\}^c$. By the definition of a stopping time, the event $\{S \leq n-1\}$ is $\mathcal{F}_{n-1}$-measurable. Similarly, $\{T \leq n-1\} \in \mathcal{F}_{n-1}$, which implies its complement is also in $\mathcal{F}_{n-1}$. Since $\mathcal{F}_{n-1}$ is a σ-algebra, the intersection of these events is $\mathcal{F}_{n-1}$-measurable. This establishes that the process $(H_n)_{n\geq 1}$ is predictable. Since $\mathcal{F}_{n-1} \subseteq \mathcal{F}_n$, the variable H_n is also $\mathcal{F}_n$-measurable. As X_n is adapted to $\mathcal{F}_n$, the product $Y_n = X_n H_n$ is $\mathcal{F}_n$-measurable and integrable, satisfying $\mathbb{E}[|Y_n|] \leq \mathbb{E}[|X_n|] < \infty$. We proceed to compute the conditional expectation of Y_n given the filtration $\mathcal{F}_{n-1}$

$$\mathbb{E}[Y_n \mid \mathcal{F}_{n-1}] = \mathbb{E}[X_n \cdot H_n \mid \mathcal{F}_{n-1}]. \tag{9}$$

Because H_n is $\mathcal{F}_{n-1}$-measurable, it acts as a constant with respect to the conditional expectation and can be factored out. This yields

$$\mathbb{E}[Y_n \mid \mathcal{F}_{n-1}] = H_n \cdot \mathbb{E}[X_n \mid \mathcal{F}_{n-1}]. \tag{10}$$

By hypothesis, (X_n) is a martingale difference sequence, so $\mathbb{E}[X_n \mid \mathcal{F}_{n-1}] = 0$ almost surely. Therefore, $\mathbb{E}[Y_n \mid \mathcal{F}_{n-1}] = 0$. This confirms that (Y_n) is a martingale difference sequence. The equality of the summations follows directly from the definition of the indicator function, which takes the value 1 when the index n falls between the stopping times S and T, and 0 otherwise. □

B.2 Theorem of Vector-Valued Freedman Inequality

Theorem B.2 (Vector-Valued Freedman Inequality)**.** Let $(\mathcal{G}, \|\cdot\|_{\mathcal{G}})$ be a κ-smooth normed space. Let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$ be a filtered probability space, and let $\{Y_t\}_{t\geq 1}$ be an $\mathcal{G}$-valued martingale difference sequence adapted to the filtration $\{\mathcal{F}_t\}_{t\geq 1}$. Assume there exists a sequence of non-negative predictable random variables $\{\Sigma_t\}_{t\geq 1}$, where each Σ_t is $\mathcal{F}_{t-1}$-measurable. Assume that on the $\mathcal{F}_{t-1}$-measurable event $\{\Sigma_t = 0\}$, $Y_t = 0$ almost surely.[a] Assume that the conditional light-tail bound evaluates to

$$\mathbb{E}\left[\exp\left(\frac{\|Y_t\|_{\mathcal{G}}^2}{\Sigma_t^2}\right)\middle|\mathcal{F}_{t-1}\right] \leq 2, \tag{11}$$

almost surely for all $t \geq 1$. Define the accumulated predictable variance proxy as $\mathfrak{V}_t^2 = \sum_{j=1}^t \Sigma_j^2$, and let $M_t = \sum_{j=1}^t Y_j$ denote the martingale. For any deterministic variance budget $V > 0$ and any deviation parameter $\gamma \geq 0$, the following high-probability bound holds for all $t \geq 1$:

$$\mathbb{P}\left(\|M_t\|_{\mathcal{G}} \geq \left(\sqrt{\kappa} + \gamma\right)\sqrt{V} \quad \text{and} \quad \mathfrak{V}_t^2 \leq V\right) \leq \exp\left(-\frac{\gamma^2}{3}\right). \tag{12}$$

Hence, for any deterministic variance budget $V > 0$ and any target confidence level $\delta \in (0, 1)$, the following concentration bound holds for all $t \geq 1$:

$$\mathbb{P}\left(\|M_t\|_{\mathcal{G}} \leq C(\delta, \kappa)\sqrt{V}\right) \geq 1 - \delta - \mathbb{P}\left(\mathfrak{V}_t^2 > V\right), \tag{13}$$

where $C(\delta, \kappa) := \sqrt{\kappa} + \sqrt{3 \ln \frac{1}{\delta}}$.

[a] We adopt the analytic convention that $0/0 = 0$; such a requirement is essential for our ensuing assumptions on sub-Gaussian notions of Lipschitzness where the variance proxy includes terms such as $\|w - w'\|$ in its denominator.

Geometric Implications. Let us first remark. Note that the $(2, D)$-smoothness of Pinelis [110] and the κ-regularity of Juditsky and Nemirovski [77] are connected via $\kappa = D^2$ (see the definition of κ-smoothness in Appendix A). Specifically, for a normed space $(\mathcal{G}, \|\cdot\|_{\mathcal{G}})$, [77] requires the existence of a proxy function $V(\cdot)$ with a κ-Lipschitz gradient, whereas [110] defines smoothness through the inequality $\|x + y\|_{\mathcal{G}}^2 + \|x - y\|_{\mathcal{G}}^2 \leq 2\|x\|_{\mathcal{G}}^2 + 2D^2\|y\|_{\mathcal{G}}^2$. For the Lebesgue space L_q with $q \geq 2$, both frameworks converge on the optimal regularity parameter $\kappa = q - 1$. Furthermore, for Hilbert spaces (e.g., d-dimensional Euclidean space), we have $\kappa = D = 1$. Further discussion is deferred to the Appendix A.

Proof. Following standard concentration techniques for martingales with predictable variance sequences, e.g. [110], we define the stopping time τ with respect to the filtration $\{\mathcal{F}_t\}_{t\geq 0}$ as

$$\tau = \inf\left\{k \geq 0 : \sum_{j=1}^{k+1} \Sigma_j^2 > V\right\}, \tag{14}$$

where $\inf \emptyset = \infty$ (indicating the ideal case). The equivalence $\{\tau \leq t - 1\} = \left\{\sum_{j=1}^t \Sigma_j^2 > V\right\}$ guarantees the event is $\mathcal{F}_{t-1}$-measurable, ensuring τ is a valid stopping time. We subsequently analyze the stopped martingale sequence $M_{t\wedge\tau}$, where $t \wedge \tau$ is a shorthand for $\min(t, \tau)$.

We now follow the approach of [77] to construct a scalar supermartingale utilizing a proxy function mapping derived from the dual geometry of $\mathcal{G}$. Let $\|\cdot\|_{\mathcal{G}^*}$ denote the dual norm. We define the base proxy function $V : \mathcal{G} \to \mathbb{R}$ via the Legendre transform

$$V(z) = \sup_{\|\eta\|_{\mathcal{G}^*} \leq 1}\left[\langle \eta, z\rangle - \frac{\kappa}{2}\|\eta\|_{\mathcal{G}^*}^2\right]. \tag{15}$$

Because the space $\mathcal{G}$ is κ-smooth, the dual penalty $v(\eta) = \frac{\kappa}{2}\|\eta\|_{\mathcal{G}^*}^2$ is globally 1-strongly convex (Page 22 of [77]). Standard convex analysis (see our detailed standard analysis for Equations 15,16,17,18 in Remark B.3 and Page 22 of [77]) establishes that $V(\cdot)$ is continuously differentiable, its gradient satisfies $\|V'(z)\|_{\mathcal{G}^*} \leq 1$, and V' is globally 1-Lipschitz continuous, ensuring

$$V(z + \Delta) \leq V(z) + \langle V'(z), \Delta\rangle + \frac{1}{2}\|\Delta\|_{\mathcal{G}}^2. \tag{16}$$

Further, the primal geometric offset is bounded by

$$\|z\|_{\mathcal{G}} = \sup_{\|\eta\|_{\mathcal{G}^*} \leq 1} \langle \eta, z \rangle \leq \frac{\kappa}{2} + V(z). \tag{17}$$

Following the proof of Theorem 4.1 in [77], for any fixed scalar $\beta > 0$ (to be optimized later), the scaled proxy function $V_\beta(z) = \beta V(z/\beta)$ preserves the gradient bound and modifies the spatial constraints to

$$\|z\|_{\mathcal{G}} \leq \frac{\kappa\beta}{2} + V_\beta(z) \quad \text{and} \quad V_\beta(z+\Delta) \leq V_\beta(z) + \langle V'_\beta(z), \Delta \rangle + \frac{1}{2\beta}\|\Delta\|^2_{\mathcal{G}}. \tag{18}$$

By defining the scalar increments $\psi_i = V_\beta(M_i) - V_\beta(M_{i-1})$, the condition $\|V'_\beta\|_{\mathcal{G}^*} \leq 1$ (i.e. V_β having a Lipschitz constant of 1) restricts the sequence to $|\psi_i| \leq \|Y_i\|_{\mathcal{G}}$. The Lipschitz continuity of V'_β with constant $(1/\beta)$ and $M_i = M_{i-1} + Y_i$ yields

$$\psi_i = V_\beta(M_i) - V_\beta(M_{i-1}) \leq \langle V'_\beta(M_{i-1}), Y_i \rangle + \frac{1}{2\beta}\|Y_i\|^2_{\mathcal{G}}, \tag{19}$$

such that after taking the expectation and noting Y_i is a martingale difference

$$\mathbb{E}[\psi_i \mid \mathcal{F}_{i-1}] \leq \frac{1}{2\beta}\mathbb{E}[\|Y_i\|^2_{\mathcal{G}} \mid \mathcal{F}_{i-1}]. \tag{20}$$

Applying the inequality $1 + x \leq \exp(x)$ at $x = \|Y_i\|^2_{\mathcal{G}}/\Sigma_i^2$ provides $\mathbb{E}[\|Y_i\|^2_{\mathcal{G}} \mid \mathcal{F}_{i-1}] \leq \Sigma_i^2$. Consequently, the predictable drift is bounded by

$$\mathbb{E}[\psi_i \mid \mathcal{F}_{i-1}] \leq \mu_i \coloneqq \frac{1}{2\beta}\Sigma_i^2. \tag{21}$$

To bound the exponential moments, the condition $|\psi_i| \leq \|Y_i\|_{\mathcal{G}}$ and Eq. (11) enforce $\mathbb{E}[\exp(\psi_i^2/\Sigma_i^2) \mid \mathcal{F}_{i-1}] \leq 2$. Using the inequality $\exp(s) \leq s + \exp(9s^2/16)$ which holds for any $s \in \mathbb{R}$ and Eq. (21), we evaluate the moment generating function for any $\lambda \geq 0$

$$\mathbb{E}[\exp(\lambda\psi_i) \mid \mathcal{F}_{i-1}] \leq \lambda\mu_i + \mathbb{E}\left[\exp\left(\frac{9\lambda^2\psi_i^2}{16}\right)\middle|\mathcal{F}_{i-1}\right]. \tag{22}$$

As standard in deriving Bernstein-type inequalities or uniform sub-Gaussian bounds, we need to handle two cases to separately deal with convex and concave cases. We first consider the latter, which corresponds to $0 \leq \lambda \leq \frac{4}{3\Sigma_i}$. Defining $\theta = \frac{9\lambda^2\Sigma_i^2}{16} \in [0,1]$ and applying Jensen's inequality to the concave function $z \mapsto z^\theta$ yields $\mathbb{E}[\exp(\psi_i^2/\Sigma_i^2)^\theta \mid \mathcal{F}_{i-1}] \leq \exp(\theta)$. Because $\lambda\mu_i$ and $9\lambda^2\Sigma_i^2/16$ are non-negative, $x + \exp(y) \leq \exp(x+y)$ establishes the majorant $\exp(\lambda\mu_i + \frac{3}{4}\lambda^2\Sigma_i^2)$.

For $\lambda > \frac{4}{3\Sigma_i}$ (the convex case), applying Young's inequality $\lambda x \leq \frac{3}{8}\lambda^2\Sigma_i^2 + \frac{2}{3}\frac{x^2}{\Sigma_i^2}$ at $x = \psi_i$ gives

$$\mathbb{E}[\exp(\lambda\psi_i) \mid \mathcal{F}_{i-1}] \leq \exp\left(\frac{3}{8}\lambda^2\Sigma_i^2\right)\mathbb{E}\left[\exp\left(\frac{2}{3}\frac{\psi_i^2}{\Sigma_i^2}\right)\middle|\mathcal{F}_{i-1}\right]. \tag{23}$$

Now we can apply Jensen's inequality for the exponent $2/3 \in [0,1]$ to bound the inner expectation by $\exp(2/3)$. The domain constraint necessitates $\frac{3}{8}\lambda^2\Sigma_i^2 > \frac{2}{3}$, ensuring $\exp(2/3) \leq \exp(\frac{3}{8}\lambda^2\Sigma_i^2)$. Thus, the uniform bound

$$\mathbb{E}[\exp(\lambda\psi_i) \mid \mathcal{F}_{i-1}] \leq \exp(\lambda\mu_i + \frac{3}{4}\lambda^2\Sigma_i^2) \tag{24}$$

holds globally for all λ. Furthermore, integrating the stopping time, we define the masked variables

$$\tilde{\psi}_i = \psi_i \mathbb{I}_{\{i \leq \tau\}}, \quad \tilde{\mu}_i = \mu_i \mathbb{I}_{\{i \leq \tau\}}, \quad \tilde{\Sigma}_i^2 = \Sigma_i^2 \mathbb{I}_{\{i \leq \tau\}}. \tag{25}$$

Since $\{i \leq \tau\}$ is $\mathcal{F}_{i-1}$-measurable, the conditional expectation bound in Eq. (24) is preserved , i.e.[3]

$$\mathbb{E}[\exp(\lambda\tilde{\psi}_t) \mid \mathcal{F}_{t-1}] \leq \exp(\lambda\tilde{\mu}_t + \frac{3}{4}\lambda^2\tilde{\Sigma}_t^2). \tag{26}$$

[3]To see this, note that on event $\{t \leq \tau\}$ the indicator evaluates to $I_t = 1$ and (26) reduces to (24) while on event $\{t > \tau\}$ both sides equate to 1 as the indicator evaluates to $I_t = 0$.

To construct the unconditional bound, we define the variance-penalized process

$$Z_t = \exp\left(\lambda \sum_{i=1}^{t} \tilde{\psi}_i - \lambda \sum_{i=1}^{t} \tilde{\mu}_i - \frac{3}{4}\lambda^2 \sum_{i=1}^{t} \tilde{\Sigma}_i^2\right). \tag{27}$$

To verify that Z_t is a supermartingale adapted to $\mathcal{F}_t$, we isolate the terminal increment following an approach similar to Lemma 20.2 in [83], i.e.

$$Z_t = Z_{t-1} \exp\left(\lambda\tilde{\psi}_t - \lambda\tilde{\mu}_t - \frac{3}{4}\lambda^2\tilde{\Sigma}_t^2\right). \tag{28}$$

Because τ is a predictable stopping time, $\tilde{\mu}_t$ and $\tilde{\Sigma}_t^2$ are $\mathcal{F}_{t-1}$-measurable. Factoring these out of the conditional expectation yields

$$\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] = Z_{t-1} \exp\left(-\lambda\tilde{\mu}_t - \frac{3}{4}\lambda^2\tilde{\Sigma}_t^2\right) \mathbb{E}\left[\exp(\lambda\tilde{\psi}_t)\Big|\mathcal{F}_{t-1}\right]. \tag{29}$$

Substituting the bound from Eq. (26) cancels the variance penalty terms, establishing $\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] \leq Z_{t-1}$. Thus, by tower expectation, $\mathbb{E}[\mathbb{E}[Z_k \mid \mathcal{F}_{k-1}]] \leq \mathbb{E}[Z_{k-1}]$, and doing so iteratively yields $\mathbb{E}[Z_t] \leq \mathbb{E}[Z_0] = 1$.

By definitions in Eq. (14) and Eq. (25), the cumulative constraints $\sum_{i=1}^{t} \tilde{\Sigma}_i^2 \leq V$ and $\sum_{i=1}^{t} \tilde{\mu}_i \leq \frac{1}{2\beta}V$ hold deterministically. On the event $\sum_{i=1}^{t} \tilde{\psi}_i \geq \frac{1}{2\beta}V + r$, the supermartingale ensures $Z_t \geq \exp(\lambda r - \frac{3}{4}\lambda^2 V)$. Markov's inequality yields

$$\mathbb{P}\left(\sum_{i=1}^{t} \tilde{\psi}_i \geq \frac{1}{2\beta}V + r\right) \leq \exp\left(-\lambda r + \frac{3}{4}\lambda^2 V\right). \tag{30}$$

Minimizing the quadratic exponent at $\lambda = \frac{2r}{3V}$ produces the optimal limit $\exp(-\frac{r^2}{3V})$.

On the intersection event where $\mathfrak{V}_t^2 \leq V$, we have $\tau \geq t$, and thus $\sum_{i=1}^{t} \tilde{\psi}_i = \sum_{i=1}^{t} \psi_i$.

Thus, using the definition $\psi_i = V_\beta(M_i) - V_\beta(M_{i-1})$ and a telescoping sum we have $V_\beta(M_t) \equiv \sum_{i=1}^{t} \tilde{\psi}_i$. Utilizing the first part of (18), the norm satisfies

$$\|M_t\|_{\mathcal{G}} \leq \frac{\kappa\beta}{2} + \sum_{i=1}^{t} \tilde{\psi}_i \leq \frac{\kappa\beta}{2} + \frac{1}{2\beta}V + r. \tag{31}$$

Optimizing with respect to β via differentiation yields the unique minimum $\beta = \sqrt{\frac{V}{\kappa}}$. Substituting this parameter establishes

$$\frac{\kappa}{2}\sqrt{\frac{V}{\kappa}} + \frac{1}{2}\sqrt{\kappa V} = \sqrt{\kappa V}. \tag{32}$$

Parameterizing the deviation in Eq. (30) as $r = \gamma\sqrt{V}$, the deviation simplifies to $(\sqrt{\kappa} + \gamma)\sqrt{V}$, and the corresponding tail probability evaluates to $\exp(-\gamma^2/3)$, thereby furnishing the proof. □

Remark B.3 (Detailed Standard Convex Analysis)**.** Below, we provide the detailed, step-by-step derivation of Equations 15,16,17,18 for the proof of Theorem B.2 (Vector-Valued Freedman Inequality) in Appendix B.2. This demonstrates that the result follows purely from standard convex analysis without any additional regularity conditions.

For a κ-smooth normed space $(\mathcal{G}, \|\cdot\|_{\mathcal{G}})$, its dual space $(\mathcal{G}^*, \|\cdot\|_{\mathcal{G}^*})$, and their canonical duality pairing $\langle\cdot,\cdot\rangle : \mathcal{G}^* \times \mathcal{G} \to \mathbb{R}$, it is well known that the convex conjugate of the convex function $\phi(z) := \frac{1}{2\kappa}\|z\|_{\mathcal{G}}^2$ is $v(\eta) := \phi^*(\eta) \equiv \sup_{z\in\mathcal{G}}\langle\eta, z\rangle - \phi(z) = \frac{\kappa}{2}\|\eta\|_{\mathcal{G}^*}^2$ no matter whether $\mathcal{G}$ is finite-dimensional or infinite-dimensional, without requiring any extra regularity conditions.

Since $\mathcal{G}$ is κ-smooth (Def. A.1), $\phi(z)$ is continuously differentiable ($\eta = \nabla\phi(z) \Leftrightarrow \eta \in \partial\phi(z) \equiv \{\nabla\phi(z)\}$), and for any $z, \Delta \in \mathcal{G}$:

$$\phi(z + \Delta) \leq \phi(z) + \langle\nabla\phi(z), \Delta\rangle + \frac{1}{2}\|\Delta\|_{\mathcal{G}}^2$$

We now derive Eqs (15)-(18), showing $v(\eta) \equiv \phi^*(\eta)$ is *1-strongly convex* and the proxy $V(z)$, defined as **(Eq. 15)**:

$$V(z) := \sup_{\eta \in \mathcal{B}^*} \langle \eta, z \rangle - v(\eta), \quad \mathcal{B}^* := \{\eta \in \mathcal{G}^* : \|\eta\|_{\mathcal{G}^*} \leq 1\}$$

is continuously differentiable ($\partial V(z) = \{V'(z)\}$) and $\|V'(z)\|_{\mathcal{G}^*} \leq 1$, with dual unit ball $\mathcal{B}^*$.

1-Strong Convexity of $v(\cdot)$

For base point $z \in \mathcal{G}$, let $\eta := \nabla\phi(z) \in \mathcal{G}^*$, and introduce concave $\psi_\eta(\tilde{z}) := \langle \eta, \tilde{z} \rangle - \phi(\tilde{z})$ for test point $\tilde{z} \in \mathcal{G}$. Its gradient $\nabla\psi_\eta(\tilde{z}) = \eta - \nabla\phi(\tilde{z})$ is 0 at $\tilde{z} = z$. Hence, the *Fenchel-Young* equality holds:

$$\psi_\eta(\tilde{z}) \leq \psi_\eta(z) + \langle \psi_\eta(z), \tilde{z} - z \rangle = \psi_\eta(z) \implies v(\eta) \equiv \sup_{\tilde{z} \in \mathcal{G}} \psi_\eta(\tilde{z}) = \psi_\eta(z) \equiv \langle \eta, z \rangle - \phi(z)$$

For test point $\tilde{\eta} \in \mathcal{G}^*$, $v(\tilde{\eta}) \equiv \sup_{\tilde{z} \in \mathcal{G}} \langle \tilde{\eta}, \tilde{z} \rangle - \phi(\tilde{z}) = \sup_{\Delta \in \mathcal{G}} \langle \tilde{\eta}, z + \Delta \rangle - \phi(z + \Delta)$ letting $\Delta = \tilde{z} - z$. By κ-smoothness, $\phi(z + \Delta) \leq \phi(z) + \langle \eta, \Delta \rangle + \frac{1}{2}\|\Delta\|_{\mathcal{G}}^2$. Using $v(\eta) = \langle \eta, z \rangle - \phi(z)$ and $\frac{1}{2}\|\tilde{\eta} - \eta\|_{\mathcal{G}^*}^2 = \sup_{\Delta \in \mathcal{G}} \langle \tilde{\eta} - \eta, \Delta \rangle - \frac{1}{2}\|\Delta\|_{\mathcal{G}}^2$:

$$v(\tilde{\eta}) \geq \sup_{\Delta \in \mathcal{G}} \langle \tilde{\eta}, z + \Delta \rangle - \phi(z) - \langle \eta, \Delta \rangle - \frac{1}{2}\|\Delta\|_{\mathcal{G}}^2 = v(\eta) + \langle \tilde{\eta} - \eta, z \rangle + \frac{1}{2}\|\tilde{\eta} - \eta\|_{\mathcal{G}^*}^2$$

Thus $v(\tilde{\eta}) \geq v(\eta) + \langle \tilde{\eta} - \eta, z \rangle$, implying $z \in \partial v(\eta)$. Repeating this for any element in $\partial v(\eta)$ yields the same inequality, proving $v(\cdot)$ is 1-strongly convex.

$V(z)$ is Continuously Differentiable and $\|V'(z)\|_{\mathcal{G}^*} \leq 1$

Since $\mathcal{B}^* \equiv \{\eta \in \mathcal{G}^* : \|\eta\|_{\mathcal{G}^*} \leq 1\}$ is compact under the weak-* topology (Banach-Alaoglu theorem) and $v(\eta)$ is lower semicontinuous, the function $\langle \eta, z \rangle - v(\eta)$ is upper semicontinuous under the weak-* topology, which guarantees the existence of $\arg\max_{\eta \in \mathcal{B}^*}\{\langle \eta, z \rangle - v(\eta)\}$ (Generalized Weierstrass theorem). Since $v(\cdot)$ is 1-strongly convex, $\langle \eta, z \rangle - v(\eta)$ is strictly concave over convex $\mathcal{B}^*$, ensuring a unique maximizer η_z:

$$\eta_z := \arg\max_{\eta \in \mathcal{B}^*}\{\langle \eta, z \rangle - v(\eta)\} \in \mathcal{B}^*$$

The *Fenchel-Young* relation implies $V(\tilde{z}) \geq V(z) + \langle \eta_z, \tilde{z} - z \rangle$, namely:

$$V(z) = \sup_{\eta \in \mathcal{B}^*} \langle \eta, z \rangle - v(\eta) = \langle \eta_z, z \rangle - v(\eta_z) \implies \eta_z \in \partial V(z)$$

$z \mapsto \eta_z$ is 1-Lipschitz

$\Psi_z(\eta) := \langle \eta, z \rangle - v(\eta)$ is 1-strongly concave with unique maximizer η_z in convex $\mathcal{B}^*$. For $\alpha \in (0, 1)$ and test point $\tilde{\eta} \in \mathcal{B}^*$:

$$\Psi_z(\eta_z) \geq \Psi_z(\alpha\eta_z + (1 - \alpha)\tilde{\eta}) \geq \alpha\Psi_z(\eta_z) + (1 - \alpha)\Psi_z(\tilde{\eta}) + \frac{\alpha(1 - \alpha)}{2}\|\eta_z - \tilde{\eta}\|_{\mathcal{G}^*}^2$$

Rearranging, dividing by $1 - \alpha$, and taking $\alpha \to 1$ yields:

$$v(\tilde{\eta}) - v(\eta_z) - \langle \tilde{\eta} - \eta_z, z \rangle \geq \frac{1}{2}\|\eta_z - \tilde{\eta}\|_{\mathcal{G}^*}^2$$

Let $\eta_1 := \eta_{z_1}$ and $\eta_2 := \eta_{z_2}$. Summing the two inequalities given by $(z \leftarrow z_1, \eta_z \leftarrow \eta_1, \tilde{\eta} \leftarrow \eta_2)$ and $(z \leftarrow z_2, \eta_z \leftarrow \eta_2, \tilde{\eta} \leftarrow \eta_1)$ yields:

$$\|\eta_1 - \eta_2\|_{\mathcal{G}^*}^2 \leq \langle \eta_1 - \eta_2, z_1 - z_2 \rangle \leq \|\eta_1 - \eta_2\|_{\mathcal{G}^*}\|z_1 - z_2\|_{\mathcal{G}} \implies \|\eta_1 - \eta_2\|_{\mathcal{G}^*} \leq \|z_1 - z_2\|_{\mathcal{G}}$$

$\partial V(z) = \{\eta_z\}$ is a singleton

For $\eta \in \partial V(z)$ and $\Delta \in \mathcal{G}$, since $V(z) \geq \langle \eta_{z+\Delta}, z \rangle - v(\eta_{z+\Delta}) = V(z + \Delta) - \langle \eta_{z+\Delta}, \Delta \rangle$:

$$\langle \eta, \Delta \rangle \leq V(z + \Delta) - V(z) \leq \langle \eta_{z+\Delta}, \Delta \rangle$$

By the 1-Lipschitz continuity of $z \mapsto \eta_z$:

$$\langle \eta - \eta_z, \Delta \rangle \leq \langle \eta_{z+\Delta} - \eta_z, \Delta \rangle \leq \|\Delta\|_{\mathcal{G}}^2$$

Letting $\Delta \leftarrow -\Delta$ gives $-\|\Delta\|_{\mathcal{G}}^2 \leq \langle \eta - \eta_z, \Delta \rangle$. For fixed direction $d \in \mathcal{G}$ and $t > 0$, substituting $\Delta = td$:

$$t|\langle \eta - \eta_z, d \rangle| \leq t^2 \|d\|_{\mathcal{G}}^2 \implies |\langle \eta - \eta_z, d \rangle| \leq \lim_{t \to 0} t\|d\|_{\mathcal{G}}^2 = 0$$

forcing $\langle \eta - \eta_z, d \rangle = 0$ for all d.

Thus $\eta = \eta_z$, $\partial V(z) = \{\eta_z\}$, and $V'(z) = \eta_z$ exists. Since $\eta_z \in \mathcal{B}^*$ and $z \mapsto \eta_z$ is 1-Lipschitz:

$$\|V'(z)\|_{\mathcal{G}^*} = \|\eta_z\|_{\mathcal{G}^*} \leq 1, \quad \|V'(z_1) - V'(z_2)\|_{\mathcal{G}^*} \leq \|z_1 - z_2\|_{\mathcal{G}} \implies V(z) \text{ is continuously differentiable}$$

Lipschitz Gradient and Smoothness (Eq. 16)

For $z, \Delta \in \mathcal{G}$ and $t \geq 0$, via 1-Lipschitz continuity of $V'(z) \equiv \eta_z$:

$$\langle V'(z + t\Delta) - V'(z), \Delta \rangle \leq t\|\Delta\|_{\mathcal{G}}^2$$

By the Fundamental Theorem of Calculus:

$$V(\xi + \Delta) = V(\xi) + \langle V'(\xi), \Delta \rangle + \int_0^1 \langle V'(\xi + t\Delta) - V'(\xi), \Delta \rangle dt \leq V(\xi) + \langle V'(\xi), \Delta \rangle + \frac{1}{2}\|\Delta\|_{\mathcal{G}}^2$$

Primal Geometric Offset (Eq. 17)

Since $\|z\|_{\mathcal{G}} \equiv \sup_{\eta \in \mathcal{B}^*} \langle \eta, z \rangle$ with $\mathcal{B}^* \equiv \{\eta \in \mathcal{G}^* : \|\eta\|_{\mathcal{G}^*} \leq 1\}$. For $v(\eta) \equiv \frac{\kappa}{2}\|\eta\|_{\mathcal{G}^*}^2$ and $V(z) \equiv \sup_{\eta \in \mathcal{B}^*} \langle \eta, z \rangle - v(\eta)$, applying $\sup(A + B) \leq \sup A + \sup B$ yields:

$$\|z\|_{\mathcal{G}} \leq \sup_{\eta \in \mathcal{B}^*} \{\langle \eta, z \rangle - v(\eta)\} + \sup_{\eta \in \mathcal{B}^*} v(\eta) = V(z) + \frac{\kappa}{2}$$

Scaling via β (Eq. 18)

For scale $\beta > 0$, $V_\beta(z) := \beta V(z/\beta)$. By the chain rule, $V'_\beta(z) = V'(z/\beta)$. Since $\|V'(\cdot)\|_{\mathcal{G}^*} \leq 1$, $\|V'_\beta(z)\|_{\mathcal{G}^*} \leq 1$. Evaluating Eq. 17 at z/β and multiplying by β gives the scaled offset via absolute homogeneity:

$$\|z\|_{\mathcal{G}} = \beta \left\| \frac{z}{\beta} \right\|_{\mathcal{G}} \leq \beta \left(V\left(\frac{z}{\beta}\right) + \frac{\kappa}{2} \right) = V_\beta(z) + \frac{\kappa\beta}{2}$$

Evaluating Eq. 16 at base z/β with displacement Δ/β, and scaling by β gives:

$$V_\beta(z+\Delta) = \beta V\left(\frac{z}{\beta} + \frac{\Delta}{\beta}\right) \leq \beta \left(V\left(\frac{z}{\beta}\right) + \left\langle V'\left(\frac{z}{\beta}\right), \frac{\Delta}{\beta} \right\rangle + \frac{1}{2}\left\|\frac{\Delta}{\beta}\right\|_{\mathcal{G}}^2 \right) = V_\beta(z) + \langle V'_\beta(z), \Delta \rangle + \frac{1}{2\beta}\|\Delta\|_{\mathcal{G}}^2$$

Remark B.4 (Optimization Implications)**.** The formulation of Theorem B.2 exhibits a strict topological decoupling between the martingale geometry and the variance proxy constructions. The κ-smoothness condition is imposed exclusively on the space $(\mathcal{G}, \|\cdot\|_{\mathcal{G}})$ governing the martingale difference sequence Y_t. The predictable variance proxy Σ_t operates purely as an $\mathcal{F}_{t-1}$-measurable scalar bound. Consequently, if Σ_t is parameterized by auxiliary algorithmic quantities, e.g., $\|U\|_q$ or $\|w - w'\|_r$ for arbitrary $q, r \in [1, \infty]$, the probabilistic concentration mechanism remains completely agnostic to these auxiliary topologies. However, integrating this concentration into the ensuing application of optimization analysis imposes certain structural constraints on the permissible ℓ_p norms, dictated entirely by the nature of the stochastic sequence:

- **Concentration of Gradients (Typical Application):** If the martingale models stochastic gradient residuals, the sequence inherently resides in the topological dual space, $\mathcal{G} = \mathcal{X}^*$. To invoke the theorem natively without dimension-dependent distortion, $\mathcal{X}^*$ must be κ-smooth, which necessitates an ℓ_p norm geometry where $p \in [2, \infty)$. By strict Hölder conjugacy ($1/p + 1/q = 1$), the primal iterate space $\mathcal{X}$ is formally restricted to an ℓ_q geometry where $q \in (1, 2]$.
- **Concentration of Function Values (SGM Application):** If the martingale models scalar stochastic noise in the objective evaluations, the sequence resides in $\mathcal{G} = \mathbb{R}$. The absolute value norm constitutes a trivially 1-smooth Hilbert geometry ($\kappa = 1$). The resulting concentration is unconditionally dimension-free. Under this regime, the primal iterate space $\mathcal{W}$ may be equipped with any arbitrary norm without inducing geometric penalties, as the concentration bound does not interact with the dual pairing of the algorithmic step.

B.3 Proof for Theorem 4.2 of Unified Estimation Error Bound

Proof of Theorem 4.2. For any horizon $T \geq 2$, we rigorously establish a uniform upper bound for the estimation error $\|e_t\|_{\mathcal{G}}$ that holds simultaneously for all $t \in [T]$ with probability at least $1 - \delta$.

I. Resets, Stopping Time, and Epoch Index

- *Filtration:* Let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t=0}^T, \mathbb{P})$ be a filtered space.
- *Resets:* Let $b_t \in \{0, 1\}$ be an $\mathcal{F}_t$-measurable random variable representing the reset decision at time t with $b_0 = 1$. Let $N_t := \sum_{s=1}^t b_s$ denote the cumulative number of resets up to time t.
- *Stopping Time:* Let $\tau_m := \inf\{t > \tau_{m-1} : b_t = 1\}$ with $\tau_0 = 0$. Since $\{\tau_m \leq t\} \equiv \{N_t \geq m\} \in \mathcal{F}_t$, τ_m is a valid stopping time with respect to the filtration.
- *Epoch Index:* For $t \geq 1$, the predictable ($\mathcal{F}_{t-1}$-measurable) active epoch index is $m(t) := N_{t-1} + 1$. The most recent prior reset is $\tau_{m(t)-1}$, meaning $b_j = 0$ for $j \in (\tau_{m(t)-1}, t-1]$.

II. Error Dynamics

The estimation error e_t (with initial value $e_0 = G(w_0, \xi_0^B) - g(w_0)$ and $b_0 = 1$) satisfies the recursive dynamic:

$$e_t = b_t E_{B,t} + (1 - b_t)(\beta_t e_{t-1} + \mu_t + \mathcal{Z}_t)$$

where:

- *Batch Error:* $E_{B,t} := G(w_t, \xi_t^B) - g(w_t)$ satisfies the sub-Gaussianity condition of Assumption 3.2, where $G(w_t, \xi_t^B) := \frac{1}{B}\sum_{i=1}^B G(w_t, \xi_t^{(i)})$ denotes the empirical average over a batch of independent samples $\xi_t^B := (\xi_t^{(i)})_{i=1}^B$.
- *Decay:* Predictable factors $\beta_t \in (0, 1]$ and scalars $\Lambda_{t,j} := \prod_{k=j+1}^t \beta_k = A_t^{-1} A_j$, with $A_t^{-1} \equiv \prod_{k=1}^t \beta_k$.
- *Innovation:* $\mathcal{I}_t(\mathcal{T}) \equiv \mu_t + \mathcal{Z}_t$. The predictable bias satisfies $\|\mu_t\|_{\mathcal{G}} \leq \mathcal{B}_t$. The noise $\mathcal{Z}_t$ is a martingale difference ($\mathbb{E}[\mathcal{Z}_t \mid \mathcal{F}_{t-1}] = 0$) with predictable variance proxy Σ_t^2.

Unrolling e_{t-1} from $\tau_{m(t)-1}$ and noting $b_j = 0$ for $j \in (\tau_{m(t)-1}, t-1]$ isolates the components:

$$e_t = \underbrace{b_t E_{B,t} + (1-b_t)\Lambda_{t,\tau_{m(t)-1}} E_{B,\tau_{m(t)-1}}}_{\text{Batch Error}} + (1-b_t) \underbrace{\sum_{j=\tau_{m(t)-1}+1}^{t} \Lambda_{t,j}\mu_j}_{\text{Predictable Bias}} + (1-b_t)\, \underbrace{A_t^{-1} \sum_{j=\tau_{m(t)-1}+1}^{t} A_j \mathcal{Z}_j}_{\text{Stochastic Noise}}$$

Analyzing b_t isolates the most recent reset at τ_{N_t}. Summing the disjoint states ($b_t = 1$ where $\tau_{N_t} = t$; $b_t = 0$ where $\tau_{N_t} = \tau_{m(t)-1}$) and defining empty sums as zero ($\sum_{j=s}^t(\cdot) \equiv 0$ if $s > t$) establishes the identity:

$$e_t = \underbrace{\Lambda_{t,\tau_{N_t}} E_{B,\tau_{N_t}}}_{\text{Batch Error}} + \underbrace{\sum_{j=\tau_{N_t}+1}^{t} \Lambda_{t,j}\mu_j}_{\text{Predictable Bias}} + \underbrace{A_t^{-1} \sum_{j=\tau_{N_t}+1}^{t} A_j \mathcal{Z}_j}_{\text{Stochastic Noise}}$$

(Mapping to Localized Martingales). For epoch $m \in \{1, \ldots, T\}$, define $I_j^{(m)} := \mathbb{I}\{\tau_{m-1} < j \leq \tau_m\} \equiv \mathbb{I}\{N_{j-1} = m-1\}$. Since $I_j^{(m)}$ depends only on history up to $j-1$, it is predictable.

Thus, increments $I_j^{(m)} A_j \mathcal{Z}_j$ form a martingale difference sequence (see the similar argument in Lemma B.1):

$$\mathbb{E}\left[I_j^{(m)} A_j \mathcal{Z}_j \mid \mathcal{F}_{j-1}\right] = I_j^{(m)} A_j \mathbb{E}[\mathcal{Z}_j \mid \mathcal{F}_{j-1}] = 0.$$

Hence, $M_t^{(m)} := \sum_{j=1}^{t} I_j^{(m)} A_j \mathcal{Z}_j$ is a martingale satisfying $\mathbb{E}[M_t^{(m)} \mid \mathcal{F}_{t-1}] = M_{t-1}^{(m)}$.

At step t, the accumulated noise corresponds to epoch $N_t + 1 \equiv m(t+1)$. Since $j \le \tau_{N_t+1}$ holds for $j \le t$, the sum maps to:

$$\sum_{j=\tau_{N_t}+1}^{t} A_j \mathcal{Z}_j = \sum_{j=1}^{t} \mathbb{I}\{\tau_{N_t} < j \le \tau_{N_t+1}\} A_j \mathcal{Z}_j \equiv M_t^{(N_t+1)}$$

Applying the triangle inequality and $\|\mu_j\|_{\mathcal{G}} \le \mathcal{B}_j$ gives:

$$\|e_t\|_{\mathcal{G}} \le \Lambda_{t,\tau_{N_t}} \left\|E_{B,\tau_{N_t}}\right\|_{\mathcal{G}} + \sum_{j=\tau_{N_t}+1}^{t} \Lambda_{t,j} \mathcal{B}_j + A_t^{-1} \left\|M_t^{(N_t+1)}\right\|_{\mathcal{G}}$$

III. High-Probability Bounds

(Batch Concentration). Since $\tau_{N_t} \le t \le T$, any realization of τ_{N_t} maps to the deterministic sequence $\{0, \dots, T\}$ of size $T+1$. Allocating local confidence $\frac{\delta}{2T}$ per evaluation, applying a union bound yields the success event $\mathcal{E}_B$:

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_B^c) &= \mathbb{P}\left(\exists t \in [T] : \|E_{B,\tau_{N_t}}\|_{\mathcal{G}} > C\left(\frac{\delta}{2T}, \kappa\right) \frac{\sigma}{\sqrt{B}}\right) \\
&\le \mathbb{P}\left(\bigcup_{k=0}^{T} \left\{\|E_{B,k}\|_{\mathcal{G}} > C\left(\frac{\delta}{2T}, \kappa\right) \frac{\sigma}{\sqrt{B}}\right\}\right) \le \sum_{k=0}^{T} \left(\frac{\delta}{2T}\right) = \frac{T+1}{2T}\delta
\end{aligned}$$

(Martingale Configuration). At each step $t \in [T]$, the cumulative resets satisfy $N_t \le t$, meaning $N_t + 1$ takes possible deterministic values $m \in \{1, \dots, t+1\}$.

Because $N_{j-1} \le j-1$, the indicator $\mathbb{I}\{N_{j-1} = m-1\}$ within $M_t^{(m)} := \sum_{j=1}^{t} \mathbb{I}\{N_{j-1} = m-1\} A_j \mathcal{Z}_j$ requires $m \le j \le t$ to be non-zero. Consequently, $M_t^{(m)} \equiv 0$ almost surely for all $m > t$. Therefore, any realization of the random pair $(N_t + 1, t) = (m, t)$ is restricted by the deterministic, upper-triangular grid $1 \le m \le t \le T$. The exact number of these deterministic pairs is $\sum_{t=1}^{T} \sum_{m=1}^{t} 1 = \sum_{t=1}^{T} t = \frac{T(T+1)}{2}$.

Allocating local confidence $\frac{\delta}{2T(T+1)}$ and substituting coordinate (m, t) with envelope $V_t \ge \sum_{j=\tau_{m-1}+1}^{t} A_j^2 \Sigma_j^2$ into Theorem B.2 of Vector-Valued Freedman Inequality guarantees:

$$\mathbb{P}\left(\|M_t^{(m)}\|_{\mathcal{G}} > C\left(\frac{\delta}{2T(T+1)}, \kappa\right) \sqrt{V_t}\right) \le \frac{\delta}{2T(T+1)}$$

Taking the union bound over this grid envelopes the realized noise, yielding event $\mathcal{E}_M$:

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_M^c) &= \mathbb{P}\left(\exists t \in [T] : \|M_t^{(N_t+1)}\|_{\mathcal{G}} > C\left(\frac{\delta}{2T(T+1)}, \kappa\right) \sqrt{V_t}\right) \\
&\le \mathbb{P}\left(\bigcup_{t=1}^{T} \bigcup_{m=1}^{t} \left\{\|M_t^{(m)}\|_{\mathcal{G}} > C\left(\frac{\delta}{2T(T+1)}, \kappa\right) \sqrt{V_t}\right\}\right) \le \sum_{t=1}^{T} \sum_{m=1}^{t} \frac{\delta}{2T(T+1)} = \frac{\delta}{4}
\end{aligned}$$

Global Bound. The global success event

$$\mathcal{E}^* := \mathcal{E}_B \cap \mathcal{E}_M = \left\{\forall t \in [T], \left\|E_{B,\tau_{N_t}}\right\|_{\mathcal{G}} \le C\left(\frac{\delta}{2T}, \kappa\right) \frac{\sigma}{\sqrt{B}}, \left\|M_t^{(N_t+1)}\right\|_{\mathcal{G}} \le C\left(\frac{\delta}{2T(T+1)}, \kappa\right) \sqrt{V_t}\right\}$$

holds with failure probability bounded by the direct sum for $T \geq 2$:

$$\mathbb{P}((\mathcal{E}^*)^c) \leq \frac{T+1}{2T}\delta + \frac{1}{4}\delta = \frac{3T+2}{4T}\delta \leq \delta \implies \mathbb{P}(\mathcal{E}^*) \geq 1-\delta$$

Conditioned on $\mathcal{E}^*$, this holds for all $t \in [T]$:

$$\|e_t\|_{\mathcal{G}} \leq \underbrace{\Lambda_{t,\tau_{N_t}} C\left(\frac{\delta}{2T}, \kappa\right) \frac{\sigma}{\sqrt{B}}}_{\text{Decayed Batch Error}} + \underbrace{\sum_{j=\tau_{N_t}+1}^{t} \Lambda_{t,j}\mathcal{B}_j}_{\text{Predictable Bias}} + \underbrace{A_t^{-1} C\left(\frac{\delta}{2T(T+1)}, \kappa\right)\sqrt{V_t}}_{\text{Martingale Noise}}$$

Discussion on Bounds

By the definition $C(\delta, \kappa) := \sqrt{\kappa} + \sqrt{3\ln(1/\delta)}$ and observing:

$$\ln \frac{2T(T+1)}{\delta} = \ln \frac{1}{\delta} + \ln(2T) + \ln(T+1) \leq 2\left(\ln \frac{1}{\delta} + \ln(2T)\right) = 2\ln \frac{2T}{\delta}$$

we can bound $C\left(\frac{\delta}{2T(T+1)}, \kappa\right)$ by $C\left(\frac{\delta}{2T}, \kappa\right)$:

$$C\left(\frac{\delta}{2T}, \kappa\right) \leq C\left(\frac{\delta}{2T(T+1)}, \kappa\right) \leq \sqrt{2} C\left(\frac{\delta}{2T}, \kappa\right)$$

Thus, the simplified bound for all $t \in [T]$ with $T \geq 2$ is:

$$\|e_t\|_{\mathcal{G}} \leq \underbrace{\Lambda_{t,\tau_{N_t}} C\left(\frac{\delta}{2T}, \kappa\right) \frac{\sigma}{\sqrt{B}}}_{\text{Decayed Batch Error}} + \underbrace{\sum_{j=\tau_{N_t}+1}^{t} \Lambda_{t,j}\mathcal{B}_j}_{\text{Predictable Bias}} + \underbrace{\sqrt{2} A_t^{-1} C\left(\frac{\delta}{2T}, \kappa\right)\sqrt{V_t}}_{\text{Martingale Noise}}$$

□

Remark B.5 (Tightened Upper Bound for Case 1 with $p_t = 0$)**.** When $p_t = 0$ for all $t \in [T]$, the reset indicator $b_t \sim \text{Bernoulli}(p_t)$ satisfies $b_t = 0$ for all $t \in [T]$ with a probability of 1. Conditioned on the event $\{b_t = 0, \forall t \in [T]\}$, the cumulative number of resets $N_t := \sum_{s=1}^{t} b_s$ (with $N_0 = 0$) and the epoch index $m(t) := N_{t-1} + 1$ evaluate to:

$$N_t = 0, m(t) = 0 + 1 = 1, \quad \forall t \in [T] \implies \tau_{N_t} = \tau_{N_{t-1}} = \tau_{m(t)-1} = \tau_0 = 0$$

By examining the expression for e_t and substituting $\tau_{N_t} = \tau_{m(t)-1} = 0$, we obtain:

$$e_t = \underbrace{\Lambda_{t,0} E_{B,0}}_{\text{Batch Error}} + \underbrace{\sum_{j=1}^{t} \Lambda_{t,j}\mu_j}_{\text{Predictable Bias}} + \underbrace{A_t^{-1} M_t^{(1)}}_{\text{Stochastic Noise}}$$

Allocating a confidence of $\frac{\delta}{2}$ to bound the batch concentration term $\|E_{B,0}\|_{\mathcal{G}}$, and a confidence of $\frac{\delta}{2T}$ to bound each martingale term $M_t^{(1)}$ for all $t \in [T]$, the applications of Theorem B.2 (Vector-Valued Freedman Inequality) yield:

$$\|E_{B,0}\|_{\mathcal{G}} \leq C\left(\frac{\delta}{2}, \kappa\right)\frac{\sigma}{\sqrt{B}}, \quad \|M_t^{(1)}\|_{\mathcal{G}} \leq C\left(\frac{\delta}{2T}, \kappa\right)\sqrt{V_t}, \forall t \in [T]$$

with probability at least $1 - \left(\frac{\delta}{2} + T \times \frac{\delta}{2T}\right) = 1 - \delta$.

Applying these inequalities alongside the bound $\|\mu_j\|_{\mathcal{G}} \leq \mathcal{B}_j$, we derive a tightened upper bound for the special case of $p_t = 0$:

$$\|e_t\|_{\mathcal{G}} \leq \underbrace{\Lambda_{t,0} C\left(\frac{\delta}{2}, \kappa\right) \frac{\sigma}{\sqrt{B}}}_{\text{Decayed Batch Error}} + \underbrace{\sum_{j=1}^{t} \Lambda_{t,j} \mathcal{B}_j}_{\text{Predictable Bias}} + \underbrace{A_t^{-1} C\left(\frac{\delta}{2T}, \kappa\right) \sqrt{V_t}}_{\text{Martingale Noise}}$$

In this tightened upper bound, the coefficient for the decayed batch error is tightened from $C\left(\frac{\delta}{2T}, \kappa\right)$ to the smaller $C\left(\frac{\delta}{2}, \kappa\right)$, and the $\sqrt{2}$ factor in the martingale noise term is dropped because we are bounding a single martingale sequence within one reset epoch rather than accounting for multiple reset epochs. While the tightened bound is more precise, the original general upper bound remains valid (albeit slightly looser than necessary) for this specific case.

C Proofs for Instantiations of the Unified Estimation Error Bound

C.1 Propositions for Three Families with Zeroth / First / Second Order Corrections

The strength of Theorem 4.2 lies in its ability to recover distinct convergence mechanisms via specific choices of the control sequences $\{p_t, \beta_t, \eta_t, \mathcal{T}_t\}$. In this section, we consider three special cases corresponding to the three families of estimators we discussed previously.

Proposition C.1 (**Family 1** with Zeroth-Order Correction)**.** *Suppose Assumptions 2.3, 3.1, 3.2, and 5.1 hold. Then, for the zeroth-order family, $\mathcal{B}_t = \beta_t \eta_{t-1} GL$ and $\Sigma_t^2 = (1-\beta_t)^2\sigma^2$. Consequently, Theorem 4.2 yields*

$$\|e_t\|_{\mathcal{G}} \le \Lambda_{t,\tau_{N_t}} C\left(\frac{\delta}{2T}, \kappa\right) \sigma/\sqrt{B} + LG \sum_{j=\tau_{N_t}+1}^{t} \Lambda_{t,j}\beta_j\eta_{j-1} + \sqrt{2}C\left(\frac{\delta}{2T}, \kappa\right) A_t^{-1}\sqrt{V_t}, \quad (33)$$

whenever $\sum_{j=\tau_{N_t}+1}^{t} A_j^2(1-\beta_j)^2\sigma^2 \le V_t$.

Proposition C.1 shows that the zeroth-order family attenuates the oracle noise by the factor $(1-\beta_t)$, but this variance reduction comes at the cost of a tracking bias proportional to the step displacement. This is precisely the classical tradeoff in momentum-style estimation: larger momentum suppresses fresh noise, while simultaneously increasing the lag induced by the moving target.

Proposition C.2 (**Family 2** with First-Order Correction)**.** *Suppose Assumptions 2.3, 3.1, 3.2, and 5.2 hold. Then, for the first-order family, $\mathcal{B}_t = 0$ and $\Sigma_t^2 = 2(1-\beta_t)^2\sigma^2 + 2\beta_t^2\ell^2\eta_{t-1}^2G^2$. Consequently, Theorem 4.2 yields*

$$\|e_t\|_{\mathcal{G}} \le \Lambda_{t,\tau_{N_t}} C\left(\frac{\delta}{2T}, \kappa\right) \sigma/\sqrt{B} + \sqrt{2}C\left(\frac{\delta}{2T}, \kappa\right) A_t^{-1}\sqrt{V_t}, \quad (34)$$

whenever $\sum_{j=\tau_{N_t}+1}^{t} A_j^2\Big(2(1-\beta_j)^2\sigma^2 + 2\beta_j^2\ell^2\eta_{j-1}^2G^2\Big) \le V_t$.

Proposition C.2 isolates the essential advantage of first-order differential estimators: the conditional bias disappears entirely. The price is that the variance term now depends on the centered stochastic difference, whose scale is proportional to the distance $\|w_t - w_{t-1}\|$. In other words, this family replaces tracking lag with local difference noise.

Proposition C.3 (**Family 3** with Second-Order Correction)**.** *Suppose Assumptions 2.3, 3.1, 3.2, 5.3, and 5.4 hold. Then, for the second-order family, $\mathcal{B}_t = \frac{\alpha}{2}\beta_t\eta_{t-1}^2G^2$ and $\Sigma_t^2 = 2(1-\beta_t)^2\sigma^2 + 2\beta_t^2\gamma^2\eta_{t-1}^2G^2$. Consequently, Theorem 4.2 yields*

$$\|e_t\|_{\mathcal{G}} \le \Lambda_{t,\tau_{N_t}} C\left(\frac{\delta}{2T}, \kappa\right) \sigma/\sqrt{B} + \frac{\alpha G^2}{2} \sum_{j=\tau_{N_t}+1}^{t} \Lambda_{t,j}\beta_j\eta_{j-1}^2 + \sqrt{2}C\left(\frac{\delta}{2T}, \kappa\right) A_t^{-1}\sqrt{V_t}, \quad (35)$$

whenever $\sum_{j=\tau_{N_t}+1}^{t} A_j^2\Big(2(1-\beta_j)^2\sigma^2 + 2\beta_j^2\gamma^2\eta_{j-1}^2G^2\Big) \le V_t$.

Proposition C.3 makes explicit the gain obtained from using curvature information. Relative to the zeroth-order family, the deterministic bias is reduced from first order in the displacement to second order. Relative to the first-order family, the remaining stochastic fluctuation is now tied to Hessian-vector product noise rather than to stochastic differences. Thus, the second-order correction interpolates between low drift and curvature-sensitive variance control.

Before providing the proofs for the propositions of three families, we first state two facts which will be used in the proofs.

Fact 1. Under Assumption 2.3 for U_t in the mirror descent $w_{t+1} = w_t - \eta_t P(w_t, U_t, \eta_t)$, then $\|w_t - w_{t-1}\| \le \eta_{t-1}G$ is bounded almost surely.

Fact 2. For the random vectors $X, Y \in \mathcal{G}$ that admit variance proxies Σ_X^2, Σ_Y^2 such that $\mathbb{E}[\exp(\|X\|_{\mathcal{G}}^2/\Sigma_X^2)] \le 2$ and $\mathbb{E}[\exp(\|Y\|_{\mathcal{G}}^2/\Sigma_Y^2)] \le 2$, then $X+Y$ possesses a variance proxy $2\Sigma_X^2 + 2\Sigma_Y^2$ such that $\mathbb{E}[\exp(\|X+Y\|_{\mathcal{G}}^2/(2\Sigma_X^2 + 2\Sigma_Y^2))] \le 2$.

Fact 1 is established by noting $\|w_t - w_{t-1}\| = \|\eta_{t-1} P(w_{t-1}, U_{t-1}, \eta_{t-1})\| \le \eta_{t-1}\|U_{t-1}\|_*$ from Lemma A.7 and $\|U_{t-1}\|_* \le G$ from Assumption 2.3.

For fact 2, we establish a deterministic upper bound. For any two arbitrary, potentially dependent random vectors X and Y admitting variance proxies Σ_X^2 and Σ_Y^2, the triangle inequality and the geometric inequality $\|X+Y\|_{\mathcal{G}}^2 \le (\|X\|_{\mathcal{G}} + \|Y\|_{\mathcal{G}})^2 \le 2\|X\|_{\mathcal{G}}^2 + 2\|Y\|_{\mathcal{G}}^2$, paired with the convexity of the exponential function, dictates that

$$\mathbb{E}\left[\exp\left(\frac{\|X+Y\|_{\mathcal{G}}^2}{2\Sigma_X^2+2\Sigma_Y^2}\right)\right] \le \frac{\Sigma_X^2}{\Sigma_X^2+\Sigma_Y^2}\mathbb{E}\left[\exp\left(\frac{\|X\|_{\mathcal{G}}^2}{\Sigma_X^2}\right)\right] + \frac{\Sigma_Y^2}{\Sigma_X^2+\Sigma_Y^2}\mathbb{E}\left[\exp\left(\frac{\|Y\|_{\mathcal{G}}^2}{\Sigma_Y^2}\right)\right] \le 2.$$

Consequently, the summation of two conditionally dependent sub-Gaussian vectors inherently possesses a variance proxy bounded by $2\Sigma_X^2 + 2\Sigma_Y^2$.

Proof of Proposition C.1. Under the zeroth-order correction, the effective innovation becomes

$$\mathcal{I}_t(\mathcal{T}) = (1-\beta_t)\big(G(w_t,\xi_t) - g(w_t)\big) - \beta_t\big(g(w_t) - g(w_{t-1})\big).$$

Taking conditional expectation, using Assumption 5.1, and noting fact 1 $\|w_t - w_{t-1}\| \le \eta_{t-1}G$ under Assumption 2.3

$$\begin{aligned}\|\mathbb{E}[\mathcal{I}_t(\mathcal{T}) \mid \mathcal{F}_{t-1}]\|_{\mathcal{G}} &= \beta_t \|\mathbb{E}[g(w_t) - g(w_{t-1}) \mid \mathcal{F}_{t-1}]\|_{\mathcal{G}} \\ &\overset{\text{Jensen's}}{\le} \beta_t \mathbb{E}[\|g(w_t) - g(w_{t-1})\|_{\mathcal{G}} \mid \mathcal{F}_{t-1}] \\ &\overset{\text{As. 5.1}}{\le} \beta_t L \mathbb{E}[\|w_t - w_{t-1}\| \mid \mathcal{F}_{t-1}] \le \beta_t \eta_{t-1} G L =: \mathcal{B}_t,\end{aligned}$$

The centered innovation is

$$\mathcal{Z}_t = \mathcal{I}_t(\mathcal{T}) - \mathbb{E}[\mathcal{I}_t(\mathcal{T}) \mid \mathcal{F}_{t-1}] = (1-\beta_t)\big(G(w_t,\xi_t) - g(w_t)\big),$$

Assumption 4.1 implies $\Sigma_t^2 = (1-\beta_t)^2\sigma^2$. Substituting these quantities into Theorem 4.2 gives (33). □

Proof of Proposition C.2. For $\mathcal{T}_t = 0$, the effective innovation is

$$\mathcal{I}_t(\mathcal{T}) = (1-\beta_t)\big(G(w_t,\xi_t) - g(w_t)\big) + \beta_t \Delta(w_t, w_{t-1}, \xi_t).$$

Both terms have zero conditional mean given $\mathcal{F}_{t-1}$, so $\mathcal{B}_t = \|\mathbb{E}[\mathcal{I}_t(\mathcal{T}) \mid \mathcal{F}_{t-1}]\|_{\mathcal{G}} = 0$. It remains to control the centered innovation, which in this case equals $\mathcal{I}_t(\mathcal{T})$ itself.

To construct the variance proxy Σ_t^2, we must account for the conditional dependence of the noise sources all evaluated at the same random realization ξ_t. Because these components share the same underlying random seed conditionally on $\mathcal{F}_{t-1}$, we cannot sum their variance proxies as if they were conditionally independent or orthogonal martingale sequences.

By applying the above fact 1 and fact 2 and utilizing the Assumption 2.3, the variance proxy for $\mathcal{Z}_t$ is bounded by

$$\Sigma_t^2 = 2(1-\beta_t)^2\sigma^2 + 2\beta_t^2\ell^2\eta_{t-1}^2 G^2.$$

Substituting $\mathcal{B}_t = 0$ and this variance proxy into Theorem 4.2 gives (34). □

Proof of Proposition C.3. Substituting the second-order correction into the effective innovation gives

$$\mathcal{I}_t(\mathcal{T}) = (1-\beta_t)\big(G(w_t,\xi_t) - g(w_t)\big) + \beta_t\big(g(w_{t-1}) - g(w_t) - \nabla G(w_t,\xi_t)(w_{t-1} - w_t)\big).$$

Taking conditional expectation and using $\mathbb{E}[\nabla G(w_t,\xi_t) \mid \mathcal{F}_{t-1}] = \nabla g(w_t)$, we obtain

$$\|\mathbb{E}[\mathcal{I}_t(\mathcal{T}) \mid \mathcal{F}_{t-1}]\|_{\mathcal{G}} = \beta_t \|g(w_{t-1}) - g(w_t) - \nabla g(w_t)(w_{t-1} - w_t)\|_{\mathcal{G}}.$$

Assumption 5.4 yields the second-order Taylor remainder bound, and noting fact 1 under Assumption 2.3.

$$\|\mathbb{E}[\mathcal{I}_t(\mathcal{T}) \mid \mathcal{F}_{t-1}]\|_{\mathcal{G}} \le \frac{\alpha}{2}\beta_t\|w_t - w_{t-1}\|^2 \le \frac{\alpha}{2}\beta_t\eta_{t-1}^2 G^2 =: \mathcal{B}_t.$$

The centered innovation $\mathcal{Z}_t = \mathcal{I}_t(\mathcal{T}) - \mathbb{E}[\mathcal{I}_t(\mathcal{T}) \mid \mathcal{F}_{t-1}]$ consequently distills into two purely stochastic components, representing the direct gradient evaluation noise and the error in the Hessian-vector product

$$\mathcal{Z}_t = (1-\beta_t)(G(w_t,\xi_t) - g(w_t)) - \beta_t(\nabla G(w_t,\xi_t) - \nabla g(w_t))(w_{t-1} - w_t).$$

By Assumption 5.3 and fact 1 under Assumption 2.3, the Hessian-vector product noise evaluated along the update direction $w_{t-1} - w_t$ is sub-Gaussian with a variance proxy bounded by $\gamma^2\|w_t - w_{t-1}\|^2 \leq \gamma^2\eta_{t-1}^2 G^2$. Summing the proxies and using our previous argument of fact 2, we ascertain the uniform predictable variance proxy for $\mathcal{Z}_t$ as

$$\Sigma_t^2 = 2(1-\beta_t)^2\sigma^2 + 2\beta_t^2\gamma^2\eta_{t-1}^2 G^2.$$

Substitution into Theorem 4.2 completes the proof. □

In the following subsections, we instantiate the unified estimation error bound for each of the three families under specific parameter regimes. We provide detailed derivations and highlight the implications of different choices of $\{p_t, \beta_t, \eta_t\}$ on the error bounds. For the case 1 of each family, we can drop the $\sqrt{2}$ factor in the last term of the bound in Theorem 4.2 by the finer analysis in Remark B.5 on the tightened upper Bound for case 1 with $p_t = 0$.

For the case 2 and case 3 of Family 1, because variance proxy remains zero when $\beta_t = 1$, then the last term related to the variance proxy of the upper bound in Theorem 4.2 can be dropped. Also, for the case 2 and case 3 of Family 2 and Family 3, we can drop the $\sqrt{2}$ factor in the last term of the bound in Theorem 4.2 as well, since for these two cases, we always have $\beta_t = 1$ and therefore the additional factor 2 introduced in bounding the variance proxy in Fact 2 is not needed.

C.2 Zeroth-Order Correction (Standard Recursion and Resetting)

Recall that for the zeroth-order family, the correction map is defined as $\mathcal{T}_t(w_t, w_{t-1}, \xi_t) := \beta_t\big(G(w_{t-1},\xi_t) - G(w_t,\xi_t)\big)$.

C.2.1 The case where $p_t = 0$, $\beta_t = \beta$, and $\eta_t = \eta$ deterministically

In this regime, the estimator is initialized with a large batch sample $G(w_0, \xi_0^B)$ and the recursion proceeds deterministically as

$$v_t = (1-\beta)G(w_t,\xi_t) + \beta v_{t-1}. \tag{36}$$

The absence of stochastic resets implies the execution consists of a single continuous epoch where $m = 1$, $\tau_0 = 0$, and $\tau_1 = T$. Applying the constant parameters, the accumulated variance proxy evaluates to

$$\mathfrak{V}_t^2 = \sum_{j=1}^{t} A_j^2\Sigma_j^2 = \sum_{j=1}^{t}\beta^{-2j}(1-\beta)^2\sigma^2 = (1-\beta)^2\sigma^2\frac{\beta^{-2t}-1}{1-\beta^2}. \tag{37}$$

We set the deterministic variance budget V_t to this exact bound. Utilizing the multiplier sequence $\Lambda_{t,\tau_0} = A_t^{-1} = \beta^t$, the cumulative deterministic bias is bounded by

$$\sum_{j=1}^{t}\Lambda_{t,j}\mathcal{B}_j = \sum_{j=1}^{t} LG\beta\eta\beta^{t-j} = LG\beta\eta\frac{1-\beta^t}{1-\beta} \leq \frac{LG\eta}{1-\beta}. \tag{38}$$

Simultaneously, by letting $V_t = \mathfrak{V}_t^2$, the scaled deviation radius simplifies to

$$A_t^{-1}\sqrt{V_t} = \beta^t(1-\beta)\sigma\sqrt{\frac{\beta^{-2t}-1}{1-\beta^2}} = \sigma\sqrt{\frac{(1-\beta)(1-\beta^{2t})}{1+\beta}} \leq \sigma\sqrt{\frac{1-\beta}{1+\beta}} \leq \sigma\sqrt{1-\beta}. \tag{39}$$

Therefore, applying Theorem 4.2, we establish that with probability at least $1-\delta$, the estimation error satisfies

$$\|e_t\|_{\mathcal{G}} \leq \beta^t C\left(\frac{\delta}{2T},\kappa\right)\sigma/\sqrt{B} + \frac{GL\eta}{1-\beta} + C\left(\frac{\delta}{2T},\kappa\right)\sigma\sqrt{1-\beta}. \tag{40}$$

C.2.2 The case where $p_t = p$, $\beta_t = 1$, and $\eta_t = \eta$ deterministically

Under a probabilistic reset mechanism with full momentum, the estimator update takes the form

$$v_t = \begin{cases} G(w_t, \xi_t^B), & \text{w.p. } p, \\ v_{t-1}, & \text{w.p. } 1-p. \end{cases} \tag{41}$$

The trajectory is partitioned into a random number of epochs, where the temporal length of each epoch is distributed geometrically with success probability p. Setting $\beta_t = 1$ enforces $A_j = 1$ and $\Lambda_{t,j} = 1$ universally. Furthermore, the variance proxy evaluates to $\mathfrak{V}_t^2 = 0$, as the direct stochastic noise is fully suppressed by the deterministic momentum weight. Assigning $V_t = \mathfrak{V}_t^2 = 0$, the cumulative bias inside the random epoch mapping to time t becomes the sole source of structural error:

$$\sum_{j=\tau_{m-1}+1}^{t} \Lambda_{t,j}\mathcal{B}_j = \sum_{j=\tau_{m-1}+1}^{t} LG\eta = LG\eta(t - \tau_{m-1}) \leq LG\eta(\tau_m - \tau_{m-1}). \tag{42}$$

Let $E_m \coloneqq \tau_m - \tau_{m-1}$ for all $m \in [T]$ denote the duration of the m-th epoch. The maximum epoch length across the horizon is controlled via a standard union bound over the geometric tails

$$\mathbb{P}\left(\max_{m\in[T]} E_m \geq \mathfrak{E}_{max}\right) \leq \sum_{m=1}^{T} \mathbb{P}\left(E_m \geq \mathfrak{E}_{max}\right) = T(1-p)^{\mathfrak{E}_{max}}. \tag{43}$$

Enforcing the strict inequality $\log(1/(1-p)) > p$, we constrain the failure probability to $\delta/4$, which yields $\mathfrak{E}_{max} \leq \frac{1}{p}\log\frac{4T}{\delta}$. Integrating this maximum accumulation span, we deduce that with probability at least $1-\delta$, the estimation error respects the bound

$$\|e_t\|_{\mathcal{G}} \leq C\left(\frac{\delta}{4T}, \kappa\right)\sigma/\sqrt{B} + \frac{GL\eta}{p}\log\frac{4T}{\delta}. \tag{44}$$

C.2.3 The case where $p_t = \mathbb{I}_{\{t \mod E = 0\}}$, $\beta_t = 1$, and $\eta_t = \eta$ deterministically

Enforcing a deterministic periodic reset schedule restrains the maximum epoch length to E iterations. The estimator update evaluates as

$$v_t = \begin{cases} G(w_t, \xi_t^B), & t \mod E = 0, \\ v_{t-1}, & \text{otherwise.} \end{cases} \tag{45}$$

Assuming the total horizon factors as $T = E \times N$ for some positive integer N, the algorithm generates $N = T/E$ deterministic epochs. Because the variance proxy $\mathfrak{V}_t^2$ remains zero, by letting $V_t = \mathfrak{V}_t^2 = 0$, the error is bounded entirely by the initialization noise and the maximum deterministic drift over the fixed interval E. Because the total number of reset nodes is reduced from T to T/E, the union bound covering the reset initializations requires a refined confidence adjustment of $\delta/(T/E) = \delta E/T$. Consequently, with probability at least $1-\delta$, the estimation error satisfies

$$\|e_t\|_{\mathcal{G}} \leq C\left(\frac{\delta}{2T/E}, \kappa\right)\sigma/\sqrt{B} + GL\eta E. \tag{46}$$

C.3 First-Order Correction (Differential Recursion and Resetting)

Recall that for this family of estimators, the correction term is set to zero, yielding $\mathcal{T}_t(w_t, w_{t-1}, \xi_t) \coloneqq 0$. The effective innovation, therefore, consists exclusively of the direct noise and the difference noise, expressed as

$$\mathcal{I}_t(\mathcal{T}) = (1-\beta_t)(G(w_t, \xi_t) - g(w_t)) + \beta_t\Delta(w_t, w_{t-1}, \xi_t). \tag{47}$$

C.3.1 The case where $p_t = 0$, $\beta_t = \beta$, and $\eta_t = \eta$ deterministically

In this regime, the estimator initializes with a large batch sample $G(w_0, \xi_0^B)$ and processes deterministically without stochastic resets:

$$v_t = G(w_t, \xi_t) + \beta(v_{t-1} - G(w_{t-1}, \xi_t)). \tag{48}$$

Analogous to the zeroth-order family, the execution consists of a single continuous epoch where $m = 1$, $\tau_0 = 0$, and $\tau_1 = T$. Applying the constant parameters to the derived variance proxy, the accumulated proxy scales according to the predefined multipliers $A_j = \beta^{-j}$. The total variance proxy is evaluated as

$$\mathfrak{V}_t^2 = \sum_{j=1}^{t} A_j^2 \Sigma_j^2 = \sum_{j=1}^{t} \beta^{-2j} \left(2(1-\beta)^2\sigma^2 + 2\beta^2\ell^2\eta^2 G^2\right). \tag{49}$$

Evaluating the geometric series, we define the deterministic variance budget V_t by letting $V_t = \mathfrak{V}_t^2$

$$V_t = \left(2(1-\beta)^2\sigma^2 + 2\beta^2\ell^2\eta^2 G^2\right) \frac{\beta^{-2t} - 1}{1 - \beta^2}. \tag{50}$$

Because the systematic bias $\mathcal{B}_j$ is zero, the cumulative bias component in the overarching theorem vanishes entirely, yielding $\sum_{j=1}^{t} \Lambda_{t,j}\mathcal{B}_j = 0$. The concentration term governed by the martingale difference relies on the modulated variance budget, which simplifies to

$$A_t^{-1}\sqrt{V_t} = \beta^t\sqrt{V_t} = \sqrt{\frac{1-\beta^{2t}}{1-\beta^2}}\sqrt{2(1-\beta)^2\sigma^2 + 2\beta^2\ell^2\eta^2G^2} \leq \sqrt{\frac{2(1-\beta)^2\sigma^2 + 2\beta^2\ell^2\eta^2G^2}{1-\beta^2}}. \tag{51}$$

Invoking Theorem 4.2 we assert that with probability at least $1 - \delta$, the estimation error satisfies

$$\|e_t\|_{\mathcal{G}} \leq \beta^t C\left(\frac{\delta}{2T}, \kappa\right)\sigma/\sqrt{B} + C\left(\frac{\delta}{2T}, \kappa\right)\sqrt{\frac{2(1-\beta)^2\sigma^2 + 2\beta^2\ell^2\eta^2G^2}{1-\beta^2}}. \tag{52}$$

C.3.2 The case where $p_t = p$, $\beta_t = 1$, and $\eta_t = \eta$ deterministically

By assigning full weight to the differential update via $\beta_t = 1$, the estimator replicates the behavior of the PAGE algorithm, resetting memory probabilistically with rate p:

$$v_t := \begin{cases} G(w_t, \xi_t^B), & w.p. \quad p, \\ v_{t-1} + G(w_t, \xi_t) - G(w_{t-1}, \xi_t), & w.p. \quad 1-p, \end{cases} \tag{53}$$

The explicit choice of full momentum inherently nullifies the direct noise variance contribution, leaving the variance proxy dependent on the geometric continuity of the gradient space. Substituting $\beta_t = 1$ yields a constant variance proxy uniformly bounded by $\Sigma_j^2 \leq 2\ell^2\eta^2G^2$. The corresponding multipliers resolve to unity ($A_j = 1$ and $\Lambda_{t,j} = 1$), dictating that the variance accumulation grows linearly with the duration of the random epoch $E_m := \tau_m - \tau_{m-1}$. The accumulated variance proxy inside the epoch thus evaluates to

$$\mathfrak{V}_t^2 = \sum_{j=\tau_{m-1}+1}^{t} 2\ell^2\eta^2G^2 = 2\ell^2\eta^2G^2(t - \tau_{m-1}) \leq 2\ell^2\eta^2G^2E_m. \tag{54}$$

As established previously, the maximal epoch length $\mathfrak{E}_{max}$ governed by the underlying geometric distribution satisfies $\mathfrak{E}_{max} \leq \frac{1}{p}\log\frac{4T}{\delta}$ with high probability. Setting the deterministic variance budget V_t to $2\ell^2\eta^2G^2\frac{1}{p}\log\frac{4T}{\delta}$, and recalling the zero-bias condition $\mathcal{B}_t = 0$, the structure of the error bound eliminates the linear growth associated with deterministic drift. For this configuration, with probability at least $1 - \delta$, the terminal estimation error is subsequently bounded by

$$\|e_t\|_{\mathcal{G}} \leq C\left(\frac{\delta}{4T}, \kappa\right)\sigma/\sqrt{B} + C\left(\frac{\delta}{4T}, \kappa\right)\ell\eta G\sqrt{\frac{2}{p}\log\frac{4T}{\delta}}. \tag{55}$$

C.3.3 The case where $p_t = \mathbb{I}_{\{t \mod E = 0\}}$, $\beta_t = 1$, and $\eta_t = \eta$ deterministically

Under a cyclic deterministic reset mechanism, the temporal length of each estimation epoch is constrained to E iterations:

$$v_t := \begin{cases} G(w_t, \xi_t^B), & t \mod E = 0, \\ v_{t-1} + G(w_t, \xi_t) - G(w_{t-1}, \xi_t), & \text{otherwise}. \end{cases} \tag{56}$$

Assuming the horizon $T = E \times N$ aligns perfectly with these periodic boundaries, the stochastic behavior is constrained similarly to the probabilistic reset case, yet enforces deterministic termination bounds for each tracking segment. Given $\beta_t = 1$, the variance proxy bound simplifies to $\Sigma_j^2 \leq 2\ell^2\eta^2G^2$. Because the length of the interval $t - \tau_{m-1}$ is deterministically upper-bounded by E, the variance budget across any arbitrary epoch deterministically accumulates to

$$\mathfrak{V}_t^2 = \sum_{j=\tau_{m-1}+1}^{t} 2\ell^2\eta^2G^2 \leq 2\ell^2\eta^2G^2E. \tag{57}$$

Setting the strict variance budget $V_t = 2\ell^2\eta^2G^2E$ and combining this with the intrinsic zero-bias property characterizing this estimator family, we embed these components into the unified framework. Consequently, incorporating the deterministic epoch structure and applying the appropriate uniform bounds over the respective intervals, with probability at least $1 - \delta$, the estimation error satisfies

$$\|e_t\|_{\mathcal{G}} \leq C\left(\frac{\delta}{2T/E}, \kappa\right)\sigma/\sqrt{B} + C\left(\frac{\delta}{2T}, \kappa\right)\ell\eta G\sqrt{2E}. \tag{58}$$

C.4 Second-Order Correction (Second-Order Recursion and Resetting)

Recall the explicit formulation of the second-order correction mechanism designed to correct the drift using local curvature, given by $\mathcal{T}_t(w_t, w_{t-1}, \xi_t) \coloneqq \beta_t\Big(G(w_{t-1}, \xi_t) - \big[G(w_t, \xi_t) + \nabla G(w_t, \xi_t)(w_{t-1} - w_t)\big]\Big)$. Substituting this correction into the general effective innovation recursion, we obtain

$$\mathcal{I}_t(\mathcal{T}) = (1-\beta_t)(G(w_t, \xi_t) - g(w_t)) + \beta_t\big(g(w_{t-1}) - g(w_t) - \nabla G(w_t, \xi_t)(w_{t-1} - w_t)\big). \tag{59}$$

C.4.1 The case where $p_t = 0$, $\beta_t = \beta$, and $\eta_t = \eta$ deterministically

Under a purely recursive update, absent random resets, the process evolves across a single continuous epoch:

$$v_t := (1-\beta_t)G(w_t, \xi_t) + \beta_t\Big(v_{t-1} - \nabla G(w_t, \xi_t)(w_{t-1} - w_t)\Big). \tag{60}$$

Imposing the fixed hyperparameters onto the variance framework, we establish the accumulated variance proxy utilizing the geometric sequence $A_j = \beta^{-j}$. The proxy integral evaluates to

$$\mathfrak{V}_t^2 = \sum_{j=1}^{t} A_j^2\Sigma_j^2 = \left(2(1-\beta)^2\sigma^2 + 2\beta^2\gamma^2\eta^2G^2\right)\frac{\beta^{-2t} - 1}{1 - \beta^2}. \tag{61}$$

Setting the deterministic variance budget $V_t = \mathfrak{V}_t^2$ to this quantity, we concurrently evaluate the cumulative deterministic bias induced by the Taylor remainder. The bias accumulates through the discounted sum defined by $\Lambda_{t,j} = \beta^{t-j}$, leading to

$$\sum_{j=1}^{t}\Lambda_{t,j}\mathcal{B}_j \leq \sum_{j=1}^{t}\beta^{t-j}\frac{\alpha}{2}\beta\eta^2G^2 = \frac{\alpha}{2}\beta\eta^2G^2\frac{1-\beta^t}{1-\beta} \leq \frac{\alpha\beta\eta^2G^2}{2(1-\beta)}. \tag{62}$$

Isolating the modulated variance term within the martingale concentration mechanism yields the upper bound $A_t^{-1}\sqrt{V_t} \leq \sqrt{\frac{2(1-\beta)^2\sigma^2 + 2\beta^2\gamma^2\eta^2G^2}{1-\beta^2}}$. Applying Theorem 4.2 over this singular deterministic epoch, we conclude that with probability at least $1 - \delta$, the estimation error satisfies

$$\|e_t\|_{\mathcal{G}} \leq \beta^t C\left(\frac{\delta}{2T}, \kappa\right)\sigma/\sqrt{B} + \frac{\alpha\beta\eta^2G^2}{2(1-\beta)} + C\left(\frac{\delta}{2T}, \kappa\right)\sqrt{\frac{2(1-\beta)^2\sigma^2 + 2\beta^2\gamma^2\eta^2G^2}{1-\beta^2}}. \tag{63}$$

C.4.2 The case where $p_t = p$, $\beta_t = 1$, and $\eta_t = \eta$ deterministically

Setting the momentum uniformly to 1 transitions the estimator into a differential tracker that leverages stochastic second-order corrections, accompanied by memory resets distributed geometrically with rate p:

$$v_t := \begin{cases} G(w_t, \xi_t^B), & w.p. \quad p, \\ v_{t-1} - \nabla G(w_t, \xi_t)(w_{t-1} - w_t), & w.p. \quad 1-p. \end{cases} \tag{64}$$

Because $\beta_t = 1$, the direct noise term evaluates to zero, coupling the variance proxy to the sub-Gaussian Hessian noise such that $\Sigma_j^2 = 2\gamma^2\eta^2G^2$. Similarly, the per-step deterministic bias stabilizes at $\mathcal{B}_j \leq \frac{\alpha}{2}\eta^2G^2$. Due to $A_j = 1$ and $\Lambda_{t,j} = 1$, the accumulation of both the bias and the variance proxy scales linearly with the duration of the current random epoch $E_m \coloneqq \tau_m - \tau_{m-1}$. Within this epoch, the summations are bounded by

$$\sum_{j=\tau_{m-1}+1}^{t} \mathcal{B}_j \leq \frac{\alpha}{2}\eta^2G^2E_m \quad \text{and} \quad \mathfrak{V}_t^2 = \sum_{j=\tau_{m-1}+1}^{t} \Sigma_j^2 \leq 2\gamma^2\eta^2G^2E_m. \tag{65}$$

Invoking the high-probability uniform bound on the maximum epoch length, we verify $\mathfrak{E}_{max} \leq \frac{1}{p}\log\frac{4T}{\delta}$. By defining the deterministic variance budget as $V_t = 2\gamma^2\eta^2G^2\frac{1}{p}\log\frac{4T}{\delta}$, and bounding the total bias over the longest possible epoch, we assemble the complete error topology. Consequently, with probability at least $1-\delta$, the terminal estimation error is bounded by

$$\|e_t\|_{\mathcal{G}} \leq C\left(\frac{\delta}{4T}, \kappa\right)\sigma/\sqrt{B} + \frac{\alpha\eta^2G^2}{2p}\log\frac{4T}{\delta} + C\left(\frac{\delta}{4T}, \kappa\right)\gamma\eta G\sqrt{\frac{2}{p}\log\frac{4T}{\delta}}. \tag{66}$$

C.4.3 The case where $p_t = \mathbb{I}_{\{t \mod E=0\}}$, $\beta_t = 1$, and $\eta_t = \eta$ deterministically

Imposing a deterministic, periodic reset schedule ensures that memory truncation occurs every E iterations:

$$v_t \coloneqq \begin{cases} G(w_t, \xi_t^B), & t \mod E = 0, \\ v_{t-1} - \nabla G(w_t, \xi_t)(w_{t-1} - w_t), & \text{otherwise.} \end{cases} \tag{67}$$

Assuming the computational horizon is an integer multiple of the epoch length, $T = E \times N$, this architecture completely arrests the unbounded accumulation of estimation lag while preserving the full corrective weight $\beta_t = 1$. The fundamental quantities parallel the probabilistic setting, maintaining $\Sigma_j^2 = 2\gamma^2\eta^2G^2$ and $\mathcal{B}_j \leq \frac{\alpha}{2}\eta^2G^2$. However, the interval lengths are now deterministically capped at $t - \tau_{m-1} \leq E$. The cumulative proxy and bias are correspondingly bounded across the fixed interval

$$\sum_{j=\tau_{m-1}+1}^{t} \mathcal{B}_j \leq \frac{\alpha}{2}\eta^2G^2E \quad \text{and} \quad \mathfrak{V}_t^2 \leq 2\gamma^2\eta^2G^2E. \tag{68}$$

Assigning this upper bound to V_t and directly incorporating the maximal deterministic bias contribution, the bounds concentrate without the logarithmic geometric penalties. We establish that with probability at least $1-\delta$, the resulting error magnitude honors the bound

$$\|e_t\|_{\mathcal{G}} \leq C\left(\frac{\delta}{2T/E}, \kappa\right)\sigma/\sqrt{B} + \frac{\alpha\eta^2G^2E}{2} + C\left(\frac{\delta}{2T}, \kappa\right)\gamma\eta G\sqrt{2E}. \tag{69}$$

C.5 Examples for Assumption 5.2 of Sub-Gaussian Lipschitzness of the Centered Differences

The sub-Gaussian tail assumptions utilized in our analysis are automatically satisfied in standard Empirical Risk Minimization (ERM) configurations. Let the function (or the gradient) be defined over a finite dataset as $g(w) = \frac{1}{n}\sum_{i=1}^{n} G_i(w)$, where the stochastic oracle samples a component index uniformly at random. If every individual component G_i is L_i-Lipschitz continuous, the discrete nature of the sample space guarantees that the centered difference noise is bounded almost surely by $2\max_i L_i\|w - w'\|$. Because any bounded random variable is sub-Gaussian, Assumption 5.2 is globally satisfied with the variance proxy parameter $\ell = 2\max_i L_i$. Nonetheless, we present concrete instantiations satisfying Assumption 5.2 next.

We leverage the moment-generating function of a scaled chi-squared distribution: for a standard Gaussian random variable $Z \sim \mathcal{N}(0,1)$, the expectation $\mathbb{E}[\exp(\lambda Z^2)] = (1-2\lambda)^{-1/2}$ is valid for all $\lambda < 1/2$. Enforcing the condition $(1-2\lambda)^{-1/2} \le 2$ necessitates $\lambda \le \frac{3}{8}$.

Recall that Assumption 5.2 requires the centered estimate difference noise $\Delta(w, w', \xi) := (G(w,\xi) - G(w',\xi)) - (g(w) - g(w'))$ to concentrate with a variance proxy bounded by $\ell^2\|w - w'\|^2$.

Example 1: Scalar Function Configuration. Let the primal space $\mathcal{X} := \mathbb{R}^d$ be equipped with the Euclidean geometry $\|\cdot\| := \|\cdot\|_2$. Consider a stochastic linear observation model where the oracle evaluates a scalar function $G : \mathbb{R}^d \times \Omega \to \mathcal{G} := \mathbb{R}$ with $\|\cdot\|_{\mathcal{G}} := |\cdot|$ defined by

$$G(w, \xi) = \langle \xi, w \rangle, \tag{70}$$

where $\xi \sim \mathcal{N}(0, \Sigma)$ is a centered Gaussian random vector. The population function is zero, $g(w) = \mathbb{E}[G(w,\xi)] = 0$. The centered difference evaluates to the inner product

$$\Delta(w, w', \xi) = \langle \xi, w - w' \rangle. \tag{71}$$

For any fixed deterministic w and w', the random variable Δ is a univariate Gaussian distributed as $\mathcal{N}(0, (w-w')^\top \Sigma (w-w'))$. The variance is bounded by the spectral norm, $\nu^2 := (w-w')^\top \Sigma (w - w') \le \|\Sigma\|_{\mathrm{op}}\|w - w'\|_2^2$. Letting $Z = \Delta/\nu \sim \mathcal{N}(0,1)$, we verify the condition

$$\mathbb{E}\left[\exp\left(\frac{|\Delta(w,w',\xi)|^2}{\ell^2\|w-w'\|_2^2}\right)\right] \le \mathbb{E}\left[\exp\left(\frac{\|\Sigma\|_{\mathrm{op}}}{\ell^2} Z^2\right)\right]. \tag{72}$$

By setting $\lambda = \|\Sigma\|_{\mathrm{op}}/\ell^2$, the sub-Gaussian Lipschitz condition is satisfied for any structural constant satisfying $\ell \ge \sqrt{8\|\Sigma\|_{\mathrm{op}}/3}$.

Example 2: Vector Field Configuration (Gradient of a Potential). Consider a stochastic quadratic objective $F : \mathbb{R}^d \times \Omega \to \mathbb{R}$ with the primal space $(\mathcal{X}, \|\cdot\|) := (\mathbb{R}^d, \|\cdot\|_2)$ defined by

$$F(w, \xi) = \frac{1}{2} w^\top (A + \xi I) w, \tag{73}$$

where $A \in \mathbb{R}^{d\times d}$ is a deterministic symmetric matrix, and $\xi \sim \mathcal{N}(0, \sigma^2)$ is a scalar Gaussian noise perturbing the curvature. The stochastic vector field is the gradient $G(w,\xi) = \nabla_w F(w,\xi) = (A + \xi I)w$, implicitly setting $\mathcal{G} := \mathbb{R}^d$ and $\|\cdot\|_{\mathcal{G}} := \|\cdot\|_2$, making the population gradient $g(w) = Aw$. The centered estimate difference noise simplifies algebraically to

$$\Delta(w, w', \xi) = \xi(w - w'). \tag{74}$$

Taking the squared dual norm (which coincides with the ℓ_2 norm), we obtain $\|\Delta(w,w',\xi)\|_2^2 = \xi^2\|w - w'\|_2^2$. The exponential moment evaluates as

$$\mathbb{E}\left[\exp\left(\frac{\|\Delta(w,w',\xi)\|_2^2}{\ell^2\|w-w'\|_2^2}\right)\right] = \mathbb{E}\left[\exp\left(\frac{\sigma^2}{\ell^2}\left(\frac{\xi}{\sigma}\right)^2\right)\right]. \tag{75}$$

Recognizing $(\xi/\sigma) \sim \mathcal{N}(0,1)$, the condition holds globally provided $\ell \ge \sigma\sqrt{8/3}$.

D Lemmas in Proofs for Two Applications

Lemma D.1 (Bound for Proximal Gradient Mapping with Normalized Direction, see similar Lemma 2 in [31]). Suppose $w \in \mathcal{W} \subseteq \mathcal{X}$, $u, \delta \in \mathcal{X}^*$, $\eta > 0$, and $\|u\|_* \neq 0, \|u+\delta\|_* \neq 0$. Then for the proximal gradient mapping P defined in Definition 2.2, we have:

$$\left\langle u, P\left(w, \frac{u+\delta}{\|u+\delta\|_*}, \eta\right)\right\rangle \geq \frac{1}{4}\left\langle u, P\left(w, \frac{u}{\|u\|_*}, \frac{\eta}{2}\right)\right\rangle - 2\|\delta\|_*$$

Proof of Lemma D.1. Noting the scaling property of the proximal gradient mapping (Lemma A.7) such that $P(w, \alpha u, \eta) = \alpha P(w, u, \alpha\eta), \forall \alpha > 0$, we have:

$$P\left(w, \frac{u+\delta}{\|u+\delta\|_*}, \eta\right) = \frac{P(w, u+\delta, \eta/\|u+\delta\|_*)}{\|u+\delta\|_*}, \quad P\left(w, \frac{u}{\|u\|_*}, \frac{\eta}{2}\right) = \frac{P(w, u, \eta/(2\|u\|_*))}{\|u\|_*}$$

Therefore, the desired inequality to be proved is equivalent to:

$$-\frac{\langle u, P(w, u+\delta, \eta/\|u+\delta\|_*)\rangle}{\|u+\delta\|_*} \leq -\frac{\langle u, P(w, u, \eta/(2\|u\|_*))\rangle}{4\|u\|_*} + 2\|\delta\|_*$$

Now we start to prove the above equivalent inequality, then the desired inequality will follow by noting the scaling property of the proximal gradient mapping. We consider the following two cases for the ease of analysis: (1) $\|\delta\|_* \leq \frac{1}{3}\|u\|_*$; (2) $\|\delta\|_* > \frac{1}{3}\|u\|_*$.

Case (1): $\|\delta\|_* \leq \frac{1}{3}\|u\|_*$.

Applying the properties of the proximal gradient mapping (Lemma A.7) such that $\|P(w, u_1, \eta) - P(w, u_2, \eta)\| \leq \|u_1 - u_2\|_*$, and $\langle \mathbf{x}, \mathbf{y}\rangle \leq \|\mathbf{x}\|_*\|\mathbf{y}\|$ for any $\mathbf{x} \in \mathcal{X}^*, \mathbf{y} \in \mathcal{X}$, we have:

$$-\langle u, P(w, u+\delta, \eta/\|u+\delta\|_*)\rangle \leq -\langle u, P(w, u, \eta/\|u+\delta\|_*)\rangle + \|u\|_*\|\delta\|_*$$

By using the condition $\|\delta\|_* \leq \frac{1}{3}\|u\|_*$ and the triangle inequality, we have:

$$\frac{2}{3}\|u\|_* \leq \|u\|_* - \|\delta\|_* \leq \|u+\delta\|_* \leq \|u\|_* + \|\delta\|_* \leq \frac{4}{3}\|u\|_*$$

Noting the monotonicity of the proximal gradient mapping (Lemma A.7) $\langle u, \eta' P(w, u, \eta')\rangle \geq \langle u, \eta P(w, u, \eta)\rangle \geq 0$ for any $\eta' \geq \eta > 0$, and using the condition $\|\delta\|_* \leq \frac{1}{3}\|u\|_*$:

$$\langle u, P(w, u, \eta/\|u+\delta\|_*)\rangle \geq \frac{\|u+\delta\|_*}{2\|u\|_*}\langle u, P(w, u, \eta/(2\|u\|_*))\rangle \geq \frac{1}{3}\langle u, P(w, u, \eta/(2\|u\|_*))\rangle$$

By combining the above three inequalities, we have the desired inequality for Case (1): $\|\delta\|_* \leq \frac{1}{3}\|u\|_*$.

$$\left[-\frac{\langle u, P(w, u+\delta, \eta/\|u+\delta\|_*)\rangle}{\|u+\delta\|_*}\right]\mathbb{1}_{\|\delta\|_* \leq \frac{1}{3}\|u\|_*} \leq \left[-\frac{\langle u, P(w, u, \eta/(2\|u\|_*))\rangle}{4\|u\|_*} + 2\|\delta\|_*\right]\mathbb{1}_{\|\delta\|_* \leq \frac{1}{3}\|u\|_*}$$

Case (2): $\|\delta\|_* > \frac{1}{3}\|u\|_*$.

Applying the properties of the proximal gradient mapping (Lemma A.7) such that $\langle u, P(w, u, \eta)\rangle \geq \|P(w, u, \eta)\|^2 \geq 0$, $\|P(w, u, \eta)\| \leq \|u\|_*$, and using $\langle \mathbf{x}, \mathbf{y}\rangle \leq \|\mathbf{x}\|_*\|\mathbf{y}\|$ for any $\mathbf{x} \in \mathcal{X}^*, \mathbf{y} \in \mathcal{X}$:

$$\begin{aligned} -\frac{\langle u, P(w, u+\delta, \eta/\|u+\delta\|_*)\rangle}{\|u+\delta\|_*} &= -\frac{\langle u+\delta, P(w, u+\delta, \frac{\eta}{\|u+\delta\|_*})\rangle}{\|u+\delta\|_*} + \frac{\langle \delta, P(w, u+\delta, \frac{\eta}{\|u+\delta\|_*})\rangle}{\|u+\delta\|_*} \\ &\leq 0 + \frac{\|\delta\|_*\|P(w, u+\delta, \frac{\eta}{\|u+\delta\|_*})\|}{\|u+\delta\|_*} \leq \|\delta\|_* \end{aligned}$$

Given that condition $\|\delta\|_* > \frac{1}{3}\|u\|_*$, and using the properties of the proximal gradient mapping (Lemma A.7) such that $\langle u, P(w, u, \eta)\rangle \leq \|u\|_*\|P(w, u, \eta)\|, \|P(w, u, \eta)\| \leq \|u\|_*$, we have:

$$\frac{\langle u, P(w, u, \eta/(2\|u\|_*))\rangle}{4\|u\|_*} \leq \frac{1}{4}\|P(w, u, \eta/(2\|u\|_*))\| \leq \frac{1}{4}\|u\|_* \leq \frac{1}{3}\|u\|_* < \|\delta\|_*$$

Combining the above two inequalities, we have the desired inequality for Case (2): $\|\delta\|_* > \frac{1}{3}\|u\|_*$.

$$\left[-\frac{\langle u, P(w, u+\delta, \eta/\|u+\delta\|_*)\rangle}{\|u+\delta\|_*}\right]\mathbb{1}_{\|\delta\|_*>\frac{1}{3}\|u\|_*} \leq \left[-\frac{\langle u, P(w, u, \eta/(2\|u\|_*))\rangle}{4\|u\|_*} + 2\|\delta\|_*\right]\mathbb{1}_{\|\delta\|_*>\frac{1}{3}\|u\|_*}$$

Summing up the two inequalities for these **two cases** and noting that $1 = \mathbb{1}_{\|\delta\|_* \leq \frac{1}{3}\|u\|_*} + \mathbb{1}_{\|\delta\|_*>\frac{1}{3}\|u\|_*}$, we complete the proof of the equivalent inequality, and the desired inequality follows. □

Lemma D.2 (Optimal Decoupled Bounds with Selection of Step Size and Momentum Parameter)**.** For constants $C_1, \ldots, C_6 \geq 0$, and $C_2C_4C_5C_6 \neq 0$, and horizon $T \geq 1$, define the objective functions:

$$f_1(\eta, 1-\beta) = \frac{C_1}{\eta T} + C_2\eta + \frac{C_3}{(1-\beta)T} + C_4\sqrt{1-\beta} + \frac{C_5\eta}{1-\beta}$$
$$f_2(\eta, 1-\beta) = \frac{C_1}{\eta T} + C_2\eta + \frac{C_3}{(1-\beta)T} + C_4\sqrt{1-\beta} + \frac{C_5\eta}{\sqrt{1-\beta}}$$
$$f_3(\eta, 1-\beta) = f_2(\eta, 1-\beta) + \frac{C_6\eta^2}{1-\beta}$$

Letting $\eta = \min\left\{\left(\frac{C_1}{C_2}\right)^{\frac{1}{2}} T^{-\frac{1}{2}},\ \left(\frac{C_1^3}{C_4^2C_5}\right)^{\frac{1}{4}} T^{-\frac{3}{4}}\right\}, 1-\beta = \max\left\{\left(\frac{C_3}{C_4}\right)^{\frac{2}{3}} T^{-\frac{2}{3}},\ \left(\frac{C_1C_5}{C_4^2}\right)^{\frac{1}{2}} T^{-\frac{1}{2}}\right\}$

$$f_1(\eta, 1-\beta) \leq 2\left(\frac{C_1C_2}{T}\right)^{\frac{1}{2}} + 2\left(\frac{C_3C_4^2}{T}\right)^{\frac{1}{3}} + 3\left(\frac{C_1C_4^2C_5}{T}\right)^{\frac{1}{4}} \textbf{ (Family 1)}$$

when $T \geq \max\left\{\frac{C_3}{C_4}, \frac{C_1C_5}{C_4^2}\right\}$.

Letting $\eta = \min\left\{\left(\frac{C_1}{C_2}\right)^{\frac{1}{2}} T^{-\frac{1}{2}},\ \left(\frac{C_1^2}{C_4C_5}\right)^{\frac{1}{3}} T^{-\frac{2}{3}}\right\}, 1-\beta = \max\left\{\left(\frac{C_3}{C_4}\right)^{\frac{2}{3}} T^{-\frac{2}{3}},\ \left(\frac{C_1C_5}{C_4^2}\right)^{\frac{2}{3}} T^{-\frac{2}{3}}\right\}$

$$f_2(\eta, 1-\beta) \leq 2\left(\frac{C_1C_2}{T}\right)^{\frac{1}{2}} + 2\left(\frac{C_3C_4^2}{T}\right)^{\frac{1}{3}} + 3\left(\frac{C_1C_4C_5}{T}\right)^{\frac{1}{3}} \textbf{ (Family 2)}$$

when $T \geq \max\left\{\frac{C_3}{C_4}, \frac{C_1C_5}{C_4^2}\right\}$.

Letting $\eta = \min\left\{\left(\frac{C_1}{C_2}\right)^{\frac{1}{2}} T^{-\frac{1}{2}},\ \left(\frac{C_1^2}{C_4C_5}\right)^{\frac{1}{3}} T^{-\frac{2}{3}},\ \left(\frac{C_1^3}{C_4^2C_6}\right)^{\frac{1}{5}} T^{-\frac{3}{5}}\right\}$,
$1-\beta = \max\left\{\left(\frac{C_3}{C_4}\right)^{\frac{2}{3}} T^{-\frac{2}{3}},\ \left(\frac{C_1C_5}{C_4^2}\right)^{\frac{2}{3}} T^{-\frac{2}{3}},\ \left(\frac{C_1^2C_6}{C_4^3}\right)^{\frac{2}{5}} T^{-\frac{4}{5}}\right\}$

$$f_3(\eta, 1-\beta) \leq 2\left(\frac{C_1C_2}{T}\right)^{\frac{1}{2}} + 2\left(\frac{C_3C_4^2}{T}\right)^{\frac{1}{3}} + 3\left(\frac{C_1C_4C_5}{T}\right)^{\frac{1}{3}} + 3\left(\frac{C_1^2C_4^2C_6}{T^2}\right)^{\frac{1}{5}} \textbf{ (Family 3)}$$

when $T \geq \max\left\{\frac{C_3}{C_4}, \frac{C_1C_5}{C_4^2}, \left(\frac{C_1^2C_6}{C_4^3}\right)^{\frac{1}{2}}\right\}$.

Proof. To simplify the analysis, we introduce $y = 1-\beta$. We denote the fundamental structural limits present in the bounds as:

$$L_1 = \left(\frac{C_1C_2}{T}\right)^{\frac{1}{2}}, L_2 = \left(\frac{C_3C_4^2}{T}\right)^{\frac{1}{3}}, L_3 = \left(\frac{C_1C_4^2C_5}{T}\right)^{\frac{1}{4}}, L_4 = \left(\frac{C_1C_4C_5}{T}\right)^{\frac{1}{3}}, L_5 = \left(\frac{C_1^2C_4^2C_6}{T^2}\right)^{\frac{1}{5}}$$

We define the unscaled parameter candidate blocks present in the Lemma selections as follows:

$$\eta_A = \left(\frac{C_1}{C_2}\right)^{\frac{1}{2}} T^{-\frac{1}{2}},\ \eta_B = \left(\frac{C_1^3}{C_4^2C_5}\right)^{\frac{1}{4}} T^{-\frac{3}{4}},\ \eta_C = \left(\frac{C_1^2}{C_4C_5}\right)^{\frac{1}{3}} T^{-\frac{2}{3}},\ \eta_D = \left(\frac{C_1^3}{C_4^2C_6}\right)^{\frac{1}{5}} T^{-\frac{3}{5}}$$

$$y_A = \left(\frac{C_3}{C_4}\right)^{\frac{2}{3}} T^{-\frac{2}{3}},\ y_B = \left(\frac{C_1C_5}{C_4^2}\right)^{\frac{1}{2}} T^{-\frac{1}{2}},\ y_C = \left(\frac{C_1C_5}{C_4^2}\right)^{\frac{2}{3}} T^{-\frac{2}{3}},\ y_D = \left(\frac{C_1^2C_6}{C_4^3}\right)^{\frac{2}{5}} T^{-\frac{4}{5}}$$

For **Family 1**, substituting $\eta = \min\{\eta_A, \eta_B\}$ and $y = \max\{y_A, y_B\}$ into the objective function and using $\eta \le \eta_A$, $\frac{1}{\eta} \le \frac{1}{\eta_A} + \frac{1}{\eta_B}$, $\sqrt{y} \le \sqrt{y_A} + \sqrt{y_B}$, $\frac{1}{y} \le \frac{1}{y_A}$, $\frac{\eta}{y} \le \frac{\eta_B}{y_B}$ yields the following inequality:

$$\begin{aligned}
f_1(\eta, 1-\beta) &= \frac{C_1}{\eta T} + C_2\eta + \frac{C_3}{yT} + C_4\sqrt{y} + \frac{C_5\eta}{y} \\
&\le \left(\frac{C_1}{\eta_A T} + \frac{C_1}{\eta_B T}\right) + C_2\eta_A + \frac{C_3}{y_A T} + C_4\left(\sqrt{y_A} + \sqrt{y_B}\right) + \frac{C_5\eta_B}{y_B} \\
&= \underbrace{\left(\frac{C_1}{\eta_A T} + C_2\eta_A\right)}_{2L_1} + \underbrace{\left(\frac{C_3}{y_A T} + C_4\sqrt{y_A}\right)}_{2L_2} + \underbrace{\left(\frac{C_1}{\eta_B T} + C_4\sqrt{y_B} + \frac{C_5\eta_B}{y_B}\right)}_{3L_3} \\
&= 2L_1 + 2L_2 + 3L_3.
\end{aligned}$$

For **Family 2**, substituting $\eta = \min\{\eta_A, \eta_C\}$ and $y = \max\{y_A, y_C\}$ and using $\eta \le \eta_A$, $\frac{1}{\eta} \le \frac{1}{\eta_A} + \frac{1}{\eta_C}$, $\sqrt{y} \le \sqrt{y_A} + \sqrt{y_C}$, $\frac{1}{y} \le \frac{1}{y_A}$, $\frac{\eta}{\sqrt{y}} \le \frac{\eta_C}{\sqrt{y_C}}$ yields the following inequality:

$$\begin{aligned}
f_2(\eta, 1-\beta) &= \frac{C_1}{\eta T} + C_2\eta + \frac{C_3}{yT} + C_4\sqrt{y} + \frac{C_5\eta}{\sqrt{y}} \\
&\le \left(\frac{C_1}{\eta_A T} + \frac{C_1}{\eta_C T}\right) + C_2\eta_A + \frac{C_3}{y_A T} + C_4\left(\sqrt{y_A} + \sqrt{y_C}\right) + \frac{C_5\eta_C}{\sqrt{y_C}} \\
&= \underbrace{\left(\frac{C_1}{\eta_A T} + C_2\eta_A\right)}_{2L_1} + \underbrace{\left(\frac{C_3}{y_A T} + C_4\sqrt{y_A}\right)}_{2L_2} + \underbrace{\left(\frac{C_1}{\eta_C T} + C_4\sqrt{y_C} + \frac{C_5\eta_C}{\sqrt{y_C}}\right)}_{3L_4} \\
&= 2L_1 + 2L_2 + 3L_4.
\end{aligned}$$

For **Family 3**, substituting $\eta = \min\{\eta_A, \eta_C, \eta_D\}$ and $y = \max\{y_A, y_C, y_D\}$ and using $\eta \le \eta_A$, $\frac{1}{\eta} \le \frac{1}{\eta_A} + \frac{1}{\eta_C} + \frac{1}{\eta_D}$, $\sqrt{y} \le \sqrt{y_A} + \sqrt{y_C} + \sqrt{y_D}$, $\frac{1}{y} \le \frac{1}{y_A}$, $\frac{\eta}{\sqrt{y}} \le \frac{\eta_C}{\sqrt{y_C}}$, $\frac{\eta^2}{y} \le \frac{\eta_D^2}{y_D}$ extends the previous inequality by incorporating the quadratic term:

$$\begin{aligned}
f_3(\eta, 1-\beta) &= f_2(\eta, 1-\beta) + \frac{C_6\eta^2}{y} \\
&\le 2L_1 + 2L_2 + 3L_4 + \underbrace{\left(\frac{C_1}{\eta_D T} + C_4\sqrt{y_D} + \frac{C_6\eta_D^2}{y_D}\right)}_{3L_5} \\
&= 2L_1 + 2L_2 + 3L_4 + 3L_5.
\end{aligned}$$

This strictly recovers the bounds stated in the Lemma. □

Lemma D.3 (Optimal Bound with Selection of Step Size, Reset Probability and Batch Size)**.** Let $B, E, T \in \mathbb{Z}_+$, $p = \frac{1}{E} \in (0, 1]$, $\eta > 0$. For $C_1, \ldots, C_6 \ge 0$ with $C_1C_2C_4C_5C_6 \ne 0$ and $C_3 = 0$, define the objective functions as follows:

$$\begin{aligned}
f_1(\eta, p)_{C_3=0} &= \frac{C_1}{\eta T} + C_2\eta + 0 + C_4\sqrt{p} + \frac{C_5\eta}{p} \\
f_2(\eta, p)_{C_3=0} &= \frac{C_1}{\eta T} + C_2\eta + 0 + C_4\sqrt{p} + \frac{C_5\eta}{\sqrt{p}} \\
f_3(\eta, p)_{C_3=0} &= f_2(\eta, p) + \frac{C_6\eta^2}{p}
\end{aligned}$$

Letting $\eta = \min\left\{\left(\frac{C_1}{C_2T}\right)^{\frac{1}{2}}, \left(\frac{2C_1^3}{C_4^2C_5T^3}\right)^{\frac{1}{4}}\right\}$ and $p = \frac{1}{E}$ with $B = E = \left\lceil\left(\frac{C_4^2T}{8C_1C_5}\right)^{\frac{1}{2}}\right\rceil$, when $T \ge \frac{8C_1C_5}{C_4^2}$

$$f_1(\eta, p)_{C_3=0} \le 2\left(\frac{C_1C_2}{T}\right)^{\frac{1}{2}} + 4\left(\frac{C_1C_4^2C_5}{2T}\right)^{\frac{1}{4}} \textbf{ (Family 1)}$$

Letting $\eta = \min\left\{\left(\frac{C_1}{C_2 T}\right)^{\frac{1}{2}}, \left(\frac{C_1^2}{\sqrt{2}C_4 C_5 T^2}\right)^{\frac{1}{3}}\right\}$ and $p = \frac{1}{E}$ with $B = E = \left\lceil \left(\frac{C_4^4 T^2}{2C_1^2 C_5^2}\right)^{\frac{1}{3}} \right\rceil$, when $T \geq \frac{\sqrt{2}C_1 C_5}{C_4^2}$

$$f_2(\eta, p)_{C_3=0} \leq 2\left(\frac{C_1 C_2}{T}\right)^{\frac{1}{2}} + 3\left(\frac{\sqrt{2}C_1 C_4 C_5}{T}\right)^{\frac{1}{3}} \textbf{ (Family 2)}$$

Letting $\eta = \min\left\{\left(\frac{C_1}{C_2 T}\right)^{\frac{1}{2}}, \left(\frac{C_1^2}{\sqrt{2}C_4 C_5 T^2}\right)^{\frac{1}{3}}, \left(\frac{C_1^3}{4C_4^2 C_6 T^3}\right)^{\frac{1}{5}}\right\}$ and $p = \frac{1}{E}$ with $B = E = \left\lceil \min\left\{\left(\frac{C_4^4 T^2}{2C_1^2 C_5^2}\right)^{\frac{1}{3}}, \left(\frac{C_4^6 T^4}{16 C_1^4 C_6^2}\right)^{\frac{1}{5}}\right\} \right\rceil$, when $T \geq \max\left\{\frac{\sqrt{2}C_1 C_5}{C_4^2}, 2\left(\frac{C_1^2 C_6}{C_4^3}\right)^{\frac{1}{2}}\right\}$

$$f_3(\eta, p)_{C_3=0} \leq 2\left(\frac{C_1 C_2}{T}\right)^{\frac{1}{2}} + 3\left(\frac{\sqrt{2}C_1 C_4 C_5}{T}\right)^{\frac{1}{3}} + 5\left(\frac{C_1^2 C_4^2 C_6}{8T^2}\right)^{\frac{1}{5}} \textbf{ (Family 3)}$$

Proof. Before establishing the bounds, we define the following variables $\eta_A, \eta_B, \eta_C, \eta_D$, the continuous batch size roots E_B, E_C, E_D, and the AM-GM minimal values M_1, M_2, M_3, M_4:

$$\begin{aligned}
\eta_A &= \left(\frac{C_1}{C_2 T}\right)^{\frac{1}{2}} & & & M_1 &= \left(\frac{C_1 C_2}{T}\right)^{\frac{1}{2}} \\
\eta_B &= \left(\frac{2C_1^3}{C_4^2 C_5 T^3}\right)^{\frac{1}{4}} & E_B &= \left(\frac{C_4^2 T}{8C_1 C_5}\right)^{\frac{1}{2}} & M_2 &= \left(\frac{C_1 C_4^2 C_5}{2T}\right)^{\frac{1}{4}} \\
\eta_C &= \left(\frac{C_1^2}{\sqrt{2}C_4 C_5 T^2}\right)^{\frac{1}{3}} & E_C &= \left(\frac{C_4^4 T^2}{2C_1^2 C_5^2}\right)^{\frac{1}{3}} & M_3 &= \left(\frac{\sqrt{2}C_1 C_4 C_5}{T}\right)^{\frac{1}{3}} \\
\eta_D &= \left(\frac{C_1^3}{4C_4^2 C_6 T^3}\right)^{\frac{1}{5}} & E_D &= \left(\frac{C_4^6 T^4}{16 C_1^4 C_6^2}\right)^{\frac{1}{5}} & M_4 &= \left(\frac{C_1^2 C_4^2 C_6}{8T^2}\right)^{\frac{1}{5}}
\end{aligned}$$

For **Family 1**, substituting $\eta = \min\{\eta_A, \eta_B\}, p = \frac{1}{E} = \frac{1}{\lceil E_B \rceil}$, and noting that $E_B \geq 1$ when $T \geq \frac{8C_1 C_5}{C_4^2}$ therefore $E_B \leq \lceil E_B \rceil < E_B + 1 \leq 2E_B$, and using $\eta \leq \eta_A, \frac{1}{\eta} \leq \frac{1}{\eta_A} + \frac{1}{\eta_B}, \sqrt{p} = \frac{1}{\sqrt{E}} \leq \frac{1}{\sqrt{E_B}}, \frac{\eta}{p} \leq \eta_B \cdot 2E_B$, we have:

$$\begin{aligned}
f_1(\eta, p)_{C_3=0} &= \frac{C_1}{\eta T} + C_2\eta + 0 + C_4\sqrt{p} + \frac{C_5\eta}{p} \\
&\leq \left(\frac{C_1}{\eta_A T} + \frac{C_1}{\eta_B T}\right) + C_2\eta_A + 0 + \frac{C_4}{\sqrt{E_B}} + 2C_5\eta_B E_B \\
&= \underbrace{\left(\frac{C_1}{\eta_A T} + C_2\eta_A\right)}_{2M_1} + \underbrace{\left(\frac{C_1}{\eta_B T} + \frac{C_4/2}{\sqrt{E_B}} + \frac{C_4/2}{\sqrt{E_B}} + 2C_5\eta_B E_B\right)}_{4M_2} \\
&= 2M_1 + 4M_2.
\end{aligned}$$

For **Family 2**, substituting $\eta = \min\{\eta_A, \eta_C\}, p = \frac{1}{E} = \frac{1}{\lceil E_C \rceil}$, and noting that $E_C \geq 1$ when $T \geq \frac{\sqrt{2}C_1 C_5}{C_4^2}$ therefore $E_C \leq \lceil E_C \rceil < E_C + 1 \leq 2E_C$, and using $\eta \leq \eta_A, \frac{1}{\eta} \leq \frac{1}{\eta_A} + \frac{1}{\eta_C}, \sqrt{p} =$

$\frac{1}{\sqrt{E}} \le \frac{1}{\sqrt{E_C}}, \frac{\eta}{\sqrt{p}} \le \eta_C \cdot \sqrt{2E_C}$, we have:

$$\begin{aligned} f_2(\eta, p)_{C_3=0} &= \frac{C_1}{\eta T} + C_2\eta + 0 + C_4\sqrt{p} + \frac{C_5\eta}{\sqrt{p}} \\ &\le \left(\frac{C_1}{\eta_A T} + \frac{C_1}{\eta_C T}\right) + C_2\eta_A + 0 + \frac{C_4}{\sqrt{E_C}} + C_5\eta_C\sqrt{2E_C} \\ &= \underbrace{\left(\frac{C_1}{\eta_A T} + C_2\eta_A\right)}_{2M_1} + \underbrace{\left(\frac{C_1}{\eta_C T} + \frac{C_4}{\sqrt{E_C}} + C_5\eta_C\sqrt{2E_C}\right)}_{3M_3} \\ &= 2M_1 + 3M_3. \end{aligned}$$

For **Family 3**, substituting $\eta = \min\{\eta_A, \eta_C, \eta_D\}, p = \frac{1}{E} = \frac{1}{\min\{\lceil E_C\rceil, \lceil E_D\rceil\}} = \frac{1}{\lceil \min\{E_C, E_D\}\rceil}$, and noting that $\min\{E_C, E_D\} \ge 1$ when $T \ge \max\left\{\frac{\sqrt{2}C_1C_5}{C_4^2}, 2\left(\frac{C_1^2C_6}{C_4^3}\right)^{\frac{1}{2}}\right\}$ therefore $E_C \le \lceil E_C\rceil < E_C + 1 \le 2E_C$, $E_D \le \lceil E_D\rceil < E_D + 1 \le 2E_D$, and using $\eta \le \eta_A$, $\frac{1}{\eta} \le \frac{1}{\eta_A} + \frac{1}{\eta_C} + \frac{1}{\eta_D}$, $\sqrt{p} = \frac{1}{\sqrt{E}} \le \frac{1}{\sqrt{E_C}} + \frac{1}{\sqrt{E_D}}$, $\frac{\eta}{\sqrt{p}} \le \eta_C \cdot \sqrt{2E_C}$, $\frac{\eta^2}{p} \le \eta_D^2 \cdot 2E_D$, we have:

$$\begin{aligned} f_3(\eta, p)_{C_3=0} &= f_2(\eta, p)_{C_3=0} + \frac{C_6\eta^2}{p} \\ &\le 2M_1 + 3M_3 + \underbrace{\left(\frac{C_1/2}{\eta_D T} + \frac{C_1/2}{\eta_D T} + \frac{C_4/2}{\sqrt{E_D}} + \frac{C_4/2}{\sqrt{E_D}} + 2C_6\eta_D^2 E_D\right)}_{5M_4} \\ &= 2M_1 + 3M_3 + 5M_4. \end{aligned}$$

□

Lemma D.4 (Iteration Complexity with Logarithmic Dependence)**.** For $\varepsilon > 0, \delta > 0$, and any $q > 0$, choosing

$$T \ge \frac{\mathrm{e}^{q^2}}{\varepsilon}\log^q\left[\mathrm{e} \vee \frac{1}{\varepsilon\delta}\right] \implies \frac{\log^q(T/\delta)}{T} \le \varepsilon$$

.

Proof. From the premise, we multiply both sides by ε, take the $1/q$-th power, and isolate the maximum operator to obtain

$$\log\left[\mathrm{e} \vee \frac{1}{\varepsilon\delta}\right] = \max\left(1, \log\frac{1}{\varepsilon\delta}\right) \le \mathrm{e}^{-q}(T\varepsilon)^{1/q}$$

The above inequality guarantee implies $T\varepsilon \ge \mathrm{e}^{q^2}$. Because the maximum of the function $f(x) = \frac{\log x}{x^{1/q}}$ is bounded by $1 - \mathrm{e}^{-q}$ for any $x \ge \mathrm{e}^{q^2}$, we have:

$$\log(T\varepsilon) \le (1 - \mathrm{e}^{-q})(T\varepsilon)^{1/q}$$

By leveraging $y \le \max(1, y)$, we can safely absorb any values of $\varepsilon\delta$ and evaluate the logarithmic terms in a single, unbroken inequality chain:

$$\log\left(\frac{T}{\delta}\right) = \log\left(\frac{1}{\varepsilon\delta}\right) + \log(T\varepsilon) \le \max\left(1, \log\frac{1}{\varepsilon\delta}\right) + \log(T\varepsilon) \le (T\varepsilon)^{1/q}$$

Since both sides are strictly positive, we raise the entire inequality to the power of q: $\log^q\left(\frac{T}{\delta}\right) \le T\varepsilon$, dividing both sides by T completes the proof: $\frac{\log^q(T/\delta)}{T} \le \varepsilon$. □

E Application 1: Stochastic Optimization with Mirror Descent

E.1 Theorem of First-Order Stationarity for Stochastic Optimization

Theorem E.1 (First-Order Stationarity for Stochastic Optimization with Mirror Descent)**.** Consider solving (†), where $\inf_{w\in\mathcal{W}} f(w) > -\infty$, via mirror descent $w_{t+1} = w_t - \eta_t P(w_t, U_t, \eta_t)$ using the normalized gradient estimator $U_t = v_t/\|v_t\|_*$. We establish the oracle and iteration complexities required to achieve ε-stationarity ($\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(w_t)\|_* \|P_t\|^2 \le \varepsilon$) with probability at least $1-\delta$ under various configurations of the correction term, momentum β_t, and reset probability p_t. To highlight structural similarities, we extract a shared **base rate** ρ per family, use the $\sigma_L, \sigma_\ell, \sigma_\gamma, \sigma_\alpha$ constants defined in (4), and introduce $\Theta_p(\cdot)$ to hide $\log\frac{T}{\delta}$ and κ dependencies.

Family 1 (zeroth-order correction). Under Assumptions 3.1,3.2,5.1, using constant step size $\eta_t = \eta = \frac{1}{L}\Theta_p(\sigma_L T^{-\frac{1}{2}} \wedge (\sigma_L^3/\sigma)^{\frac{1}{2}} T^{-\frac{3}{4}})$ and base rate $\rho = \frac{\sigma_L}{\sigma} T^{-\frac{1}{2}}$:

- **Momentum / SGD-M** ($\beta_t = \beta, p_t = 0$): With momentum parameter $1-\beta = \Theta_p(T^{-\frac{1}{3}} \vee \rho)$ and $B = 1$, oracle/iteration complexity is $N,\ T = \mathcal{O}\left(\frac{\sigma_L^2\sigma^2}{\varepsilon^4}\log\frac{\sigma_L^2\sigma^2}{\varepsilon^4\delta} \vee \frac{\sigma^3}{\varepsilon^3}\log^{\frac{3}{2}}\frac{\sigma^3}{\varepsilon^3\delta}\right)$.
- **Probabilistic Momentum** ($\beta_t = 1, p_t = p$): With reset probability $p = \Theta_p(\rho)$ and $B = 1/p$, oracle/iteration complexity is $N,\ T = \mathcal{O}\left(\frac{\sigma_L^2\sigma^2}{\varepsilon^4}\log^2\frac{\sigma_L^2\sigma^2}{\varepsilon^4\delta}\right)$.
- **Periodic Momentum** ($\beta_t = 1, p_t = \mathbb{I}_{\{t \mod E=0\}}$): With reset epoch $1/E = \Theta_p(\rho)$ and $B = E$, oracle/iteration complexity is $N,\ T = \mathcal{O}\left(\frac{\sigma_L^2\sigma^2}{\varepsilon^4}\log\frac{\sigma_L^2\sigma^2}{\varepsilon^4\delta}\right)$.

Family 2 (first-order correction). Under Assumptions 3.1,3.2,5.1,5.2, using constant step size $\eta_t = \eta = \Theta_p(\frac{1}{L}\sigma_L T^{-\frac{1}{2}} \wedge \frac{1}{\ell}(\sigma_\ell^4/\sigma)^{\frac{1}{3}} T^{-\frac{2}{3}})$ and base rate $\rho = \left(\frac{\sigma_\ell}{\sigma}\right)^{\frac{4}{3}} T^{-\frac{2}{3}}$:

- **STORM** ($\beta_t = \beta, p_t = 0$): With momentum parameter $1-\beta = \Theta_p(T^{-\frac{2}{3}} \vee \rho)$ and $B = 1$, oracle/iteration complexity is $N,\ T = \mathcal{O}\left(\frac{\sigma^3}{\varepsilon^3}\log^{\frac{3}{2}}\frac{\sigma^3}{\varepsilon^3\delta} \vee \frac{\sigma_\ell^2\sigma}{\varepsilon^3}\log\frac{\sigma_\ell^2\sigma}{\varepsilon^3\delta} \vee \frac{\sigma_L^2}{\varepsilon^2}\right)$.
- **PAGE / Loopless SVRG** ($\beta_t = 1, p_t = p$): With reset probability $p = \Theta_p(\rho)$ and $B = 1/p$, oracle/iteration complexity is $N,\ T = \mathcal{O}\left(\frac{\sigma_\ell^2\sigma}{\varepsilon^3}\log^{\frac{3}{2}}\frac{\sigma_\ell^2\sigma}{\varepsilon^3\delta} \vee \frac{\sigma_L^2}{\varepsilon^2}\right)$.
- **SPIDER / SVRG** ($\beta_t = 1, p_t = \mathbb{I}_{\{t \mod E=0\}}$): With reset epoch $1/E = \Theta_p(\rho)$ and $B = E$, oracle/iteration complexity is $N,\ T = \mathcal{O}\left(\frac{\sigma_\ell^2\sigma}{\varepsilon^3}\log\frac{\sigma_\ell^2\sigma}{\varepsilon^3\delta} \vee \frac{\sigma_L^2}{\varepsilon^2}\right)$.

Family 3 (second-order correction). Under Assumptions 3.1,3.2,5.1,5.3,5.4, using constant step size $\eta_t = \eta = \Theta_p(\frac{1}{L}\sigma_L T^{-\frac{1}{2}} \wedge \frac{1}{\gamma}(\sigma_\gamma^4/\sigma)^{\frac{1}{3}} T^{-\frac{2}{3}} \wedge \sqrt{\frac{1}{\alpha}(\sigma_\alpha^9/\sigma^4)^{\frac{1}{5}} T^{-\frac{3}{5}}})$ and base rate $\rho = \left(\frac{\sigma_\gamma}{\sigma}\right)^{\frac{4}{3}} T^{-\frac{2}{3}} \vee \left(\frac{\sigma_\alpha}{\sigma}\right)^{\frac{6}{5}} T^{-\frac{4}{5}}$:

- **Second Order Momentum** ($\beta_t = \beta, p_t = 0$):
  With momentum parameter $1-\beta = \Theta_p(T^{-\frac{2}{3}} \vee \rho)$ and $B = 1$, oracle/iteration complexity is $N,\ T = \mathcal{O}\left(\frac{\sigma^3}{\varepsilon^3}\log^{\frac{3}{2}}\frac{\sigma^3}{\varepsilon^3\delta} \vee \frac{\sigma_\gamma^2\sigma}{\varepsilon^3}\log\frac{\sigma_\gamma^2\sigma}{\varepsilon^3\delta} \vee \sqrt{\frac{\sigma_\alpha^3\sigma^2}{\varepsilon^5}}\log^{\frac{1}{2}}\left(\frac{1}{\delta}\sqrt{\frac{\sigma_\alpha^3\sigma^2}{\varepsilon^5}}\right) \vee \frac{\sigma_L^2}{\varepsilon^2}\right)$.
- **Second Order PAGE** ($\beta_t = 1, p_t = p$):
  With reset probability $p = \Theta_p(\rho)$ and $B = 1/p$, oracle/iteration complexity is $N,\ T = \mathcal{O}\left(\frac{\sigma_\gamma^2\sigma}{\varepsilon^3}\log^{\frac{3}{2}}\frac{\sigma_\gamma^2\sigma}{\varepsilon^3\delta} \vee \sqrt{\frac{\sigma_\alpha^3\sigma^2}{\varepsilon^5}}\log\left(\frac{1}{\delta}\sqrt{\frac{\sigma_\alpha^3\sigma^2}{\varepsilon^5}}\right) \vee \frac{\sigma_L^2}{\varepsilon^2}\right)$.
- **Second Order SPIDER** ($\beta_t = 1, p_t = \mathbb{I}_{\{t \mod E=0\}}$):
  With reset epoch $1/E = \Theta_p(\rho)$ and $B = E$, oracle/iteration complexity is $N,\ T = \mathcal{O}\left(\frac{\sigma_\gamma^2\sigma}{\varepsilon^3}\log\frac{\sigma_\gamma^2\sigma}{\varepsilon^3\delta} \vee \sqrt{\frac{\sigma_\alpha^3\sigma^2}{\varepsilon^5}}\log^{\frac{1}{2}}\left(\frac{1}{\delta}\sqrt{\frac{\sigma_\alpha^3\sigma^2}{\varepsilon^5}}\right) \vee \frac{\sigma_L^2}{\varepsilon^2}\right)$.

Table 5: Summary of Iteration (T) and Oracle (N) Complexities for Special Cases of the Unified Estimator to achieve ε-stationarity, i.e., $\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(w_t)\|_*\|P_t\|^2 \leq \varepsilon$. The $\mathcal{O}(\cdot)$ notation is omitted from the complexity bounds, these $\sigma_L, \sigma_\ell, \sigma_\gamma, \sigma_\alpha$ are defined in (4) and σ is from Assumption 3.2.

Algorithm	**Complexity** (N, T)
Family 1: Standard Recursion and Resetting (Zeroth-Order) **Asm:** 3.1,3.2, 5.1	
Momentum (SGD-M)	$\frac{\sigma_L^2\sigma^2}{\varepsilon^4}\log\frac{\sigma_L^2\sigma^2}{\varepsilon^4\delta} \vee \frac{\sigma^3}{\varepsilon^3}\log^{\frac{3}{2}}\frac{\sigma^3}{\varepsilon^3\delta}$
Probabilistic Momentum	$\frac{\sigma_L^2\sigma^2}{\varepsilon^4}\log^2\frac{\sigma_L^2\sigma^2}{\varepsilon^4\delta}$
Periodic Momentum	$\frac{\sigma_L^2\sigma^2}{\varepsilon^4}\log\frac{\sigma_L^2\sigma^2}{\varepsilon^4\delta}$
Family 2: Differential Recursion and Resetting (First-Order) **Asm:** 3.1,3.2, 5.1, 5.2	
STORM	$\frac{\sigma^3}{\varepsilon^3}\log^{\frac{3}{2}}\frac{\sigma^3}{\varepsilon^3\delta} \vee \frac{\sigma_\ell^2\sigma}{\varepsilon^3}\log\frac{\sigma_\ell^2\sigma}{\varepsilon^3\delta} \vee \frac{\sigma_L^2}{\varepsilon^2}$
PAGE / Loopless SVRG	$\frac{\sigma_\ell^2\sigma}{\varepsilon^3}\log^{\frac{3}{2}}\frac{\sigma_\ell^2\sigma}{\varepsilon^3\delta} \vee \frac{\sigma_L^2}{\varepsilon^2}$
SPIDER / SVRG	$\frac{\sigma_\ell^2\sigma}{\varepsilon^3}\log\frac{\sigma_\ell^2\sigma}{\varepsilon^3\delta} \vee \frac{\sigma_L^2}{\varepsilon^2}$
Family 3: Second-Order Recursion and Resetting (Second-Order) **Asm:** 3.1,3.2, 5.1, 5.3, 5.4	
Second-Order Momentum	$\frac{\sigma^3}{\varepsilon^3}\log^{\frac{3}{2}}\frac{\sigma^3}{\varepsilon^3\delta} \vee \frac{\sigma_\gamma^2\sigma}{\varepsilon^3}\log\frac{\sigma_\gamma^2\sigma}{\varepsilon^3\delta} \vee \sqrt{\frac{\sigma_\alpha^3\sigma^2}{\varepsilon^5}}\log^{\frac{1}{2}}\left(\frac{1}{\delta}\sqrt{\frac{\sigma_\alpha^3\sigma^2}{\varepsilon^5}}\right) \vee \frac{\sigma_L^2}{\varepsilon^2}$
Second-Order PAGE	$\frac{\sigma_\gamma^2\sigma}{\varepsilon^3}\log^{\frac{3}{2}}\frac{\sigma_\gamma^2\sigma}{\varepsilon^3\delta} \vee \sqrt{\frac{\sigma_\alpha^3\sigma^2}{\varepsilon^5}}\log\left(\frac{1}{\delta}\sqrt{\frac{\sigma_\alpha^3\sigma^2}{\varepsilon^5}}\right) \vee \frac{\sigma_L^2}{\varepsilon^2}$
Second-Order SPIDER	$\frac{\sigma_\gamma^2\sigma}{\varepsilon^3}\log\frac{\sigma_\gamma^2\sigma}{\varepsilon^3\delta} \vee \sqrt{\frac{\sigma_\alpha^3\sigma^2}{\varepsilon^5}}\log^{\frac{1}{2}}\left(\frac{1}{\delta}\sqrt{\frac{\sigma_\alpha^3\sigma^2}{\varepsilon^5}}\right) \vee \frac{\sigma_L^2}{\varepsilon^2}$

E.2 Proof for Theorem E.1 of First-Order Stationarity for Stochastic Optimization

Proof for Theorem E.1, Application 1: Stochastic Optimization with Mirror Descent. The update rule of the Mirror Descent framework is given by $w_{t+1} = w_t - \eta_t P(w_t, U_t, \eta_t)$ using proximal gradient mapping $P(w_t, U_t, \eta_t)$ with update vector U_t and step size η_t. Moreover, in Application 1: Unconstrained Optimization, the update vector $U_t = \frac{v_t}{\|v_t\|_*}$ is the normalized direction of the unified estimator v_t of the objective gradient $g(w_t) = \nabla f(w_t)$ at the t-th iteration.

$$w_{t+1} = w_t - \eta_t P(w_t, U_t, \eta_t) = w_t - \eta_t P\left(w_t, \frac{v_t}{\|v_t\|_*}, \eta_t\right)$$

By smoothness of $f(w)$ (Assumption 5.1 for $g(w) = \nabla f(w)$) and the above update rule, we have

$$f(w_{t+1}) - f(w_t) \leq -\eta_t\left\langle \nabla f(w_t), P\left(w_t, \frac{v_t}{\|v_t\|_*}, \eta_t\right)\right\rangle + \frac{L\eta_t^2}{2}\left\|P\left(w_t, \frac{v_t}{\|v_t\|_*}, \eta_t\right)\right\|^2$$

We upperbound the 1st term on the right-hand side by noting the tracking error $e_t \equiv v_t - g(w_t) = v_t - \nabla f(w_t)$ hence $v_t = \nabla f(w_t) + e_t$, and invoking Lemma D.1 with $u \leftarrow \nabla f(w_t), \delta \leftarrow e_t, \eta \leftarrow \eta_t$.

$$-\eta_t\left\langle \nabla f(w_t), P\left(w_t, \frac{v_t}{\|v_t\|_*}, \eta_t\right)\right\rangle \leq -\frac{\eta_t}{4}\left\langle \nabla f(w_t), P\left(w_t, \frac{\nabla f(w_t)}{\|\nabla f(w_t)\|_*}, \frac{\eta_t}{2}\right)\right\rangle + 2\eta_t\|e_t\|_*$$

We upperbound the 2nd term on the right-hand side by using the properties of the proximal gradient mapping (Lemma A.7) such that $\|P(w, u, \eta)\| \leq \|u\|_*$, we have:

$$\frac{L\eta_t^2}{2}\left\|P\left(w_t, \frac{v_t}{\|v_t\|_*}, \eta_t\right)\right\|^2 \leq \frac{L\eta_t^2}{2}\left\|\frac{v_t}{\|v_t\|_*}\right\|_*^2 = \frac{L\eta_t^2}{2}$$

Substituting the above upper bounds for 1st and 2nd terms on the right-hand side, multiplying by 4:

$$\eta_t\left\langle \nabla f(w_t), P\left(w_t, \frac{\nabla f(w_t)}{\|\nabla f(w_t)\|_*}, \frac{\eta_t}{2}\right)\right\rangle \leq 4\left[f(w_t) - f(w_{t+1})\right] + 2L\eta_t^2 + 8\eta_t\|e_t\|_*$$

Table 6: Configuration Details: Selection of Correction Term ($\mathcal{T}_t$), constant Step Size ($\eta_t = \eta$), and Momentum/Reset Parameters (β_t, p_t). $\Theta_p(\cdot)$ hides the dependencies of $\log \frac{T}{\delta}$ and κ, omitted for brevity and ease of comparison. To highlight the structural similarity across algorithms, we extract a shared **base rate** ρ for each family. The probabilistic reset p and periodic reset $1/E$ exactly match ρ, while the momentum parameter uses $1-\beta =$ base-decay $\vee\, \rho$.

Algorithm	β_t	p_t	**Parameter Selection**	**Batch size** B
Family 1: Standard Recursion and Resetting (Zeroth-Order) **Asm:** 3.1,3.2, 5.1				
Correction Term $\mathcal{T}_t = \beta_t\big(G(w_{t-1},\xi_t) - G(w_t,\xi_t)\big)$				
Step Size $\eta = \frac{1}{L}\Theta_p\left(\sigma_L T^{-\frac{1}{2}} \wedge \left(\frac{\sigma_L^3}{\sigma}\right)^{\frac{1}{2}} T^{-\frac{3}{4}}\right)$ **Base Rate** $\rho = \frac{\sigma_L}{\sigma} T^{-\frac{1}{2}}$				
Momentum (SGD-M)	β	0	$1-\beta = \Theta_p(T^{-\frac{1}{3}} \vee \rho)$	1
Probabilistic Momentum	1	p	$p = \Theta_p(\rho)$	$1/p$
Periodic Momentum	1	$\mathbb{I}\{t \pmod E = 0\}$	$1/E = \Theta_p(\rho)$	E
Family 2: Differential Recursion and Resetting (First-Order) **Asm:** 3.1,3.2, 5.1, 5.2				
Correction Term $\mathcal{T}_t = 0$				
Step Size $\eta = \Theta_p\left(\frac{1}{L}\sigma_L T^{-\frac{1}{2}} \wedge \frac{1}{\ell}\left(\frac{\sigma_\ell^4}{\sigma}\right)^{\frac{1}{3}} T^{-\frac{2}{3}}\right)$ **Base Rate** $\rho = \left(\frac{\sigma_\ell}{\sigma}\right)^{\frac{4}{3}} T^{-\frac{2}{3}}$				
STORM	β	0	$1-\beta = \Theta_p(T^{-\frac{2}{3}} \vee \rho)$	1
PAGE / Loopless SVRG	1	p	$p = \Theta_p(\rho)$	$1/p$
SPIDER / SVRG	1	$\mathbb{I}\{t \pmod E = 0\}$	$1/E = \Theta_p(\rho)$	E
Family 3: Second-Order Recursion and Resetting (Second-Order) **Asm:** 3.1,3.2, 5.1, 5.3, 5.4				
Correction Term $\mathcal{T}_t = \beta_t\big(G(w_{t-1},\xi_t) - [G(w_t,\xi_t) + \nabla G(w_t,\xi_t)(w_{t-1}-w_t)]\big)$				
Step Size $\eta = \Theta_p\left(\frac{1}{L}\sigma_L T^{-\frac{1}{2}} \wedge \frac{1}{\gamma}\left(\frac{\sigma_\gamma^4}{\sigma}\right)^{\frac{1}{3}} T^{-\frac{2}{3}} \wedge \sqrt{\frac{1}{\alpha}\left(\frac{\sigma_\alpha^9}{\sigma^4}\right)^{\frac{1}{5}}} T^{-\frac{3}{5}}\right)$				
Base Rate $\rho = \left(\frac{\sigma_\gamma}{\sigma}\right)^{\frac{4}{3}} T^{-\frac{2}{3}} \vee \left(\frac{\sigma_\alpha}{\sigma}\right)^{\frac{6}{5}} T^{-\frac{4}{5}}$				
Second-Order Momentum	β	0	$1-\beta = \Theta_p(T^{-\frac{2}{3}} \vee \rho)$	1
Second-Order PAGE	1	p	$p = \Theta_p(\rho)$	$1/p$
Second-Order SPIDER	1	$\mathbb{I}\{t \pmod E = 0\}$	$1/E = \Theta_p(\rho)$	E

By introducing the infimum $\inf_{w\in\mathcal{W}} f(w)$ of the objective $f(w)$ over $\mathcal{W}$, then $\inf_{w\in\mathcal{W}} f(w) \le f(w_{T+1})$, and summing up the above inequality from $t=1$ to T:

$$\sum_{t=1}^{T} \eta_t \left\langle \nabla f(w_t), P\left(w_t, \frac{\nabla f(w_t)}{\|\nabla f(w_t)\|_*}, \frac{\eta_t}{2}\right)\right\rangle \le 4\left[f(w_1) - \inf_{w\in\mathcal{W}} f(w)\right] + 2L\sum_{t=1}^{T}\eta_t^2 + 8\sum_{t=1}^{T}\eta_t\|e_t\|_*$$

Applying the inner product lower bound in properties of the proximal gradient mapping (Lemma A.7) such that $\|P(w,u,\eta)\|^2 \le \langle u, P(w,u,\eta)\rangle$, we have:

$$\sum_{t=1}^{T} \eta_t \|\nabla f(w_t)\|_* \left\|P\left(w_t, \frac{\nabla f(w_t)}{\|\nabla f(w_t)\|_*}, \frac{\eta_t}{2}\right)\right\|^2 \le 4[f(w_1) - \inf_{w\in\mathcal{W}} f(w)] + 2L\sum_{t=1}^{T}\eta_t^2 + 8\sum_{t=1}^{T}\eta_t\|e_t\|_*$$

Upper bound on stationarity for all special cases. Since $\eta_t = \eta$ are constants in all special cases, divide the inequality above by ηT:

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(w_t)\|_* \left\|P\left(w_t, \frac{\nabla f(w_t)}{\|\nabla f(w_t)\|_*}, \frac{\eta}{2}\right)\right\|^2 \le \frac{4[f(w_1) - \inf_{w\in\mathcal{W}} f(w)]}{\eta T} + 2L\eta + \frac{8}{T}\sum_{t=1}^{T}\|e_t\|_*.$$

We define these notations for brevity, $P_t \coloneqq P\left(w_t, \frac{\nabla f(w_t)}{\|\nabla f(w_t)\|_*}, \frac{\eta}{2}\right)$, $\Delta_f \coloneqq f(w_1) - \inf_{w\in\mathcal{W}} f(w)$. Also, we have the trivial bound $\frac{1}{T}\sum_{t=1}^T \|P_t\|^2 \le 1$, since $\|P(w,u,\eta)\| \le \|u\|_*$ by Lemma A.7 and therefore $\|P_t\| \le \left\|\frac{\nabla f(w_t)}{\|\nabla f(w_t)\|_*}\right\|_* = 1$.

$$\frac{1}{T}\sum_{t=1}^T \|\nabla f(w_t)\|_* \|P_t\|^2 \le \frac{4\Delta_f}{\eta T} + 2L\eta + \frac{8}{T}\sum_{t=1}^T \|e_t\|_*.$$

Substituting the upper bounds of $\sum_{t=1}^T \|e_t\|_*$ for three families with $\|U_t\|_* = \left\|\frac{v_t}{\|v_t\|_*}\right\|_* = 1$ (Assumption 2.3 is automatically satisfied with $G = 1$ and no need for extra assumption), and introducing $\sigma_L \coloneqq (\Delta_f L)^{\frac{1}{2}}$ with L from Assumption 5.1, $\sigma_\ell \coloneqq (\Delta_f \ell)^{\frac{1}{2}}$ with ℓ from Assumption 5.2, $\sigma_\gamma \coloneqq (\Delta_f \gamma)^{\frac{1}{2}}$ with γ from Assumption 5.3, $\sigma_\alpha \coloneqq (\Delta_f^2 \alpha)^{\frac{1}{3}}$ with α from Assumption 5.4, we obtain the following upperbounds of $\frac{1}{T}\sum_{t=1}^T \|\nabla f(w_t)\|_* \|P_t\|^2$ for three cases in each of three families:

Family 1, Case 1 (Momentum (SGD-M)) ($p_t = 0, \beta_t = \beta$): By invoking the established upper bound of $\|e_t\|_*$ for the estimator v_t with zeroth-order correction term $\mathcal{T}_t$, the selection of $p_t = 0, \beta_t = \beta$ and $\eta_t = \eta$ in the third-level section of Appendix C.2.1, and noting that $\sum_{t=1}^T \beta^t = \frac{\beta(1-\beta^T)}{1-\beta} \le \frac{1}{1-\beta}$ for $\beta \in (0,1)$, we show the following upper bound for $\frac{1}{T}\sum_{t=1}^T \|e_t\|_*$:

$$\frac{1}{T}\sum_{t=1}^T \|e_t\|_* \le \frac{C\left(\frac{\delta}{2T}, \kappa\right)\sigma/\sqrt{B}}{(1-\beta)T} + \frac{L\eta}{1-\beta} + C\left(\frac{\delta}{2T}, \kappa\right)\sigma\sqrt{1-\beta}$$

By selecting the batch size $B = 1$ and noting $C\left(\frac{\delta}{2T}, \kappa\right) \le 2\sqrt{2[\kappa \vee \log\frac{2T}{\delta}]}$, we have:

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^T \|\nabla f(w_t)\|_* \|P_t\|^2 \le \frac{4\Delta_f}{\eta T} + 2L\eta + 8\cdot\frac{1}{T}\sum_{t=1}^T \|e_t\|_* \\
\le\ & \frac{4\Delta_f}{\eta T} + 2L\cdot\eta + \frac{16\sigma\sqrt{2[\kappa\vee\log\frac{2T}{\delta}]}}{(1-\beta)T} + 16\sigma\sqrt{2[\kappa\vee\log\frac{2T}{\delta}]}\cdot\sqrt{1-\beta} + 8L\cdot\frac{\eta}{1-\beta}
\end{aligned}$$

By letting $C_1 = 4\Delta_f, C_2 = 2L, C_3 = 16\sigma\sqrt{2[\kappa\vee\log\frac{2T}{\delta}]}, C_4 = 16\sigma\sqrt{2[\kappa\vee\log\frac{2T}{\delta}]}, C_5 = 8L$ in Lemma D.2, selecting the step size and momentum parameter as $\eta = \min\left\{\left(\frac{C_1}{C_2 T}\right)^{\frac{1}{2}}, \left(\frac{C_1^3}{C_4^2 C_5 T^3}\right)^{\frac{1}{4}}\right\} = \frac{1}{L}\min\left\{\sqrt{2}\sigma_L T^{-\frac{1}{2}}, \left(\frac{\sigma_L^6}{64\sigma^2}\right)^{\frac{1}{4}}[\kappa\vee\log\frac{2T}{\delta}]^{-\frac{1}{4}}T^{-\frac{3}{4}}\right\}, 1-\beta = \max\left\{\left(\frac{C_3}{C_4 T}\right)^{\frac{2}{3}}, \left(\frac{C_1 C_5}{C_4^2 T}\right)^{\frac{1}{2}}\right\} = \max\left\{T^{-\frac{2}{3}}, \left(\frac{\sigma_L^2/\sigma^2}{16[\kappa\vee\log\frac{2T}{\delta}]}\right)^{\frac{1}{2}}T^{-\frac{1}{2}}\right\}$ when $T[\kappa\vee\log\frac{2T}{\delta}] \ge \frac{\sigma_L^2}{16\sigma^2}$, we have:

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^T \|\nabla f(w_t)\|_* \|P_t\|^2 \le 2\left(\frac{C_1 C_2}{T}\right)^{\frac{1}{2}} + 2\left(\frac{C_3 C_4^2}{T}\right)^{\frac{1}{3}} + 3\left(\frac{C_1 C_4^2 C_5}{T}\right)^{\frac{1}{4}} \\
\le\ & 4\sqrt{2}\sigma_L T^{-\frac{1}{2}} + 32\sqrt{2}\sigma\left(\frac{[\kappa\vee\log\frac{2T}{\delta}]^{\frac{3}{2}}}{T}\right)^{\frac{1}{3}} + 24\sqrt{\sigma_L\sigma}\left(\frac{4[\kappa\vee\log\frac{2T}{\delta}]}{T}\right)^{\frac{1}{4}}
\end{aligned}$$

To analyze the iteration complexity, the following inequalities of T are sufficient to show that $\frac{1}{T}\sum_{t=1}^T \|\nabla f(w_t)\|_* \|P_t\|^2 \le \varepsilon$ for some $\varepsilon > 0$ as 1st, 2nd and 3rd terms are upper bounded by $\varepsilon/2, \varepsilon/4$ and $\varepsilon/4$ respectively, noting $\kappa \ge 1$ and $T \ge \frac{\sigma_L^2/\sigma^2}{16\kappa}$ implies $T[\kappa\vee\log\frac{2T}{\delta}] \ge T\kappa \ge \frac{\sigma_L^2}{16\sigma^2}$:

$$\begin{aligned}
T \ge& \frac{\sigma_L^2/\sigma^2}{16\kappa} \vee \frac{128\sigma_L^2}{\varepsilon^2} \vee \frac{\kappa\sigma^3}{(\varepsilon/128\sqrt{2})^3} \vee \frac{4\kappa\cdot\sigma_L^2\sigma^2}{(\varepsilon/96)^4}, \\
&\frac{\log^{\frac{3}{2}}\frac{2T}{\delta}}{T} \le \left(\frac{\varepsilon}{128\sqrt{2}\sigma}\right)^3, \frac{\log\frac{2T}{\delta}}{T} \le \frac{1}{4}\left(\frac{\varepsilon}{96\sqrt{\sigma_L\sigma}}\right)^4
\end{aligned}$$

By substituting $\delta \leftarrow \delta/2$ and $(q,\varepsilon) \leftarrow (\frac{3}{2}, \left(\frac{\varepsilon}{128\sqrt{2}\sigma}\right)^3), (1, \frac{1}{4}\left(\frac{\varepsilon}{96\sqrt{\sigma_L\sigma}}\right)^4)$ in Lemma D.4, the following condition of T is sufficient to guarantee the above inequalities, and therefore $\frac{1}{T}\sum_{t=1}^T \|\nabla f(w_t)\|_* \|P_t\|^2 \le \varepsilon$.

$$T \ge \frac{\sigma_L^2/\sigma^2}{16\kappa} \vee \frac{128\sigma_L^2}{\varepsilon^2} \vee \frac{\sigma^3}{(\varepsilon/128\sqrt{2})^3}\left[\kappa \vee \mathrm{e}^{\frac{9}{4}} \log^{\frac{3}{2}}\left[\mathrm{e} \vee \frac{2\sigma^3}{(\varepsilon/128\sqrt{2})^3\delta}\right]\right]$$
$$\vee \frac{4\cdot\sigma_L^2\sigma^2}{(\varepsilon/96)^4}\left[\kappa \vee \mathrm{e}\log\left[\mathrm{e} \vee \frac{8\cdot\sigma_L^2\sigma^2}{(\varepsilon/96)^4\delta}\right]\right]$$

In particular, for all $\varepsilon \in \mathbb{R}_+$ is sufficiently small such that $\varepsilon \le \varepsilon_0(\sigma_L,\sigma,\kappa,\delta)$, then the lower bound on T simplifies asymptotically to

$$T \ \ge \ \max\left\{\frac{4\mathrm{e}\cdot\sigma_L^2\sigma^2}{(\varepsilon/96)^4}\log\frac{8\sigma_L^2\sigma^2}{(\varepsilon/96)^4\delta},\ \frac{\mathrm{e}^{\frac{9}{4}}\cdot\sigma^3}{(\varepsilon/128\sqrt{2})^3}\log^{\frac{3}{2}}\frac{2\sigma^3}{(\varepsilon/128\sqrt{2})^3\delta}\right\}$$

where the uniform upper bound $\varepsilon_0(\sigma_L,\sigma,\kappa,\delta)$ of ε is obtained by noting $\kappa \ge 1$ and solving the following inequalities for ε to ensure the dominated terms in the above asymptotic bound are larger than the other terms of the lower bound on T.

$$\begin{cases} \dfrac{2\sigma^3}{(\varepsilon/128\sqrt{2})^3\delta} \ge \mathrm{e},\ \dfrac{8\sigma_L^2\sigma^2}{(\varepsilon/96)^4\delta} \ge \mathrm{e} \\ \mathrm{e}\log\dfrac{8\sigma_L^2\sigma^2}{(\varepsilon/96)^4\delta} \ge \kappa,\ \mathrm{e}^{\frac{9}{4}}\log^{\frac{3}{2}}\dfrac{2\sigma^3}{(\varepsilon/128\sqrt{2})^3\delta} \ge \kappa \\ \dfrac{4\mathrm{e}\cdot\sigma_L^2\sigma^2}{(\varepsilon/96)^4} \ge \dfrac{\sigma_L^2/\sigma^2}{16\kappa},\ \dfrac{4\mathrm{e}\cdot\sigma_L^2\sigma^2}{(\varepsilon/96)^4} \ge \dfrac{128\sigma_L^2}{\varepsilon^2} \end{cases}$$

The explicit expression of $\varepsilon_0(\sigma_L,\sigma,\kappa,\delta) \equiv \sqrt{\sigma_L\sigma}\cdot\varepsilon_1(\kappa,\delta) \vee \sigma\cdot\varepsilon_2(\kappa,\delta)$ is given by:

$$\varepsilon_1(\kappa,\delta) = 96\left(\frac{8}{\delta\,\mathrm{e}^{1\vee(\kappa/\mathrm{e})}}\right)^{1/4}$$
$$\varepsilon_2(\kappa,\delta) = 1152\sqrt{2}\mathrm{e}^{1/2} \ \wedge\ 192\sqrt{2}(\mathrm{e}\kappa)^{1/4} \ \wedge\ 128\sqrt{2}\left(\frac{2}{\delta\,\mathrm{e}^{1\vee(\kappa^{2/3}/\mathrm{e}^{3/2})}}\right)^{1/3}$$

Therefore, the iteration complexity is $T = \mathcal{O}\left(\frac{\sigma_L^2\sigma^2}{\varepsilon^4}\log\frac{\sigma_L^2\sigma^2}{\varepsilon^4\delta} \vee \frac{\sigma^3}{\varepsilon^3}\log^{\frac{3}{2}}\frac{\sigma^3}{\varepsilon^3\delta}\right)$, and the oracle complexity is $N = B\cdot T = \mathcal{O}\left(\frac{\sigma_L^2\sigma^2}{\varepsilon^4}\log\frac{\sigma_L^2\sigma^2}{\varepsilon^4\delta} \vee \frac{\sigma^3}{\varepsilon^3}\log^{\frac{3}{2}}\frac{\sigma^3}{\varepsilon^3\delta}\right)$ with the selection of batch size $B = 1$ to guarantee $\frac{1}{T}\sum_{t=1}^T \|\nabla f(w_t)\|_*\|P_t\|^2 \le \varepsilon$ with probability at least $1-\delta$ for sufficiently small ε.

Family 1, Case 2 (Probabilistic Momentum) ($p_t = p, \beta_t = 1$): By invoking the established upper bound of $\|e_t\|_*$ for the estimator v_t with zeroth-order correction term $\mathcal{T}_t$, the selection of $p_t = p, \beta_t = 1$ and $\eta_t = \eta$ in the third-level section of Appendix C.2.2, we show the following upper bound of $\frac{1}{T}\sum_{t=1}^T \|e_t\|_*$:

$$\frac{1}{T}\sum_{t=1}^T \|e_t\|_* \le C\left(\frac{\delta}{4T},\kappa\right)\sigma/\sqrt{B} + \frac{L\eta}{p}\log\frac{4T}{\delta}$$

By selecting the batch size $B = \frac{1}{p}$, and noting $C\left(\frac{\delta}{4T},\kappa\right) \le 2\sqrt{2[\kappa \vee \log\frac{4T}{\delta}]}$, we have:

$$\frac{1}{T}\sum_{t=1}^T \|\nabla f(w_t)\|_*\|P_t\|^2 \le \frac{4\Delta_f}{\eta T} + 2L\eta + 8\cdot\frac{1}{T}\sum_{t=1}^T\|e_t\|_*$$
$$\le \ \frac{4\Delta_f}{\eta T} + 2L\cdot\eta + 0 + 16\sigma\sqrt{2[\kappa\vee\log\frac{4T}{\delta}]}\cdot\sqrt{p} + 8L\log\frac{4T}{\delta}\cdot\frac{\eta}{p}$$

Letting $C_1 = 4\Delta_f, C_2 = 2L, C_3 = 0, C_4 = 16\sigma\sqrt{2[\kappa\vee\log\frac{4T}{\delta}]}, C_5 = 8L\log\frac{4T}{\delta}$ in Lemma D.3, selecting step size and reset probability as $\eta = \min\left\{\left(\frac{C_1}{C_2}\right)^{\frac{1}{2}}T^{-\frac{1}{2}}, \left(\frac{2C_1^3}{C_4^2C_5}\right)^{\frac{1}{4}}T^{-\frac{3}{4}}\right\} =$

$\frac{1}{L}\min\left\{\sqrt{2}\sigma_L T^{-\frac{1}{2}}, \frac{1}{2}\left(\frac{\sigma_L^6}{2\sigma^2}\right)^{\frac{1}{4}}[\kappa \vee \log\frac{4T}{\delta}]^{-\frac{1}{4}}[\log\frac{4T}{\delta}]^{-\frac{1}{4}}T^{-\frac{3}{4}}\right\}$, $B = \frac{1}{p} = \left\lceil\left(\frac{C_4^2 T}{8C_1C_5}\right)^{\frac{1}{2}}\right\rceil = \left\lceil\frac{\sqrt{2}\sigma}{\sigma_L}\left(\frac{\kappa\vee\log\frac{4T}{\delta}}{\log\frac{4T}{\delta}}\right)^{\frac{1}{2}}T^{\frac{1}{2}}\right\rceil$, then when $T\left(\frac{\kappa\vee\log\frac{4T}{\delta}}{\log\frac{4T}{\delta}}\right) \geq \frac{\sigma_L^2}{2\sigma^2}$, we have:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(w_t)\|_*\|P_t\|^2 &\leq 2\left(\frac{C_1C_2}{T}\right)^{\frac{1}{2}} + 4\left(\frac{C_1C_4^2C_5}{2T}\right)^{\frac{1}{4}} \\ &\leq 4\sqrt{2}\sigma_L T^{-\frac{1}{2}} + 32\sqrt{\sigma_L\sigma}\left(\frac{2[\kappa\vee\log\frac{4T}{\delta}]\log\frac{4T}{\delta}}{T}\right)^{\frac{1}{4}}\end{aligned}$$

To analyze the iteration complexity, the following inequalities of T are sufficient to show $\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(w_t)\|_*\|P_t\|^2 \leq \varepsilon$ for some $\varepsilon > 0$ as 1st, 2nd terms are upper bounded by $\varepsilon/2$, and noting that $\kappa \geq 1$ and $T \geq \frac{\sigma_L^2}{2\sigma^2}$ implies $T\left(\frac{\kappa\vee\log\frac{4T}{\delta}}{\log\frac{4T}{\delta}}\right) \geq T \geq \frac{\sigma_L^2}{2\sigma^2}$:

$$T \geq \frac{\sigma_L^2}{2\sigma^2} \vee \frac{128\sigma_L^2}{\varepsilon^2}, \frac{\log\frac{4T}{\delta}}{T} \leq \frac{(\varepsilon/64)^4}{2\kappa\cdot\sigma_L^2\sigma^2}, \frac{\log^2\frac{4T}{\delta}}{T} \leq \frac{(\varepsilon/64)^4}{2\sigma_L^2\sigma^2}$$

By substituting $\delta \leftarrow \delta/4$ and $(q,\varepsilon) \leftarrow (1, \frac{(\varepsilon/64)^4}{2\kappa\cdot\sigma_L^2\sigma^2}), (2, \frac{(\varepsilon/64)^4}{2\sigma_L^2\sigma^2})$ in Lemma D.3, the following condition of T is sufficient to show $\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(w_t)\|_*\|P_t\|^2 \leq \varepsilon$:

$$T \geq \frac{\sigma_L^2}{2\sigma^2} \vee \frac{128\sigma_L^2}{\varepsilon^2} \vee \frac{2\sigma_L^2\sigma^2}{(\varepsilon/64)^4}\left[\mathrm{e}^4\log^2\left[\mathrm{e}\vee\frac{8\sigma_L^2\sigma^2}{(\varepsilon/64)^4\delta}\right] \vee \mathrm{e}\kappa\log\left[\mathrm{e}\vee\kappa\frac{8\sigma_L^2\sigma^2}{(\varepsilon/64)^4\delta}\right]\right]$$

In particular, for all $\varepsilon \in \mathbb{R}_+$ is sufficiently small such that $\varepsilon \leq \varepsilon_0(\sigma_L, \sigma, \kappa, \delta)$, then the lower bound on T simplifies asymptotically to

$$T \geq \frac{\mathrm{e}^4\cdot 2\sigma_L^2\sigma^2}{(\varepsilon/64)^4}\log^2\frac{8\sigma_L^2\sigma^2}{(\varepsilon/64)^4\delta},$$

where the uniform upper bound $\varepsilon_0(\sigma_L, \sigma, \kappa, \delta)$ is obtained by noting $\kappa \geq 1$ and solving the following inequalities for ε to ensure the dominated terms in the above asymptotic bound are larger than the other terms of the lower bound on T.

$$\begin{cases}\frac{8\sigma_L^2\sigma^2}{(\varepsilon/96)^4\delta} \geq \mathrm{e}, \frac{8\sigma_L^2\sigma^2}{(\varepsilon/96)^4\delta} \geq \kappa \\ \frac{\mathrm{e}^4}{2\mathrm{e}}\log\frac{8\sigma_L^2\sigma^2}{(\varepsilon/96)^4\delta} \geq \kappa \\ \frac{\mathrm{e}^4\cdot 2\sigma_L^2\sigma^2}{(\varepsilon/64)^4} \geq \frac{\sigma_L^2}{2\sigma^2}, \frac{\mathrm{e}^4\cdot 2\sigma_L^2\sigma^2}{(\varepsilon/64)^4} \geq \frac{128\sigma_L^2}{\varepsilon^2}\end{cases}$$

The explicit expression of $\varepsilon_0(\sigma_L, \sigma, \kappa, \delta) \equiv \sqrt{\sigma_L\sigma}\cdot\varepsilon_1(\kappa,\delta) \vee \sigma\cdot\varepsilon_2(\kappa,\delta)$ is given by:

$$\begin{aligned}\varepsilon_1(\kappa,\delta) &= 96\left(\frac{8}{\delta\left(\mathrm{e}\vee\kappa\vee\mathrm{e}^{2\kappa/\mathrm{e}^3}\right)}\right)^{1/4} \\ \varepsilon_2(\kappa,\delta) &= 64\sqrt{2}\mathrm{e} \wedge 512\mathrm{e}^2\end{aligned}$$

Therefore, the iteration complexity is $T = \mathcal{O}\left(\frac{\sigma_L^2\sigma^2}{\varepsilon^4}\log^2\frac{\sigma_L^2\sigma^2}{\varepsilon^4\delta}\right)$, and the expected value of oracle complexity is $N = [B\cdot p + 1\cdot(1-p)]T = T(2-1/B) = \mathcal{O}\left(\frac{\sigma_L^2\sigma^2}{\varepsilon^4}\log^2\frac{\sigma_L^2\sigma^2}{\varepsilon^4\delta}\right)$ with the selection of batch size $B = \frac{1}{p}$ to guarantee $\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(w_t)\|_*\|P_t\|^2 \leq \varepsilon$ with probability at least $1-\delta$ for sufficiently small ε.

Family 1, Case 3 (Periodic Momentum) ($p_t = \mathbb{I}_{\{t \bmod E=0\}}, \beta_t = 1$): By invoking the established upper bound of $\|e_t\|_*$ for the estimator v_t with zeroth-order correction term $\mathcal{T}_t$, the selection of

$p_t = \mathbb{I}_{\{t \mod E=0\}}, \beta_t = 1$ and $\eta_t = \eta$ in the third-level section of Appendix C.2.3, we show the following upper bound of $\frac{1}{T}\sum_{t=1}^{T}\|e_t\|_*$:

$$\frac{1}{T}\sum_{t=1}^{T}\|e_t\|_* \leq C\left(\frac{\delta}{2T/E}, \kappa\right)\sigma/\sqrt{B} + L\eta E$$

By selecting the batch size $B = E$, noting $C\left(\frac{\delta}{2T/E}, \kappa\right) \leq C\left(\frac{\delta}{2T}, \kappa\right) \leq 2\sqrt{2[\kappa \vee \log\frac{2T}{\delta}]}$, we have:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(w_t)\|_*\|P_t\|^2 &\leq \frac{4\Delta_f}{\eta T} + 2L\eta + 8\cdot\frac{1}{T}\sum_{t=1}^{T}\|e_t\|_* \\ &\leq \frac{4\Delta_f}{\eta T} + 2L\cdot\eta + 0 + 16\sigma\sqrt{2[\kappa\vee\log\frac{2T}{\delta}]}\cdot\frac{1}{\sqrt{E}} + 8L\cdot\eta E\end{aligned}$$

Therefore, by letting $C_1 = 4\Delta_f, C_2 = 2L, C_3 = 0, C_4 = 16\sigma\sqrt{2[\kappa\vee\log\frac{2T}{\delta}]}, C_5 = 8L$ in Lemma D.3, selecting the step size $\eta = \min\left\{\left(\frac{C_1}{C_2}\right)^{\frac{1}{2}}T^{-\frac{1}{2}}, \left(\frac{2C_1^3}{C_4^2C_5}\right)^{\frac{1}{4}}T^{-\frac{3}{4}}\right\} = \frac{1}{L}\min\{\sqrt{2}\sigma_L T^{-\frac{1}{2}}, \frac{1}{2}\left(\frac{\sigma_L^6}{2\sigma^2}\right)^{\frac{1}{4}}[\kappa\vee\log\frac{2T}{\delta}]^{-\frac{1}{4}}T^{-\frac{3}{4}}\}$ and batch size $B = E = \left\lceil\left(\frac{C_4^2T}{8C_1C_5}\right)^{\frac{1}{2}}\right\rceil = \lceil\frac{\sqrt{2}\sigma}{\sigma_L}[\kappa\vee\log\frac{2T}{\delta}]^{\frac{1}{2}}T^{\frac{1}{2}}\rceil$, then when $T[\kappa\vee\log\frac{2T}{\delta}] \geq \frac{\sigma_L^2}{2\sigma^2}$, we have:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(w_t)\|_*\|P_t\|^2 &\leq 2\left(\frac{C_1C_2}{T}\right)^{\frac{1}{2}} + 4\left(\frac{C_1C_4^2C_5}{2T}\right)^{\frac{1}{4}} \\ &\leq 4\sqrt{2}\sigma_L T^{-\frac{1}{2}} + 32\sqrt{\sigma_L\sigma}\left(\frac{2[\kappa\vee\log\frac{2T}{\delta}]}{T}\right)^{\frac{1}{4}}\end{aligned}$$

To analyze the iteration complexity, the following inequalities of T are sufficient to show $\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(w_t)\|_*\|P_t\|^2 \leq \varepsilon$ for some $\varepsilon > 0$ as 1st, 2nd terms are upper bounded by $\varepsilon/2$, and noting that $\kappa \geq 1$ and $T \geq \frac{\sigma_L^2}{2\kappa\cdot\sigma^2}$ implies that $T[\kappa\vee\log\frac{2T}{\delta}] \geq T\cdot\kappa \geq \frac{\sigma_L^2}{2\sigma^2}$:

$$T \geq \frac{\sigma_L^2}{2\kappa\cdot\sigma^2} \vee \frac{128\sigma_L^2}{\varepsilon^2} \vee \frac{2\kappa\cdot\sigma_L^2\sigma^2}{(\varepsilon/64)^4}, \frac{\log\frac{2T}{\delta}}{T} \leq \left(\frac{\varepsilon}{64\sqrt{\sigma_L\sigma}}\right)^4 = \frac{(\varepsilon/64)^4}{2\sigma_L^2\sigma^2}$$

By substituting $\delta \leftarrow \delta/2$ and $(q, \varepsilon) \leftarrow (1, \frac{(\varepsilon/64)^4}{2\sigma_L^2\sigma^2})$ in Lemma D.3, the following condition of T is sufficient to show $\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(w_t)\|_*\|P_t\|^2 \leq \varepsilon$:

$$T \geq \frac{\sigma_L^2}{2\kappa\cdot\sigma^2} \vee \frac{128\sigma_L^2}{\varepsilon^2} \vee \frac{2\sigma_L^2\sigma^2}{(\varepsilon/64)^4}\left[\kappa \vee \mathrm{e}\log\left[\mathrm{e}\vee\frac{4\sigma_L^2\sigma^2}{(\varepsilon/64)^4\delta}\right]\right]$$

In particular, for all $\varepsilon \in \mathbb{R}_+$ is sufficiently small such that $\varepsilon \leq \varepsilon_0(\sigma_L, \sigma, \kappa, \delta)$, then the lower bound on T simplifies asymptotically to

$$T \geq \frac{\mathrm{e}\cdot 2\sigma_L^2\sigma^2}{(\varepsilon/64)^4}\log\frac{4\sigma_L^2\sigma^2}{(\varepsilon/64)^4\delta},$$

where the uniform upper bound $\varepsilon_0(\sigma_L, \sigma, \kappa, \delta)$ is obtained by noting $\kappa \geq 1$ and solving the following inequalities for ε to ensure the dominated terms in the above asymptotic bound are larger than the other terms of the lower bound on T.

$$\begin{cases}\frac{4\sigma_L^2\sigma^2}{(\varepsilon/96)^4\delta} \geq \mathrm{e},\ \mathrm{e}\log\frac{4\sigma_L^2\sigma^2}{(\varepsilon/64)^4\delta} \geq \kappa \\ \frac{\mathrm{e}\cdot 2\sigma_L^2\sigma^2}{(\varepsilon/64)^4} \geq \frac{\sigma_L^2}{2\kappa\cdot\sigma^2},\ \frac{\mathrm{e}\cdot 2\sigma_L^2\sigma^2}{(\varepsilon/64)^4} \geq \frac{128\sigma_L^2}{\varepsilon^2}\end{cases}$$

The explicit expression of $\varepsilon_0(\sigma_L,\sigma,\kappa,\delta) \equiv \sqrt{\sigma_L\sigma}\cdot\varepsilon_1(\kappa,\delta)\vee\sigma\cdot\varepsilon_2(\kappa,\delta)$ is given by:

$$\varepsilon_1(\kappa,\delta) = 96\left(\frac{4}{\mathrm{e}\delta}\right)^{1/4} \wedge 64\left(\frac{4}{\delta \mathrm{e}^{\kappa/\mathrm{e}}}\right)^{1/4}$$
$$\varepsilon_2(\kappa,\delta) = 64\sqrt{2}(\mathrm{e}\kappa)^{1/4} \wedge 512\mathrm{e}^{1/2}$$

Therefore, the iteration complexity is $T = \mathcal{O}\left(\frac{\sigma_L^2\sigma^2}{\varepsilon^4}\log\frac{\sigma_L^2\sigma^2}{\varepsilon^4\delta}\right)$, and the oracle complexity is $N = B\cdot T/E + 1\cdot(T - T/E) = T(2-1/B) = \mathcal{O}\left(\frac{\sigma_L^2\sigma^2}{\varepsilon^4}\log\frac{\sigma_L^2\sigma^2}{\varepsilon^4\delta}\right)$ with the selection of batch size $B = E$ to guarantee $\frac{1}{T}\sum_{t=1}^T\|\nabla f(w_t)\|_*\|P_t\|^2 \le \varepsilon$ with probability at least $1-\delta$ for sufficiently small ε.

Family 2, Case 1 (STORM) ($p_t = 0, \beta_t = \beta$): By invoking the established upper bound of $\|e_t\|_*$ for the estimator v_t with first-order correction term $\mathcal{T}_t$, the selection of $p_t = 0, \beta_t = \beta$ and $\eta_t = \eta$ in the third-level section of Appendix C.3.1, and noting that $\sum_{t=1}^T\beta^t = \frac{\beta(1-\beta^T)}{1-\beta} \le \frac{1}{1-\beta}$, $\frac{(1-\beta)^2}{1-\beta^2} \le 1-\beta$, $\frac{\beta^2}{1-\beta^2} \le \frac{1}{1-\beta}$ for $\beta\in(0,1)$ and the elementary inequality $\sqrt{x+y}\le\sqrt{x}+\sqrt{y}, \forall x,y\ge 0$, we show the following upper bound for $\frac{1}{T}\sum_{t=1}^T\|e_t\|_*$:

$$\frac{1}{T}\sum_{t=1}^T\|e_t\|_* \le \frac{C\left(\frac{\delta}{2T},\kappa\right)\sigma/\sqrt{B}}{(1-\beta)T} + C\left(\frac{\delta}{2T},\kappa\right)\left[\sigma\sqrt{2(1-\beta)} + \ell\eta\sqrt{\frac{2}{1-\beta}}\right]$$

By selecting the batch size $B = 1$ and noting $C\left(\frac{\delta}{2T},\kappa\right) \le 2\sqrt{2[\kappa\vee\log\frac{2T}{\delta}]}$, we have:

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^T\|\nabla f(w_t)\|_*\|P_t\|^2 \le \frac{4\Delta_f}{\eta T} + 2L\eta + 8\cdot\frac{1}{T}\sum_{t=1}^T\|e_t\|_* \\
\le\ & \frac{4\Delta_f}{\eta T} + 2L\eta + \frac{16\sigma\sqrt{2[\kappa\vee\log\frac{2T}{\delta}]}}{(1-\beta)T} + 32\sigma\sqrt{\kappa\vee\log\frac{2T}{\delta}}\sqrt{1-\beta} + 32\ell\sqrt{\kappa\vee\log\frac{2T}{\delta}}\frac{\eta}{\sqrt{1-\beta}}
\end{aligned}$$

Letting $C_1 = 4\Delta_f, C_2 = 2L, C_3 = 16\sigma\sqrt{2[\kappa\vee\log\frac{2T}{\delta}]}, C_4 = 32\sigma\sqrt{\kappa\vee\log\frac{2T}{\delta}}, C_5 = 32\ell\sqrt{\kappa\vee\log\frac{2T}{\delta}}$ in Lemma D.2, selecting the step size and the momentum parameter as $\eta = \min\left\{\left(\frac{C_1}{C_2}\right)^{\frac{1}{2}}T^{-\frac{1}{2}}, \left(\frac{C_1^2}{C_4C_5}\right)^{\frac{1}{3}}T^{-\frac{2}{3}}\right\} = \min\left\{\frac{\sqrt{2}}{L}\sigma_L T^{-\frac{1}{2}}, \frac{1}{4\ell}\left(\frac{\sigma_\ell^4}{\sigma}\right)^{\frac{1}{3}}[\kappa\vee\log\frac{2T}{\delta}]^{-\frac{1}{3}}T^{-\frac{2}{3}}\right\}$, $1-\beta = \max\left\{\left(\frac{C_3}{C_4T}\right)^{\frac{2}{3}}, \left(\frac{C_1C_5}{C_4^2T}\right)^{\frac{2}{3}}\right\} = \max\left\{2^{-\frac{1}{3}}T^{-\frac{2}{3}}, \frac{1}{4}\left(\frac{(\sigma_\ell/\sigma)^4}{[\kappa\vee\log\frac{2T}{\delta}]}\right)^{\frac{1}{3}}T^{-\frac{2}{3}}\right\}$ then when $T[\kappa\vee\log\frac{2T}{\delta}]^{\frac{1}{2}} \ge \frac{\sigma_\ell^2}{8\sigma^2}$, we have:

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^T\|\nabla f(w_t)\|_*\|P_t\|^2 \le 2\left(\frac{C_1C_2}{T}\right)^{\frac{1}{2}} + 2\left(\frac{C_3C_4^2}{T}\right)^{\frac{1}{3}} + 3\left(\frac{C_1C_4C_5}{T}\right)^{\frac{1}{3}} \\
\le\ & 4\sqrt{2}\sigma_L T^{-\frac{1}{2}} + 32\sqrt{2}\sigma\left(\frac{2[\kappa\vee\log\frac{2T}{\delta}]^{\frac{3}{2}}}{T}\right)^{\frac{1}{3}} + 48(\sigma_\ell^2\sigma)^{\frac{1}{3}}\left(\frac{\kappa\vee\log\frac{2T}{\delta}}{T}\right)^{\frac{1}{3}}
\end{aligned}$$

To analyze the iteration complexity, the following inequalities of T are sufficient to show that $\frac{1}{T}\sum_{t=1}^T\|\nabla f(w_t)\|_*\|P_t\|^2 \le \varepsilon$ for some $\varepsilon > 0$ as 1st, 2nd, and 3rd terms are upper bounded by $\varepsilon/2, \varepsilon/4$ and $\varepsilon/4$ respectively, and noting that $\kappa \ge 1$ and $T \ge \frac{\sigma_\ell^2/\sigma^2}{8\sqrt{\kappa}}$ implies that $T[\kappa\vee\log\frac{2T}{\delta}]^{\frac{1}{2}} \ge T\sqrt{\kappa} \ge \frac{\sigma_\ell^2}{8\sigma^2}$.

$$T \ge \frac{\sigma_\ell^2/\sigma^2}{8\sqrt{\kappa}} \vee \frac{128\sigma_L^2}{\varepsilon^2} \vee \frac{2\kappa^{\frac{3}{2}}\cdot\sigma^3}{(\varepsilon/128\sqrt{2})^3} \vee \frac{\kappa\cdot\sigma_\ell^2\sigma}{(\varepsilon/192)^3}, \frac{\log^{\frac{3}{2}}\frac{2T}{\delta}}{T} \le \frac{1}{2}\left(\frac{\varepsilon}{128\sqrt{2}\sigma}\right)^3, \frac{\log\frac{2T}{\delta}}{T} \le \frac{(\varepsilon/192)^3}{\sigma_\ell^2\sigma}$$

By substituting $\delta \leftarrow \delta/2$ and $(q,\varepsilon) \leftarrow (\frac{3}{2}, \frac{1}{2}\left(\frac{\varepsilon}{128\sqrt{2}\sigma}\right)^3), (1, \frac{(\varepsilon/192)^3}{\sigma_\ell^2\sigma})$ in Lemma D.4, the following condition of T is sufficient to guarantee the above inequalities, and therefore $\frac{1}{T}\sum_{t=1}^T \|\nabla f(w_t)\|_*\|P_t\|^2 \le \varepsilon$.

$$T \ge \frac{\sigma_\ell^2/\sigma^2}{8\sqrt{\kappa}} \vee \frac{128\sigma_L^2}{\varepsilon^2} \vee \frac{2\sigma^3}{(\varepsilon/128\sqrt{2})^3}\left[\kappa^{\frac{3}{2}} \vee \mathrm{e}^{\frac{9}{4}}\log^{\frac{3}{2}}\left[\mathrm{e} \vee \frac{4\sigma^3}{(\varepsilon/128\sqrt{2})^3\delta}\right]\right] \vee \frac{\sigma_\ell^2\sigma}{(\varepsilon/192)^3}\left[\kappa \vee \mathrm{e}\log\left[\mathrm{e} \vee \frac{2\sigma_\ell^2\sigma}{(\varepsilon/192)^3\delta}\right]\right]$$

In particular, for all $\varepsilon \in \mathbb{R}_+$ is sufficiently small such that $\varepsilon \le \varepsilon_0(\sigma_\ell, \sigma, \kappa, \delta)$, then the lower bound on T simplifies asymptotically to

$$T \ \ge \ \max\left\{\frac{2\mathrm{e}^{\frac{9}{4}}\cdot\sigma^3}{(\varepsilon/128\sqrt{2})^3}\log^{\frac{3}{2}}\frac{4\sigma^3}{(\varepsilon/128\sqrt{2})^3\delta}, \ \frac{\mathrm{e}\cdot\sigma_\ell^2\sigma}{(\varepsilon/192)^3}\log\frac{2\sigma_\ell^2\sigma}{(\varepsilon/192)^3\delta}, \ \frac{2\sigma_L^2}{(\varepsilon/8)^2}\right\}.$$

where the uniform upper bound $\varepsilon_0(\sigma_\ell, \sigma, \kappa, \delta)$ of ε is obtained by noting $\kappa \ge 1$ and solving the following inequalities for ε to ensure the dominated terms in the above asymptotic bound are larger than the other terms of the lower bound on T.

$$\begin{cases} \dfrac{4\sigma^3}{(\varepsilon/128\sqrt{2})^3\delta} \ge \mathrm{e}, \ \dfrac{2\sigma_\ell^2\sigma}{(\varepsilon/192)^3\delta} \ge \mathrm{e} \\ \mathrm{e}\log\dfrac{2\sigma_\ell^2\sigma}{(\varepsilon/192)^3\delta} \ge \kappa, \ \mathrm{e}^{\frac{9}{4}}\log^{\frac{3}{2}}\dfrac{4\sigma^3}{(\varepsilon/128\sqrt{2})^3\delta} \ge \kappa^{\frac{3}{2}} \\ \dfrac{\mathrm{e}\cdot\sigma_\ell^2\sigma}{(\varepsilon/192)^3} \ge \dfrac{\sigma_\ell^2/\sigma^2}{8\sqrt{\kappa}} \end{cases}$$

The explicit expression of $\varepsilon_0(\sigma_\ell, \sigma, \kappa, \delta) \equiv (\sigma_\ell^2\sigma)^{\frac{1}{3}}\cdot\varepsilon_1(\kappa,\delta) \vee \sigma\cdot\varepsilon_2(\kappa,\delta)$ is given by:

$$\varepsilon_1(\kappa,\delta) = 192\left(\frac{2}{\delta\,\mathrm{e}^{1\vee(\kappa/\mathrm{e})}}\right)^{1/3}$$

$$\varepsilon_2(\kappa,\delta) = 384(\mathrm{e}\sqrt{\kappa})^{1/3} \ \wedge \ 128\sqrt{2}\left(\frac{4}{\delta\,\mathrm{e}^{1\vee(\kappa/\mathrm{e}^{3/2})}}\right)^{1/3}$$

Therefore, the iteration complexity is $T = \mathcal{O}\left(\frac{\sigma^3}{\varepsilon^3}\log^{\frac{3}{2}}\frac{\sigma^3}{\varepsilon^3\delta} \vee \frac{\sigma_\ell^2\sigma}{\varepsilon^3}\log\frac{\sigma_\ell^2\sigma}{\varepsilon^3\delta} \vee \frac{\sigma_L^2}{\varepsilon^2}\right)$, and the oracle complexity is $N = B\cdot T = \mathcal{O}\left(\frac{\sigma^3}{\varepsilon^3}\log^{\frac{3}{2}}\frac{\sigma^3}{\varepsilon^3\delta} \vee \frac{\sigma_\ell^2\sigma}{\varepsilon^3}\log\frac{\sigma_\ell^2\sigma}{\varepsilon^3\delta} \vee \frac{\sigma_L^2}{\varepsilon^2}\right)$ with the batch size $B = 1$ to guarantee $\frac{1}{T}\sum_{t=1}^T \|\nabla f(w_t)\|_*\|P_t\|^2 \le \varepsilon$ with probability at least $1-\delta$ for sufficiently small ε.

Family 2, Case 2 (PAGE / Loopless SVRG) ($p_t = p, \beta_t = 1$): By invoking the established upper bound of $\|e_t\|_*$ for the estimator v_t with first-order correction term $\mathcal{T}_t$, the selection of $p_t = p, \beta_t = 1$ and $\eta_t = \eta$ in the third-level section of Appendix C.3.2, we show the following upper bound of $\frac{1}{T}\sum_{t=1}^T \|e_t\|_*$:

$$\frac{1}{T}\sum_{t=1}^T \|e_t\|_* \le C\left(\frac{\delta}{4T}, \kappa\right)\sigma/\sqrt{B} + C\left(\frac{\delta}{4T}, \kappa\right)\ell\eta\sqrt{\frac{2}{p}\log\frac{4T}{\delta}}$$

By selecting the batch size $B = \frac{1}{p}$ and noting $C(\frac{\delta}{4T}, \kappa) \le 2\sqrt{2[\kappa \vee \log\frac{4T}{\delta}]}$, we have:

$$\begin{aligned} &\frac{1}{T}\sum_{t=1}^T \|\nabla f(w_t)\|_*\|P_t\|^2 \le \frac{4\Delta_f}{\eta T} + 2L\eta + 8\cdot\frac{1}{T}\sum_{t=1}^T \|e_t\|_* \\ \le \ &\frac{4\Delta_f}{\eta T} + 2L\cdot\eta + 0 + 16\sigma\sqrt{2[\kappa \vee \log\frac{4T}{\delta}]}\cdot\sqrt{p} + 32\ell\sqrt{\log\frac{4T}{\delta}[\kappa \vee \log\frac{4T}{\delta}]}\cdot\frac{\eta}{\sqrt{p}} \end{aligned}$$

Letting $C_1 = 4\Delta_f, C_2 = 2L, C_3 = 0, C_4 = 16\sigma\sqrt{2[\kappa \vee \log\frac{4T}{\delta}]}, C_5 = 32\ell\sqrt{\log\frac{4T}{\delta}[\kappa \vee \log\frac{4T}{\delta}]}$ in Lemma D.3, selecting the step size as $\eta = \min\left\{\left(\frac{C_1}{C_2T}\right)^{\frac{1}{2}}, \left(\frac{C_1^2}{\sqrt{2}C_4C_5T^2}\right)^{\frac{1}{3}}\right\} = \min\left\{\frac{\sqrt{2}}{L}\sigma_L T^{-\frac{1}{2}},\right.$

$\frac{1}{4\ell}\left(\frac{\sigma_\ell^4}{\sigma}\right)^{\frac{1}{3}}(\log\frac{4T}{\delta}[\kappa\vee\log\frac{4T}{\delta}]^2)^{-\frac{1}{6}}T^{-\frac{2}{3}}\Big\}$, reset probability such that $\frac{1}{p}=B=\left\lceil\left(\frac{C_4^4T^2}{2C_1^2C_5^2}\right)^{\frac{1}{3}}\right\rceil=\left\lceil 2\left(\frac{\sigma}{\sigma_\ell}\right)^{\frac{4}{3}}\left(\frac{\kappa\vee\log\frac{4T}{\delta}}{\log\frac{4T}{\delta}}\right)^{\frac{1}{3}}T^{\frac{2}{3}}\right\rceil$ when $T\left(\frac{\kappa\vee\log\frac{4T}{\delta}}{\log\frac{4T}{\delta}}\right)^{\frac{1}{2}}\geq\frac{\sigma_\ell^2}{2\sqrt{2}\sigma^2}$, we have:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(w_t)\|_*\|P_t\|^2 &\leq 2\left(\frac{C_1C_2}{T}\right)^{\frac{1}{2}}+3\left(\frac{\sqrt{2}C_1C_4C_5}{T}\right)^{\frac{1}{3}}\\ &= 4\sqrt{2}\sigma_LT^{-\frac{1}{2}}+48(\sigma_\ell^2\sigma)^{\frac{1}{3}}\left(\frac{[\log\frac{4T}{\delta}]^{\frac{1}{2}}\cdot[\kappa\vee\log\frac{4T}{\delta}]}{T}\right)^{\frac{1}{3}}\end{aligned}$$

To analyze the iteration complexity, the following inequalities of T are sufficient to show that $\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(w_t)\|_*\|P_t\|^2\leq\varepsilon$ for some $\varepsilon>0$ as 1st, 2nd terms all are upper bounded by $\varepsilon/2$, and noting $\kappa\geq 1$ and $T\geq\frac{\sigma_\ell^2}{2\sqrt{2}\sigma^2}$ implies that $T\left(\frac{\kappa\vee\log\frac{4T}{\delta}}{\log\frac{4T}{\delta}}\right)^{\frac{1}{2}}\geq T\geq\frac{\sigma_\ell^2}{2\sqrt{2}\sigma^2}$:

$$\begin{aligned}T\geq&\frac{\sigma_\ell^2}{2\sqrt{2}\sigma^2}\vee\frac{128\sigma_L^2}{\varepsilon^2},\\ &\frac{\log^{\frac{1}{2}}\frac{4T}{\delta}}{T}\leq\frac{1}{\kappa}\cdot\frac{(\varepsilon/96)^3}{\sigma_\ell^2\sigma},\ \frac{\log^{\frac{3}{2}}\frac{4T}{\delta}}{T}\leq\frac{(\varepsilon/96)^3}{\sigma_\ell^2\sigma}\end{aligned}$$

By substituting $\delta\leftarrow\delta/4$ and $(q,\varepsilon)\leftarrow(\frac{1}{2},\frac{1}{\kappa}\cdot\frac{(\varepsilon/96)^3}{\sigma_\ell^2\sigma}),(\frac{3}{2},\frac{(\varepsilon/96)^3}{\sigma_\ell^2\sigma})$ in Lemma D.4, the following condition of T is sufficient to guarantee the above inequalities, and therefore $\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(w_t)\|_*\|P_t\|^2\leq\varepsilon$.

$$T\geq\frac{\sigma_\ell^2}{2\sqrt{2}\sigma^2}\vee\frac{128\sigma_L^2}{\varepsilon^2}\vee\frac{\sigma_\ell^2\sigma}{(\varepsilon/96)^3}\left[\mathrm{e}^{\frac{9}{4}}\log^{\frac{3}{2}}\left[\mathrm{e}\vee\frac{4\sigma_\ell^2\sigma}{(\varepsilon/96)^3\delta}\right]\vee\mathrm{e}^{\frac{1}{4}}\kappa\log^{\frac{1}{2}}\left[\mathrm{e}\vee\kappa\cdot\frac{4\sigma_\ell^2\sigma}{(\varepsilon/96)^3\delta}\right]\right]$$

In particular, for all $\varepsilon\in\mathbb{R}_+$ is sufficiently small such $\varepsilon\leq\varepsilon_0(\sigma_\ell,\sigma,\kappa,\delta)$, then the lower bound on T simplifies asymptotically to

$$T\geq\max\left\{\frac{\mathrm{e}^{\frac{9}{4}}\cdot\sigma_\ell^2\sigma}{(\varepsilon/96)^3}\log^{\frac{3}{2}}\frac{4\sigma_\ell^2\sigma}{(\varepsilon/96)^3\delta},\frac{2\sigma_L^2}{(\varepsilon/8)^2}\right\}$$

where the uniform upper bound $\varepsilon_0(\sigma_\ell,\sigma,\kappa,\delta)$ of ε is obtained by noting $\kappa\geq 1$ and solving the following inequalities for ε to ensure the dominated terms in the above asymptotic bound are larger than the other terms of the lower bound on T.

$$\begin{cases}\frac{4\sigma_\ell^2\sigma}{(\varepsilon/96)^3\delta}\geq\mathrm{e},\ \frac{4\sigma_\ell^2\sigma}{(\varepsilon/96)^3\delta}\geq\kappa\\ \frac{\mathrm{e}^{\frac{9}{4}}}{\sqrt{2}\mathrm{e}^{\frac{1}{4}}}\log\frac{4\sigma_\ell^2\sigma}{(\varepsilon/96)^3\delta}\geq\kappa\\ \frac{\mathrm{e}^{\frac{9}{4}}\cdot\sigma_\ell^2\sigma}{(\varepsilon/96)^3}\geq\frac{\sigma_\ell^2}{2\sqrt{2}\sigma^2}\end{cases}$$

The explicit expression of $\varepsilon_0(\sigma_\ell,\sigma,\kappa,\delta)\equiv(\sigma_\ell^2\sigma)^{\frac{1}{3}}\cdot\varepsilon_1(\kappa,\delta)\vee\sigma\cdot\varepsilon_2(\kappa,\delta)$ is given by:

$$\begin{aligned}\varepsilon_1(\kappa,\delta)&=96\left(\frac{4}{\delta\left(\mathrm{e}\vee\kappa\vee\mathrm{e}^{\sqrt{2}\kappa/\mathrm{e}^2}\right)}\right)^{1/3}\\ \varepsilon_2(\kappa,\delta)&=96\sqrt{2}\mathrm{e}^{3/4}\end{aligned}$$

Therefore, the iteration complexity is $T=\mathcal{O}\left(\frac{\sigma_\ell^2\sigma}{\varepsilon^3}\log^{\frac{3}{2}}\frac{\sigma_\ell^2\sigma}{\varepsilon^3\delta}\vee\frac{\sigma_L^2}{\varepsilon^2}\right)$, and the expected value of oracle complexity is $N=[B\cdot p+1\cdot(1-p)]T=T(2-1/B)=\mathcal{O}\left(\frac{\sigma_\ell^2\sigma}{\varepsilon^3}\log^{\frac{3}{2}}\frac{\sigma_\ell^2\sigma}{\varepsilon^3\delta}\vee\frac{\sigma_L^2}{\varepsilon^2}\right)$ with the

selection of batch size $B = \frac{1}{p}$ to guarantee $\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(w_t)\|_* \|P_t\|^2 \leq \varepsilon$ with probability at least $1-\delta$ for sufficiently small ε.

Family 2, Case 3 (SPIDER) ($p_t = \mathbb{I}_{\{t \bmod E=0\}}, \beta_t = 1$): By invoking the established upper bound of $\|e_t\|_*$ for the estimator v_t with first-order correction term $\mathcal{T}_t$, the selection of $p_t = \mathbb{I}_{\{t \bmod E=0\}}, \beta_t = 1$ and $\eta_t = \eta$ in the third-level section of Appendix C.3.3, we show the following upper bound of $\frac{1}{T}\sum_{t=1}^{T} \|e_t\|_*$:

$$\frac{1}{T}\sum_{t=1}^{T} \|e_t\|_* \leq C\left(\frac{\delta}{2T/E}, \kappa\right)\sigma/\sqrt{B} + C\left(\frac{\delta}{2T}, \kappa\right)\ell\eta\sqrt{2E}$$

By selecting the batch size $B = E$, noting $C\left(\frac{\delta}{2T/E}, \kappa\right) \leq C\left(\frac{\delta}{2T}, \kappa\right) \leq 2\sqrt{2[\kappa \vee \log\frac{2T}{\delta}]}$, we have:

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(w_t)\|_* \|P_t\|^2 \leq \frac{4\Delta_f}{\eta T} + 2L\eta + 8 \cdot \frac{1}{T}\sum_{t=1}^{T} \|e_t\|_* \\
\leq\ & \frac{4\Delta_f}{\eta T} + 2L \cdot \eta + 0 + 16\sigma\sqrt{2[\kappa \vee \log\frac{2T}{\delta}]} \cdot \frac{1}{\sqrt{E}} + 32\ell\sqrt{\kappa \vee \log\frac{2T}{\delta}} \cdot \eta\sqrt{E}
\end{aligned}$$

Letting $C_1 = 4\Delta_f, C_2 = 2L, C_3 = 0, C_4 = 16\sigma\sqrt{2[\kappa \vee \log\frac{2T}{\delta}]}, C_5 = 32\ell\sqrt{\kappa \vee \log\frac{2T}{\delta}}$ in Lemma D.3, selecting the step size as $\eta = \min\left\{\left(\frac{C_1}{C_2 T}\right)^{\frac{1}{2}}, \left(\frac{C_1^2}{\sqrt{2}C_4 C_5 T^2}\right)^{\frac{1}{3}}\right\} = \min\left\{\frac{\sqrt{2}}{L}\sigma_L T^{-\frac{1}{2}}, \frac{1}{4\ell}\left(\frac{\sigma_\ell^4}{\sigma}\right)^{\frac{1}{3}}[\kappa \vee \log\frac{2T}{\delta}]^{-\frac{1}{3}} T^{-\frac{2}{3}}\right\}$, reset epoch as $E = B = \left\lceil\left(\frac{C_4^4 T^2}{2C_1^2 C_5^2}\right)^{\frac{1}{3}}\right\rceil = \left\lceil 2\left(\frac{\sigma}{\sigma_\ell}\right)^{\frac{4}{3}}[\kappa \vee \log\frac{2T}{\delta}]^{\frac{1}{3}} T^{\frac{2}{3}}\right\rceil$ when $T[\kappa \vee \log\frac{2T}{\delta}]^{\frac{1}{2}} \geq \frac{\sigma_\ell^2}{2\sqrt{2}\sigma^2}$, we have:

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(w_t)\|_* \|P_t\|^2 \leq 2\left(\frac{C_1 C_2}{T}\right)^{\frac{1}{2}} + 3\left(\frac{\sqrt{2}C_1 C_4 C_5}{T}\right)^{\frac{1}{3}} \\
=\ & 4\sqrt{2}\sigma_L T^{-\frac{1}{2}} + 48(\sigma_\ell^2\sigma)^{\frac{1}{3}}\left(\frac{\kappa \vee \log\frac{2T}{\delta}}{T}\right)^{\frac{1}{3}}
\end{aligned}$$

To analyze the iteration complexity, the following inequalities of T are sufficient to show that $\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(w_t)\|_* \|P_t\|^2 \leq \varepsilon$ for some $\varepsilon > 0$ as 1st, 2nd terms all are upper bounded by $\varepsilon/2$, and noting that $T \geq \frac{\sigma_\ell^2}{2\sqrt{2\kappa}\sigma^2}$ implies that $T[\kappa \vee \log\frac{2T}{\delta}]^{\frac{1}{2}} \geq T\kappa^{\frac{1}{2}} \geq \frac{\sigma_\ell^2}{2\sqrt{2}\sigma^2}$:

$$\begin{aligned}
T \geq & \frac{\sigma_\ell^2}{2\sqrt{2\kappa}\sigma^2} \vee \frac{128\sigma_L^2}{\varepsilon^2} \vee \frac{\kappa \cdot \sigma_\ell^2\sigma}{(\varepsilon/96)^3}, \\
& \frac{\log\frac{2T}{\delta}}{T} \leq \frac{(\varepsilon/96)^3}{\sigma_\ell^2\sigma}
\end{aligned}$$

By substituting $\delta \leftarrow \delta/2$ and $(q, \varepsilon) \leftarrow (1, \frac{(\varepsilon/96)^3}{\sigma_\ell^2\sigma})$ in Lemma D.4, the following condition of T is sufficient to guarantee the above inequalities, and therefore $\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(w_t)\|_* \|P_t\|^2 \leq \varepsilon$.

$$T \geq \frac{\sigma_\ell^2}{2\sqrt{2\kappa}\sigma^2} \vee \frac{128\sigma_L^2}{\varepsilon^2} \vee \frac{\sigma_\ell^2\sigma}{(\varepsilon/96)^3}\left[\kappa \vee \mathrm{e}\log\left[\mathrm{e} \vee \frac{2\sigma_\ell^2\sigma}{(\varepsilon/96)^3\delta}\right]\right]$$

In particular, for all $\varepsilon \in \mathbb{R}_+$ is sufficiently small such $\varepsilon \leq \varepsilon_0(\sigma_\ell, \sigma, \kappa, \delta)$, then the lower bound on T simplifies asymptotically to

$$T \geq \max\left\{\frac{\mathrm{e} \cdot \sigma_\ell^2\sigma}{(\varepsilon/96)^3}\log\frac{2\sigma_\ell^2\sigma}{(\varepsilon/96)^3\delta}, \frac{2\sigma_L^2}{(\varepsilon/8)^2}\right\}$$

where the uniform upper bound $\varepsilon_0(\sigma_\ell, \sigma, \kappa, \delta)$ of ε is obtained by solving the following inequalities for ε to ensure the dominated terms in the above asymptotic bound are larger than the other terms of the lower bound on T.

$$\begin{cases} \dfrac{2\sigma_\ell^2\sigma}{(\varepsilon/96)^3\delta} \geq \mathrm{e}, \\ \mathrm{e}\log\dfrac{2\sigma_\ell^2\sigma}{(\varepsilon/96)^3\delta} \geq \kappa, \\ \dfrac{\mathrm{e}\cdot\sigma_\ell^2\sigma}{(\varepsilon/96)^3} \geq \dfrac{1}{2\sqrt{2\kappa}}\dfrac{\sigma_\ell^2}{\sigma^2} \end{cases}$$

The explicit expression of $\varepsilon_0(\sigma_\ell, \sigma, \kappa, \delta) \equiv (\sigma_\ell^2\sigma)^{\frac{1}{3}} \cdot \varepsilon_1(\kappa, \delta) \vee \sigma \cdot \varepsilon_2(\kappa, \delta)$ is given by:

$$\varepsilon_1(\kappa, \delta) = 96\left(\frac{2}{\delta\,\mathrm{e}^{1\vee(\kappa/\mathrm{e})}}\right)^{1/3}$$
$$\varepsilon_2(\kappa, \delta) = 96\sqrt{2}(\mathrm{e}\sqrt{\kappa})^{1/3}$$

Therefore, the iteration complexity is $T = \mathcal{O}\left(\frac{\sigma_\ell^2\sigma}{\varepsilon^3}\log\frac{\sigma_\ell^2\sigma}{\varepsilon^3\delta} \vee \frac{\sigma_L^2}{\varepsilon^2}\right)$, and the oracle complexity is $N = B\cdot T/E + 1\cdot(T - T/E) = T(2 - 1/B) = \mathcal{O}\left(\frac{\sigma_\ell^2\sigma}{\varepsilon^3}\log\frac{\sigma_\ell^2\sigma}{\varepsilon^3\delta} \vee \frac{\sigma_L^2}{\varepsilon^2}\right)$ with the selection of batch size $B = E$ to guarantee $\frac{1}{T}\sum_{t=1}^T \|\nabla f(w_t)\|_*\|P_t\|^2 \leq \varepsilon$ with probability at least $1 - \delta$ for sufficiently small ε.

Family 3, Case 1 (Second-Order Momentum) ($p_t = 0, \beta_t = \beta$): By invoking the established upper bound of $\|e_t\|_*$ for the estimator v_t with second-order correction term $\mathcal{T}_t$, the selection of $p_t = 0, \beta_t = \beta$ and $\eta_t = \eta$ in the third-level section of Appendix C.4.1, and noting that $\sum_{t=1}^T \beta^t = \frac{\beta(1-\beta^T)}{1-\beta} \leq \frac{1}{1-\beta}, \beta \leq 1, \frac{(1-\beta)^2}{1-\beta^2} \leq 1-\beta, \frac{\beta^2}{1-\beta^2} \leq \frac{1}{1-\beta}$ for $\beta \in (0,1)$ and the elementary inequality $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}, \forall x, y \geq 0$, we show the following upper bound for $\frac{1}{T}\sum_{t=1}^T \|e_t\|_*$:

$$\frac{1}{T}\sum_{t=1}^T \|e_t\|_* \leq \frac{C\left(\frac{\delta}{2T}, \kappa\right)\sigma/\sqrt{B}}{(1-\beta)T} + \frac{\alpha\eta^2}{2(1-\beta)} + C\left(\frac{\delta}{2T}, \kappa\right)\left[\sigma\sqrt{2(1-\beta)} + \gamma\eta\sqrt{\frac{2}{1-\beta}}\right]$$

By selecting the batch size $B = 1$ and noting $C\left(\frac{\delta}{2T}, \kappa\right) \leq 2\sqrt{2[\kappa \vee \log\frac{2T}{\delta}]}$, we have:

$$\begin{aligned} &\frac{1}{T}\sum_{t=1}^T \|\nabla f(w_t)\|_*\|P_t\|^2 \leq \frac{4\Delta_f}{\eta T} + 2L\eta + 8\cdot\frac{1}{T}\sum_{t=1}^T \|e_t\|_* \\ \leq\ & \frac{4\Delta_f}{\eta T} + 2L\cdot\eta + \frac{16\sigma\sqrt{2[\kappa \vee \log\frac{2T}{\delta}]}}{(1-\beta)T} + 32\sigma\sqrt{\kappa \vee \log\frac{2T}{\delta}}\cdot\sqrt{1-\beta} \\ & + 32\gamma\sqrt{\kappa \vee \log\frac{2T}{\delta}}\cdot\frac{\eta}{\sqrt{1-\beta}} + 4\alpha\cdot\frac{\eta^2}{1-\beta} \end{aligned}$$

Subsequently, by letting $C_1 = 4\Delta_f, C_2 = 2L, C_3 = 16\sigma\sqrt{2[\kappa \vee \log\frac{2T}{\delta}]}, C_4 = 32\sigma\sqrt{\kappa \vee \log\frac{2T}{\delta}}, C_5 = 32\gamma\sqrt{\kappa \vee \log\frac{2T}{\delta}}, C_6 = 4\alpha$ in Lemma D.2, and selecting the step size η such that $\eta = \min\left\{\left(\frac{C_1}{C_2}\right)^{\frac{1}{2}}T^{-\frac{1}{2}}, \left(\frac{C_1^2}{C_4C_5}\right)^{\frac{1}{3}}T^{-\frac{2}{3}}, \left(\frac{C_1^3}{C_4^2C_6}\right)^{\frac{1}{5}}T^{-\frac{3}{5}}\right\} = \min\left\{\frac{\sqrt{2}}{L}\sigma_L T^{-\frac{1}{2}}, \frac{1}{4\gamma}\left(\frac{\sigma_\gamma^4}{\sigma}\right)^{\frac{1}{3}}[\kappa \vee \log\frac{2T}{\delta}]^{-\frac{1}{3}}T^{-\frac{2}{3}}, \frac{1}{2}\sqrt{\frac{1}{\alpha}\left(\frac{\sigma_\alpha^9}{4\sigma^4}\right)^{\frac{1}{5}}}[\kappa \vee \log\frac{2T}{\delta}]^{-\frac{1}{5}}T^{-\frac{3}{5}}\right\}$, the momentum parameter β as $1-\beta = \max\left\{\left(\frac{C_3}{C_4}\right)^{\frac{2}{3}}T^{-\frac{2}{3}}, \left(\frac{C_1C_5}{C_4^2}\right)^{\frac{2}{3}}T^{-\frac{2}{3}}, \left(\frac{C_1^2C_6}{C_4^3}\right)^{\frac{2}{5}}T^{-\frac{4}{5}}\right\} = \max\left\{2^{-\frac{1}{3}}T^{-\frac{2}{3}}, \frac{1}{4}\left(\frac{(\sigma_\gamma/\sigma)^4}{[\kappa\vee\log\frac{2T}{\delta}]}\right)^{\frac{1}{3}}T^{-\frac{2}{3}}, \frac{1}{8}\left(\frac{\sigma_\alpha}{\sqrt{2}\sigma}\right)^{\frac{6}{5}}[\kappa \vee \log\frac{2T}{\delta}]^{-\frac{3}{5}}T^{-\frac{4}{5}}\right\}$, then when $T[\kappa \vee \log\frac{2T}{\delta}]^{\frac{1}{2}} \geq \frac{\sigma_\gamma^2}{8\sigma^2}$ and

$T[\kappa \vee \log \frac{2T}{\delta}]^{\frac{3}{4}} \geq 2^{-\frac{9}{2}} \left(\frac{\sigma_\alpha}{\sigma}\right)^{\frac{3}{2}}$, we have:

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(w_t)\|_* \|P_t\|^2 \leq 2\left(\frac{C_1 C_2}{T}\right)^{\frac{1}{2}} + 2\left(\frac{C_3 C_4^2}{T}\right)^{\frac{1}{3}} + 3\left(\frac{C_1 C_4 C_5}{T}\right)^{\frac{1}{3}} + 3\left(\frac{C_1^2 C_4^2 C_6}{T^2}\right)^{\frac{1}{5}} \\
\leq\ & 4\sqrt{2}\sigma_L T^{-\frac{1}{2}} + 32\sqrt{2}\sigma\left(\frac{2[\kappa \vee \log \frac{2T}{\delta}]^{\frac{3}{2}}}{T}\right)^{\frac{1}{3}} + 48(\sigma_\gamma^2\sigma)^{\frac{1}{3}}\left(\frac{\kappa \vee \log \frac{2T}{\delta}}{T}\right)^{\frac{1}{3}} \\
&+ 24(\sigma_\alpha^3\sigma^2)^{\frac{1}{5}}\left(\frac{2[\kappa \vee \log \frac{2T}{\delta}]}{T^2}\right)^{\frac{1}{5}}
\end{aligned}$$

To analyze the iteration complexity, the following inequalities of T are sufficient to show that $\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(w_t)\|_* \|P_t\|^2 \leq \varepsilon$ for some $\varepsilon > 0$ as 1st, 2nd, 3rd and 4th terms all are upper bounded by $\varepsilon/4$, and noting that $\kappa \geq 1$ and $T \geq \frac{\sigma_\gamma^2/\sigma^2}{8\sqrt{\kappa}} \vee 2^{-\frac{9}{2}} \frac{(\sigma_\alpha/\sigma)^{\frac{3}{2}}}{\kappa^{\frac{3}{4}}}$ implies that $T[\kappa \vee \log \frac{2T}{\delta}]^{\frac{1}{2}} \geq T\kappa^{\frac{1}{2}} \geq \frac{\sigma_\gamma^2}{8\sigma^2}$ and $T[\kappa \vee \log \frac{2T}{\delta}]^{\frac{3}{4}} \geq T\kappa^{\frac{3}{4}} \geq 2^{-\frac{9}{2}}\left(\frac{\sigma_\alpha}{\sigma}\right)^{\frac{3}{2}}$:

$$\begin{aligned}
&T \geq \frac{\sigma_\gamma^2/\sigma^2}{8\sqrt{\kappa}} \vee 2^{-\frac{9}{2}}\frac{(\sigma_\alpha/\sigma)^{\frac{3}{2}}}{\kappa^{\frac{3}{4}}} \vee \frac{512\sigma_L^2}{\varepsilon^2} \vee \frac{2\kappa^{\frac{3}{2}}\cdot\sigma^3}{(\varepsilon/128\sqrt{2})^3} \vee \frac{\kappa\cdot\sigma_\gamma^2\sigma}{(\varepsilon/192)^3} \vee \sqrt{2\kappa\frac{\sigma_\alpha^3\sigma^2}{(\varepsilon/96)^5}}, \\
&\frac{\log^{\frac{3}{2}}\frac{2T}{\delta}}{T} \leq \frac{1}{2}\left(\frac{\varepsilon}{128\sqrt{2}\sigma}\right)^3, \frac{\log\frac{2T}{\delta}}{T} \leq \frac{(\varepsilon/192)^3}{\sigma_\gamma^2\sigma}, \frac{\log^{\frac{1}{2}}\frac{2T}{\delta}}{T} \leq \sqrt{\frac{(\varepsilon/96)^5}{2\sigma_\alpha^3\sigma^2}},
\end{aligned}$$

By substituting $\delta \leftarrow \delta/2$ and $(q,\varepsilon) \leftarrow (\frac{3}{2}, \frac{1}{2}\left(\frac{\varepsilon}{128\sqrt{2}\sigma}\right)^3), (1, \frac{(\varepsilon/192)^3}{\sigma_\gamma^2\sigma}), (\frac{1}{2}, \sqrt{\frac{(\varepsilon/96)^5}{2\sigma_\alpha^3\sigma^2}})$ in Lemma D.4, the following condition of T is sufficient to guarantee the above inequalities, and therefore $\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(w_t)\|_* \|P_t\|^2 \leq \varepsilon$.

$$\begin{aligned}
T \geq\ & \frac{\sigma_\gamma^2/\sigma^2}{8\sqrt{\kappa}} \vee 2^{-\frac{9}{2}}\frac{(\sigma_\alpha/\sigma)^{\frac{3}{2}}}{\kappa^{\frac{3}{4}}} \vee \frac{512\sigma_L^2}{\varepsilon^2} \\
&\vee \frac{2\sigma^3}{(\varepsilon/128\sqrt{2})^3}\left[\kappa^{\frac{3}{2}} \vee \mathrm{e}^{\frac{9}{4}}\log^{\frac{3}{2}}\left[\mathrm{e} \vee \frac{4\sigma^3}{(\varepsilon/128\sqrt{2})^3\delta}\right]\right] \\
&\vee \frac{\sigma_\gamma^2\sigma}{(\varepsilon/192)^3}\left[\kappa \vee \mathrm{e}\log\left[\mathrm{e} \vee \frac{2\sigma_\gamma^2\sigma}{(\varepsilon/192)^3\delta}\right]\right] \\
&\vee \sqrt{\frac{2\sigma_\alpha^3\sigma^2}{(\varepsilon/96)^5}}\left[\sqrt{\kappa} \vee \mathrm{e}^{\frac{1}{4}}\log^{\frac{1}{2}}\left[\mathrm{e} \vee \frac{2}{\delta}\sqrt{\frac{2\sigma_\alpha^3\sigma^2}{(\varepsilon/96)^5}}\right]\right]
\end{aligned}$$

In particular, for all $\varepsilon \in \mathbb{R}_+$ is sufficiently small such that $\varepsilon \leq \varepsilon_0(\sigma_\alpha, \sigma_\gamma, \sigma, \kappa, \delta)$, then the lower bound on T simplifies asymptotically to

$$\begin{aligned}
T \ \geq\ \max\Bigg\{ & \frac{2\mathrm{e}^{\frac{9}{4}}\cdot\sigma^3}{(\varepsilon/128\sqrt{2})^3}\log^{\frac{3}{2}}\frac{4\sigma^3}{(\varepsilon/128\sqrt{2})^3\delta}, \frac{\mathrm{e}\cdot\sigma_\gamma^2\sigma}{(\varepsilon/192)^3}\log\frac{2\sigma_\gamma^2\sigma}{(\varepsilon/192)^3\delta}, \\
& \mathrm{e}^{\frac{1}{4}}\sqrt{\frac{2\sigma_\alpha^3\sigma^2}{(\varepsilon/96)^5}}\log^{\frac{1}{2}}\left(\frac{2}{\delta}\sqrt{\frac{2\sigma_\alpha^3\sigma^2}{(\varepsilon/96)^5}}\right), \frac{2\sigma_L^2}{(\varepsilon/16)^2}\Bigg\}.
\end{aligned}$$

where the uniform upper bound $\varepsilon_0(\sigma_\alpha, \sigma_\gamma, \sigma, \kappa, \delta)$ is obtained by noting $\kappa \geq 1$ and solving the following inequalities for ε to ensure the dominated terms in the above asymptotic bound are larger than the other terms of the lower bound on T.

$$\begin{cases} \dfrac{4\sigma^3}{(\varepsilon/128\sqrt{2})^3\delta} \geq \mathrm{e},\ \dfrac{2\sigma_\gamma^2\sigma}{(\varepsilon/192)^3\delta} \geq \mathrm{e},\ \dfrac{2}{\delta}\sqrt{\dfrac{2\sigma_\alpha^3\sigma^2}{(\varepsilon/96)^5}} \geq \mathrm{e}, \\ \mathrm{e}\log\dfrac{2\sigma_\gamma^2\sigma}{(\varepsilon/192)^3\delta} \geq \kappa,\ \mathrm{e}^{\frac{1}{4}}\log^{\frac{1}{2}}\left(\dfrac{2}{\delta}\sqrt{\dfrac{2\sigma_\alpha^3\sigma^2}{(\varepsilon/96)^5}}\right) \geq \sqrt{\kappa}, \\ \mathrm{e}^{\frac{9}{4}}\log^{\frac{3}{2}}\dfrac{4\sigma^3}{(\varepsilon/128\sqrt{2})^3\delta} \geq \kappa^{\frac{3}{2}}, \\ \dfrac{\mathrm{e}\cdot\sigma_\gamma^2\sigma}{(\varepsilon/192)^3} \geq \dfrac{\sigma_\gamma^2/\sigma^2}{8\sqrt{\kappa}},\ \mathrm{e}^{\frac{1}{4}}\sqrt{\dfrac{2\sigma_\alpha^3\sigma^2}{(\varepsilon/96)^5}} \geq 2^{-\frac{9}{2}}\dfrac{(\sigma_\alpha/\sigma)^{\frac{3}{2}}}{\kappa^{\frac{3}{4}}}, \end{cases}$$

The explicit expression of

$$\varepsilon_0(\sigma_\alpha,\sigma_\gamma,\sigma,\kappa,\delta) \equiv (\sigma_\alpha^3\sigma^2)^{\frac{1}{5}}\cdot\varepsilon_1(\kappa,\delta) \vee (\sigma_\gamma^2\sigma)^{\frac{1}{3}}\cdot\varepsilon_2(\kappa,\delta) \vee \sigma\cdot\varepsilon_3(\kappa,\delta)$$

is given by:

$$\varepsilon_1(\kappa,\delta) = 96\left(\frac{8}{\delta^2\,\mathrm{e}^{2\vee(2\kappa/\sqrt{\mathrm{e}})}}\right)^{1/5}$$

$$\varepsilon_2(\kappa,\delta) = 192\left(\frac{2}{\delta\,\mathrm{e}^{1\vee(\kappa/\mathrm{e})}}\right)^{1/3}$$

$$\varepsilon_3(\kappa,\delta) = 384(\mathrm{e}\kappa^{1/2})^{1/3} \wedge 384(\mathrm{e}^{1/2}\kappa^{3/2})^{1/5} \wedge 128\sqrt{2}\left(\frac{4}{\delta\,\mathrm{e}^{1\vee(\kappa/\mathrm{e}^{3/2})}}\right)^{1/3}$$

Therefore, the iteration complexity for T and oracle complexity for $N = B\cdot T$ are given by:

$$N = B\cdot T = T = \mathcal{O}\left(\frac{\sigma^3}{\varepsilon^3}\log^{\frac{3}{2}}\frac{\sigma^3}{\varepsilon^3\delta} \vee \frac{\sigma_\gamma^2\sigma}{\varepsilon^3}\log\frac{\sigma_\gamma^2\sigma}{\varepsilon^3\delta} \vee \sqrt{\frac{\sigma_\alpha^3\sigma^2}{\varepsilon^5}}\log^{\frac{1}{2}}\left(\frac{1}{\delta}\sqrt{\frac{\sigma_\alpha^3\sigma^2}{\varepsilon^5}}\right) \vee \frac{\sigma_L^2}{\varepsilon^2}\right)$$

with the selection of batch size $B = 1$ to guarantee $\frac{1}{T}\sum_{t=1}^T \|\nabla f(w_t)\|_*\|P_t\|^2 \leq \varepsilon$ with probability at least $1-\delta$ for sufficiently small ε.

Family 3, Case 2 (Second-Order PAGE) ($p_t = p, \beta_t = 1$): By invoking the established upper bound of $\|e_t\|_*$ for the estimator v_t with second-order correction term $\mathcal{T}_t$, the selection of $p_t = p, \beta_t = 1$ and $\eta_t = \eta$ in the third-level section of Appendix C.4.2, we show the following upper bound of $\frac{1}{T}\sum_{t=1}^T \|e_t\|_*$:

$$\frac{1}{T}\sum_{t=1}^T \|e_t\|_* \leq C\left(\frac{\delta}{4T},\kappa\right)\sigma/\sqrt{B} + \frac{\alpha\eta^2}{2p}\log\frac{4T}{\delta} + C\left(\frac{\delta}{4T},\kappa\right)\gamma\eta\sqrt{\frac{2}{p}\log\frac{4T}{\delta}}$$

By selecting the batch size $B = \frac{1}{p}$, and noting $C\left(\frac{\delta}{4T},\kappa\right) \leq 2\sqrt{2[\kappa\vee\log\frac{4T}{\delta}]}$, we have:

$$\begin{aligned} &\frac{1}{T}\sum_{t=1}^T \|\nabla f(w_t)\|_*\|P_t\|^2 \leq \frac{4\Delta_f}{\eta T} + 2L\eta + 8\cdot\frac{1}{T}\sum_{t=1}^T\|e_t\|_* \\ \leq\ & \frac{4\Delta_f}{\eta T} + 2L\cdot\eta + 0 + 16\sigma\sqrt{2[\kappa\vee\log\frac{4T}{\delta}]}\cdot\sqrt{p} + 32\gamma\sqrt{\log\frac{4T}{\delta}[\kappa\vee\log\frac{4T}{\delta}]}\cdot\frac{\eta}{\sqrt{p}} \\ &+4\alpha\log\frac{4T}{\delta}\cdot\frac{\eta^2}{p} \end{aligned}$$

Therefore, by letting $C_1 = 4\Delta_f, C_2 = 2L, C_3 = 0, C_4 = 16\sigma\sqrt{2[\kappa\vee\log\frac{4T}{\delta}]}, C_5 = 32\gamma\sqrt{\log\frac{4T}{\delta}[\kappa\vee\log\frac{4T}{\delta}]}, C_6 = 4\alpha\log\frac{4T}{\delta}$ in Lemma D.3, selecting the step size η and the reset

probability p such that $\eta = \min\left\{ \left(\frac{C_1}{C_2 T}\right)^{\frac{1}{2}}, \left(\frac{C_1^2}{\sqrt{2}C_4C_5T^2}\right)^{\frac{1}{3}}, \left(\frac{C_1^3}{4C_4^2C_6T^3}\right)^{\frac{1}{5}} \right\} = \min\left\{ \frac{\sqrt{2}}{L}\sigma_L T^{-\frac{1}{2}}, \right.$ $\left. \frac{1}{4\gamma}\left(\frac{\sigma_\gamma^4}{\sigma}\right)^{\frac{1}{3}} (\log\frac{4T}{\delta}[\kappa \vee \log\frac{4T}{\delta}]^2)^{-\frac{1}{6}} T^{-\frac{2}{3}}, \frac{1}{2}\sqrt{\frac{1}{2\alpha}\left(\frac{2\sigma_\alpha^9}{\sigma^4}\right)^{\frac{1}{5}}} [\log\frac{4T}{\delta}]^{-\frac{1}{5}} [\kappa \vee \log\frac{4T}{\delta}]^{-\frac{1}{5}} T^{-\frac{3}{5}} \right\}$, the reset probability p and batch size B as $\frac{1}{p} = B = \left\lceil \min\left\{ \left(\frac{C_4^4T^2}{2C_1^2C_5^2}\right)^{\frac{1}{3}}, \left(\frac{C_4^6T^4}{16C_1^4C_6^2}\right)^{\frac{1}{5}} \right\} \right\rceil = \left\lceil \min\left\{ 2\left(\frac{\sigma}{\sigma_\gamma}\right)^{\frac{4}{3}} \left(\frac{\kappa \vee \log\frac{4T}{\delta}}{\log\frac{4T}{\delta}}\right)^{\frac{1}{3}} T^{\frac{2}{3}}, 4\left(\frac{2\sigma^6}{\sigma_\alpha^6}\right)^{\frac{1}{5}} [\kappa \vee \log\frac{4T}{\delta}]^{\frac{1}{5}} \left(\frac{\kappa \vee \log\frac{4T}{\delta}}{\log\frac{4T}{\delta}}\right)^{\frac{2}{5}} T^{\frac{4}{5}} \right\} \right\rceil$ when $T\left(\frac{\kappa \vee \log\frac{4T}{\delta}}{\log\frac{4T}{\delta}}\right)^{\frac{1}{2}} \geq \frac{\sigma_\gamma^2}{2\sqrt{2}\sigma^2}$ and $T[\kappa \vee \log\frac{4T}{\delta}]^{\frac{1}{4}} \left(\frac{\kappa \vee \log\frac{4T}{\delta}}{\log\frac{4T}{\delta}}\right)^{\frac{1}{2}} \geq 2^{-\frac{11}{4}} \left(\frac{\sigma_\alpha}{\sigma}\right)^{\frac{3}{2}}$, we have:

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(w_t)\|_* \|P_t\|^2 \leq 2\left(\frac{C_1C_2}{T}\right)^{\frac{1}{2}} + 3\left(\frac{\sqrt{2}C_1C_4C_5}{T}\right)^{\frac{1}{3}} + 5\left(\frac{C_1^2C_4^2C_6}{8T^2}\right)^{\frac{1}{5}} \\
= \;& 4\sqrt{2}\sigma_L T^{-\frac{1}{2}} + 48(\sigma_\gamma^2\sigma)^{\frac{1}{3}}\left(\frac{[\log\frac{4T}{\delta}]^{\frac{1}{2}}[\kappa \vee \log\frac{4T}{\delta}]}{T}\right)^{\frac{1}{3}} + 20(\sigma_\alpha^3\sigma^2)^{\frac{1}{5}}\left(\frac{2[\log\frac{4T}{\delta}]^{\frac{1}{2}}[\kappa \vee \log\frac{4T}{\delta}]^{\frac{1}{2}}}{T}\right)^{\frac{2}{5}}
\end{aligned}$$

To analyze the iteration complexity, the following inequalities of T are sufficient to show that $\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(w_t)\|_* \|P_t\|^2 \leq \varepsilon$ for some $\varepsilon > 0$ as 1st, 2nd and 3rd terms are upper bounded by $\varepsilon/2, \varepsilon/4$ and $\varepsilon/4$ respectively, noting that $T \geq \frac{\sigma_\gamma^2}{2\sqrt{2}\sigma^2} \vee \frac{2^{-\frac{11}{4}}}{\kappa^{\frac{1}{4}}}\left(\frac{\sigma_\alpha}{\sigma}\right)^{\frac{3}{2}}$ implies $T\left(\frac{\kappa \vee \log\frac{4T}{\delta}}{\log\frac{4T}{\delta}}\right)^{\frac{1}{2}} \geq T \geq \frac{\sigma_\gamma^2}{2\sqrt{2}\sigma^2}$ and $T[\kappa \vee \log\frac{4T}{\delta}]^{\frac{1}{4}}\left(\frac{\kappa \vee \log\frac{4T}{\delta}}{\log\frac{4T}{\delta}}\right)^{\frac{1}{2}} \geq T \cdot \kappa^{\frac{1}{4}} \geq 2^{-\frac{11}{4}}\left(\frac{\sigma_\alpha}{\sigma}\right)^{\frac{3}{2}}$:

$$\begin{aligned}
T \geq& \frac{\sigma_\gamma^2}{2\sqrt{2}\sigma^2} \vee \frac{2^{-\frac{11}{4}}}{\kappa^{\frac{1}{4}}}\left(\frac{\sigma_\alpha}{\sigma}\right)^{\frac{3}{2}} \vee \frac{128\sigma_L^2}{\varepsilon^2}, \\
& \frac{\log^{\frac{1}{2}}\frac{4T}{\delta}}{T} \leq \frac{(\varepsilon/192)^3}{\kappa \cdot \sigma_\gamma^2\sigma}, \frac{\log^{\frac{3}{2}}\frac{4T}{\delta}}{T} \leq \frac{(\varepsilon/192)^3}{\sigma_\gamma^2\sigma}, \frac{\log^{\frac{1}{2}}\frac{4T}{\delta}}{T} \leq \frac{1}{2}\sqrt{\frac{(\varepsilon/80)^5}{\kappa\sigma_\alpha^3\sigma^2}}, \frac{\log\frac{4T}{\delta}}{T} \leq \frac{1}{2}\sqrt{\frac{(\varepsilon/80)^5}{\sigma_\alpha^3\sigma^2}}
\end{aligned}$$

By substituting $\delta \leftarrow \delta/4$ and $(q, \varepsilon) \leftarrow (\frac{1}{2}, \frac{(\varepsilon/192)^3}{\kappa\cdot\sigma_\gamma^2\sigma}), (\frac{3}{2}, \frac{(\varepsilon/192)^3}{\sigma_\gamma^2\sigma}), (\frac{1}{2}, \frac{1}{2}\sqrt{\frac{(\varepsilon/80)^5}{\kappa\cdot\sigma_\alpha^3\sigma^2}}), (1, \frac{1}{2}\sqrt{\frac{(\varepsilon/80)^5}{\sigma_\alpha^3\sigma^2}})$ in Lemma D.4, the following condition of T is sufficient to guarantee the above inequalities, and therefore $\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(w_t)\|_* \|P_t\|^2 \leq \varepsilon$.

$$\begin{aligned}
T \geq& \frac{\sigma_\gamma^2}{2\sqrt{2}\sigma^2} \vee \frac{2^{-\frac{11}{4}}}{\kappa^{\frac{1}{4}}}\left(\frac{\sigma_\alpha}{\sigma}\right)^{\frac{3}{2}} \vee \frac{128\sigma_L^2}{\varepsilon^2} \\
& \vee \frac{\sigma_\gamma^2\sigma}{(\varepsilon/192)^3}\left[\mathrm{e}^{\frac{9}{4}}\log^{\frac{3}{2}}\left[\mathrm{e} \vee \frac{4\sigma_\gamma^2\sigma}{(\varepsilon/192)^3\delta}\right] \vee \mathrm{e}^{\frac{1}{4}}\kappa\log^{\frac{1}{2}}\left[\mathrm{e} \vee \kappa \cdot \frac{4\sigma_\gamma^2\sigma}{(\varepsilon/192)^3\delta}\right]\right] \\
& \vee 2\sqrt{\frac{\sigma_\alpha^3\sigma^2}{(\varepsilon/80)^5}}\left[\mathrm{e}\log\left[\mathrm{e} \vee \left(\frac{8}{\delta}\sqrt{\frac{\sigma_\alpha^3\sigma^2}{(\varepsilon/80)^5}}\right)\right] \vee \mathrm{e}^{\frac{1}{4}}\sqrt{\kappa}\log^{\frac{1}{2}}\left[\mathrm{e} \vee \sqrt{\kappa}\left(\frac{8}{\delta}\sqrt{\frac{2\sigma_\alpha^3\sigma^2}{(\varepsilon/80)^5}}\right)\right]\right]
\end{aligned}$$

In particular, for all $\varepsilon \in \mathbb{R}_+$ is sufficiently small such that $\varepsilon \leq \varepsilon_0(\sigma_\alpha, \sigma_\gamma, \sigma, \kappa, \delta)$, then the lower bound on T simplifies asymptotically to

$$T \geq \max\left\{ \frac{\mathrm{e}^{\frac{9}{4}}\sigma_\gamma^2\sigma}{(\varepsilon/192)^3}\log^{\frac{3}{2}}\frac{4\sigma_\gamma^2\sigma}{(\varepsilon/192)^3\delta}, 2\mathrm{e}\sqrt{\frac{\sigma_\alpha^3\sigma^2}{(\varepsilon/80)^5}}\log\left(\frac{8}{\delta}\sqrt{\frac{\sigma_\alpha^3\sigma^2}{(\varepsilon/80)^5}}\right), \frac{2\sigma_L^2}{(\varepsilon/8)^2} \right\}$$

where the uniform upper bound $\varepsilon_0(\sigma_\alpha, \sigma_\gamma, \sigma, \kappa, \delta)$ of ε is obtained by noting $\kappa \geq 1$ and solving the following inequalities for ε to ensure the dominated terms in the above asymptotic bound are larger than the other terms of the lower bound on T.

$$\begin{cases} \dfrac{4\sigma_\gamma^2\sigma}{(\varepsilon/192)^3\delta} \geq \max(\mathrm{e}, \kappa), \ \dfrac{8}{\delta}\sqrt{\dfrac{\sigma_\alpha^3\sigma^2}{(\varepsilon/80)^5}} \geq \max(\mathrm{e}, \sqrt{\kappa}) \\ \dfrac{\mathrm{e}^{\frac{9}{4}}}{\mathrm{e}^{\frac{1}{4}}\sqrt{2}} \log \dfrac{4\sigma_\gamma^2\sigma}{(\varepsilon/192)^3\delta} \geq \kappa, \ \dfrac{\mathrm{e}}{\mathrm{e}^{\frac{1}{4}}\sqrt{2}} \log^{\frac{1}{2}} \left(\dfrac{8}{\delta}\sqrt{\dfrac{\sigma_\alpha^3\sigma^2}{(\varepsilon/80)^5}}\right) \geq \sqrt{\kappa} \\ \dfrac{\mathrm{e}^{\frac{9}{4}} \cdot \sigma_\gamma^2\sigma}{(\varepsilon/192)^3} \geq \dfrac{\sigma_\gamma^2}{2\sqrt{2}\sigma^2}, \ 2\mathrm{e}\sqrt{\dfrac{\sigma_\alpha^3\sigma^2}{(\varepsilon/80)^5}} \geq \dfrac{2^{-\frac{11}{4}}}{\kappa^{\frac{1}{4}}} \left(\dfrac{\sigma_\alpha}{\sigma}\right)^{\frac{3}{2}} \end{cases}$$

The explicit expression of

$$\varepsilon_0(\sigma_\alpha, \sigma_\gamma, \sigma, \kappa, \delta) \equiv (\sigma_\alpha^3\sigma^2)^{\frac{1}{5}} \cdot \varepsilon_1(\kappa, \delta) \vee (\sigma_\gamma^2\sigma)^{\frac{1}{3}} \cdot \varepsilon_2(\kappa, \delta) \vee \sigma \cdot \varepsilon_3(\kappa, \delta)$$

is given by:

$$\varepsilon_1(\kappa, \delta) = 80 \left(\frac{64}{\delta^2 \left(\mathrm{e}^2 \vee \kappa \vee \mathrm{e}^{4\kappa/\mathrm{e}^{3/2}}\right)}\right)^{1/5}$$

$$\varepsilon_2(\kappa, \delta) = 192 \left(\frac{4}{\delta \left(\mathrm{e} \vee \kappa \vee \mathrm{e}^{\sqrt{2}\kappa/\mathrm{e}^2}\right)}\right)^{1/3}$$

$$\varepsilon_3(\kappa, \delta) = 192\sqrt{2}\mathrm{e}^{3/4} \ \wedge \ 160\sqrt{2}\mathrm{e}^{2/5}\kappa^{1/10}$$

Therefore, the iteration complexity for T and the expected value of oracle complexity for $N = [B \cdot p + 1 \cdot (1-p)]T = (2 - 1/B)T$ are given by:

$$N, T = \mathcal{O}\left(\frac{\sigma_\gamma^2\sigma}{\varepsilon^3} \log^{\frac{3}{2}} \frac{\sigma_\gamma^2\sigma}{\varepsilon^3\delta} \vee \sqrt{\frac{\sigma_\alpha^3\sigma^2}{\varepsilon^5}} \log\left(\frac{1}{\delta}\sqrt{\frac{\sigma_\alpha^3\sigma^2}{\varepsilon^5}}\right) \vee \frac{\sigma_L^2}{\varepsilon^2}\right)$$

with the selection of batch size $B = \frac{1}{p}$ to guarantee $\frac{1}{T}\sum_{t=1}^T \|\nabla f(w_t)\|_* \|P_t\|^2 \leq \varepsilon$ with probability at least $1 - \delta$ for sufficiently small ε.

Family 3, Case 3 (Second-Order SPIDER) ($p_t = \mathbb{I}_{\{t \bmod E = 0\}}, \beta_t = 1$): By invoking the established upper bound of $\|e_t\|_*$ for the estimator v_t with second-order correction term $\mathcal{T}_t$, the selection of $p_t = \mathbb{I}_{\{t \bmod E = 0\}}, \beta_t = 1$ and $\eta_t = \eta$ in the third-level section of Appendix C.4.3, we show the following upper bound of $\frac{1}{T}\sum_{t=1}^T \|e_t\|_*$:

$$\frac{1}{T}\sum_{t=1}^T \|e_t\|_* \leq C\left(\frac{\delta}{2T/E}, \kappa\right)\sigma/\sqrt{B} + \frac{\alpha\eta^2 E}{2} + C\left(\frac{\delta}{2T}, \kappa\right)\gamma\eta\sqrt{2E}$$

By selecting the batch size $B = E$, noting $C\left(\frac{\delta}{2T/E}, \kappa\right) \leq C\left(\frac{\delta}{2T}, \kappa\right) \leq 2\sqrt{2[\kappa \vee \log\frac{2T}{\delta}]}$, we have:

$$\begin{aligned} & \frac{1}{T}\sum_{t=1}^T \|\nabla f(w_t)\|_* \|P_t\|^2 \leq \frac{4\Delta_f}{\eta T} + 2L\eta + 8 \cdot \frac{1}{T}\sum_{t=1}^T \|e_t\|_* \\ \leq \ & \frac{4\Delta_f}{\eta T} + 2L \cdot \eta + 0 + 16\sigma\sqrt{2[\kappa \vee \log\frac{2T}{\delta}]} \cdot \frac{1}{\sqrt{E}} + 32\gamma\sqrt{\kappa \vee \log\frac{2T}{\delta}} \cdot \eta\sqrt{E} + 4\alpha \cdot \eta^2 E \end{aligned}$$

Set $C_1 = 4\Delta_f, C_2 = 2L, C_3 = 0, C_4 = 16\sigma\sqrt{2[\kappa \vee \log\frac{2T}{\delta}]}, C_5 = 32\gamma\sqrt{\kappa \vee \log\frac{2T}{\delta}}, C_6 = 4\alpha$ in Lemma D.3, selecting $\eta = \min\left\{\left(\frac{C_1}{C_2 T}\right)^{\frac{1}{2}}, \left(\frac{C_1^2}{\sqrt{2}C_4C_5T^2}\right)^{\frac{1}{3}}, \left(\frac{C_1^3}{4C_4^2C_6T^3}\right)^{\frac{1}{5}}\right\} = \min\left\{\frac{\sqrt{2}}{L}\sigma_L T^{-\frac{1}{2}}, \frac{1}{4\gamma}\left(\frac{\sigma_\gamma^4}{\sigma}\right)^{\frac{1}{3}}[\kappa \vee \log\frac{2T}{\delta}]^{-\frac{1}{3}}T^{-\frac{2}{3}}, \frac{1}{2}\sqrt{\frac{1}{2\alpha}\left(\frac{2\sigma_\alpha^9}{\sigma^4}\right)^{\frac{1}{5}}[\kappa \vee \log\frac{2T}{\delta}]^{-\frac{1}{5}}T^{-\frac{3}{5}}}\right\}$ and the reset epoch E and batch size B as $E = B = \left\lceil \min\left\{\left(\frac{C_4^4T^2}{2C_1^2C_5^2}\right)^{\frac{1}{3}}, \left(\frac{C_4^6T^4}{16C_1^4C_6^2}\right)^{\frac{1}{5}}\right\}\right\rceil$

$= \left\lceil \min\left\{2\left(\frac{\sigma}{\sigma_\gamma}\right)^{\frac{4}{3}}\left[\kappa \vee \log\frac{2T}{\delta}\right]^{\frac{1}{3}} T^{\frac{2}{3}}, 4\left(\frac{2\sigma^6}{\sigma_\alpha^6}\right)^{\frac{1}{5}}\left[\kappa \vee \log\frac{2T}{\delta}\right]^{\frac{3}{5}} T^{\frac{4}{5}}\right\}\right\rceil$ when $T\left[\kappa \vee \log\frac{2T}{\delta}\right]^{\frac{1}{2}} \geq \frac{\sigma_\gamma^2}{2\sqrt{2}\sigma^2}$ and $T[\kappa \vee \log\frac{2T}{\delta}]^{\frac{3}{4}} \geq 2^{-\frac{11}{4}}\left(\frac{\sigma_\alpha}{\sigma}\right)^{\frac{3}{2}}$, we have:

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(w_t)\|_*\|P_t\|^2 \leq 2\left(\frac{C_1C_2}{T}\right)^{\frac{1}{2}} + 3\left(\frac{\sqrt{2}C_1C_4C_5}{T}\right)^{\frac{1}{3}} + 5\left(\frac{C_1^2C_4^2C_6}{8T^2}\right)^{\frac{1}{5}} \\
= \quad & 4\sqrt{2}\sigma_L T^{-\frac{1}{2}} + 48(\sigma_\gamma^2\sigma)^{\frac{1}{3}}\left(\frac{\kappa \vee \log\frac{2T}{\delta}}{T}\right)^{\frac{1}{3}} + 20(\sigma_\alpha^3\sigma^2)^{\frac{1}{5}}\left(\frac{2[\kappa \vee \log\frac{2T}{\delta}]^{\frac{1}{2}}}{T}\right)^{\frac{2}{5}}
\end{aligned}$$

To analyze the iteration complexity, the following inequalities of T are sufficient to show that $\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(w_t)\|_*\|P_t\|^2 \leq \varepsilon$ for some $\varepsilon > 0$ as 1st, 2nd and 3rd terms are upper bounded by $\varepsilon/2, \varepsilon/4$ and $\varepsilon/4$ respectively, noting that $T \geq \frac{\sigma_\gamma^2}{2\sqrt{2\kappa}\sigma^2} \vee \frac{2^{-\frac{11}{4}}}{\kappa^{\frac{3}{4}}}\left(\frac{\sigma_\alpha}{\sigma}\right)^{\frac{3}{2}}$ implies $T\left[\kappa \vee \log\frac{2T}{\delta}\right]^{\frac{1}{2}} \geq T\kappa^{\frac{1}{2}} \geq \frac{\sigma_\gamma^2}{2\sqrt{2}\sigma^2}$ and $T[\kappa \vee \log\frac{2T}{\delta}]^{\frac{3}{4}} \geq T\kappa^{\frac{3}{4}} \geq 2^{-\frac{11}{4}}\left(\frac{\sigma_\alpha}{\sigma}\right)^{\frac{3}{2}}$:

$$\begin{aligned}
T \geq & \frac{\sigma_\gamma^2}{2\sqrt{2\kappa}\sigma^2} \vee \frac{2^{-\frac{11}{4}}}{\kappa^{\frac{3}{4}}}\left(\frac{\sigma_\alpha}{\sigma}\right)^{\frac{3}{2}} \vee \frac{128\sigma_L^2}{\varepsilon^2} \vee \frac{\kappa\cdot\sigma_\gamma^2\sigma}{(\varepsilon/192)^3} \vee 2\sqrt{\kappa}\sqrt{\frac{\sigma_\alpha^3\sigma^2}{(\varepsilon/80)^5}}, \\
& \frac{\log\frac{2T}{\delta}}{T} \leq \frac{(\varepsilon/192)^3}{\sigma_\gamma^2\sigma}, \ \frac{\log^{\frac{1}{2}}\frac{2T}{\delta}}{T} \leq \frac{1}{2}\sqrt{\frac{(\varepsilon/80)^5}{\sigma_\alpha^3\sigma^2}}
\end{aligned}$$

By substituting $\delta \leftarrow \delta/2$ and $(q,\varepsilon) \leftarrow (1, \frac{(\varepsilon/192)^3}{\sigma_\gamma^2\sigma}), (\frac{1}{2}, \frac{1}{2}\sqrt{\frac{(\varepsilon/80)^5}{\sigma_\alpha^3\sigma^2}})$ in Lemma D.4, the following condition of T is sufficient to guarantee the above inequalities, and therefore $\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(w_t)\|_*\|P_t\|^2 \leq \varepsilon$.

$$\begin{aligned}
T \geq & \frac{\sigma_\gamma^2}{2\sqrt{2\kappa}\sigma^2} \vee \frac{2^{-\frac{11}{4}}}{\kappa^{\frac{3}{4}}}\left(\frac{\sigma_\alpha}{\sigma}\right)^{\frac{3}{2}} \vee \frac{128\sigma_L^2}{\varepsilon^2} \vee \frac{\sigma_\gamma^2\sigma}{(\varepsilon/192)^3}\left[\kappa \vee \mathrm{e}\log\left[\mathrm{e} \vee \frac{2\sigma_\gamma^2\sigma}{(\varepsilon/192)^3\delta}\right]\right] \\
& \vee 2\sqrt{\frac{\sigma_\alpha^3\sigma^2}{(\varepsilon/80)^5}}\left[\sqrt{\kappa} \vee \mathrm{e}^{\frac{1}{4}}\log^{\frac{1}{2}}\left[\mathrm{e} \vee \left(\frac{4}{\delta}\sqrt{\frac{\sigma_\alpha^3\sigma^2}{(\varepsilon/80)^5}}\right)\right]\right]
\end{aligned}$$

In particular, for all $\varepsilon \in \mathbb{R}_+$ is sufficiently small such that $\varepsilon \leq \varepsilon_0(\sigma_\alpha, \sigma_\gamma, \sigma, \kappa, \delta)$, then the lower bound on T simplifies asymptotically to

$$T \geq \max\left\{\frac{\mathrm{e}\cdot\sigma_\gamma^2\sigma}{(\varepsilon/192)^3}\log\frac{2\sigma_\gamma^2\sigma}{(\varepsilon/192)^3\delta}, 2\mathrm{e}^{\frac{1}{4}}\sqrt{\frac{\sigma_\alpha^3\sigma^2}{(\varepsilon/80)^5}}\log^{\frac{1}{2}}\left(\frac{4}{\delta}\sqrt{\frac{\sigma_\alpha^3\sigma^2}{(\varepsilon/80)^5}}\right), \frac{2\sigma_L^2}{(\varepsilon/8)^2}\right\}$$

where the uniform upper bound $\varepsilon_0(\sigma_\alpha, \sigma_\gamma, \sigma, \kappa, \delta)$ of ε is obtained by noting $\kappa \geq 1$ and solving the following inequalities for ε to ensure the dominated terms in the above asymptotic bound are larger than the other terms of the lower bound on T.

$$\begin{cases}
\frac{2\sigma_\gamma^2\sigma}{(\varepsilon/192)^3\delta} \geq \mathrm{e}, \ \frac{4}{\delta}\sqrt{\frac{\sigma_\alpha^3\sigma^2}{(\varepsilon/80)^5}} \geq \mathrm{e}, \\
\mathrm{e}\log\frac{2\sigma_\gamma^2\sigma}{(\varepsilon/192)^3\delta} \geq \kappa, \mathrm{e}^{\frac{1}{4}}\log^{\frac{1}{2}}\left(\frac{4}{\delta}\sqrt{\frac{\sigma_\alpha^3\sigma^2}{(\varepsilon/80)^5}}\right) \geq \sqrt{\kappa}, \\
\frac{\mathrm{e}\cdot\sigma_\gamma^2\sigma}{(\varepsilon/192)^3} \geq \frac{1}{2\sqrt{2\kappa}}\frac{\sigma_\gamma^2}{\sigma^2}, 2\mathrm{e}^{\frac{1}{4}}\sqrt{\frac{\sigma_\alpha^3\sigma^2}{(\varepsilon/80)^5}} \geq \frac{2^{-\frac{11}{4}}}{\kappa^{\frac{3}{4}}}\left(\frac{\sigma_\alpha}{\sigma}\right)^{\frac{3}{2}}
\end{cases}$$

The explicit expression of

$$\varepsilon_0(\sigma_\alpha, \sigma_\gamma, \sigma, \kappa, \delta) \equiv (\sigma_\alpha^3\sigma^2)^{\frac{1}{5}}\cdot\varepsilon_1(\kappa,\delta) \vee (\sigma_\gamma^2\sigma)^{\frac{1}{3}}\cdot\varepsilon_2(\kappa,\delta) \vee \sigma\cdot\varepsilon_3(\kappa,\delta)$$

is given by:

$$\varepsilon_1(\kappa,\delta) = 80\left(\frac{16}{\delta^2 \mathrm{e}^{2\vee(2\kappa/\sqrt{\mathrm{e}})}}\right)^{1/5}$$
$$\varepsilon_2(\kappa,\delta) = 192\left(\frac{2}{\delta \mathrm{e}^{1\vee(\kappa/\mathrm{e})}}\right)^{1/3}$$
$$\varepsilon_3(\kappa,\delta) = 192\sqrt{2}(\mathrm{e}\kappa^{1/2})^{1/3} \ \wedge \ 160\sqrt{2}(\mathrm{e}^{1/2}\kappa^{3/2})^{1/5}$$

Therefore, the iteration complexity for T and the oracle complexity for $N = B{\cdot}T/E{+}1{\cdot}(T{-}T/E) = T(2-1/B)$ are given by:

$$N,\, T = \mathcal{O}\left(\frac{\sigma_\gamma^2\sigma}{\varepsilon^3}\log\frac{\sigma_\gamma^2\sigma}{\varepsilon^3\delta} \vee \sqrt{\frac{\sigma_\alpha^3\sigma^2}{\varepsilon^5}}\log^{\frac{1}{2}}\left(\frac{1}{\delta}\sqrt{\frac{\sigma_\alpha^3\sigma^2}{\varepsilon^5}}\right) \vee \frac{\sigma_L^2}{\varepsilon^2}\right)$$

with the selection of batch size $B = E$ to guarantee $\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(w_t)\|_*\|P_t\|^2 \leq \varepsilon$ with probability at least $1-\delta$ for sufficiently small ε. □

F Application 2: Expectation-Constrained Stochastic Optimization

Optimization Problem We aim to solve the stochastic constrained optimization problem:

$$\min_{w\in\mathcal{W}} f(w) \coloneqq \mathbb{E}[F(w,\xi)] \quad \text{s.t.} \quad h(w) \coloneqq \mathbb{E}[H(w,\xi)] \le 0,$$

where $F(w,\xi) : \mathcal{W} \to \mathbb{R}$ and $H(w,\xi) : \mathcal{W} \to \mathbb{R}$ are convex, $\mathcal{W}$ is a closed, convex subset of $\mathcal{X}$.

The Unified Estimator and SGM Update Rule

- **Update Direction:** $U_t \coloneqq \mathbb{I}_{\{v_t\le\varepsilon\}} F'(w_t,\zeta_t) + (1-\mathbb{I}_{\{v_t\le\varepsilon\}}) H'(w_t,\zeta_t)$.
- **Mirror Descent Step:**

$$w_{t+1} = \arg\min_{w\in\mathcal{W}} \{\langle \eta_t U_t, w\rangle + D_\Phi(w,w_t)\} = w_t - \eta_t P(w_t,U_t,\eta_t)$$

  where $\eta_t = \eta$ is the constant step size.

The active-phase feasible and non-feasible index sets

$$\mathcal{A} \coloneqq \{t\in[T] : v_t \le \varepsilon\}, \qquad \mathcal{B} \coloneqq [T]\setminus\mathcal{A}.$$

Theorem F.1 (Constrained SGM via Unified Estimation, Theorem 6.3)**.** Suppose Assumption 6.2 hold. Suppose there exists a deterministic quantity $\mathcal{E} \ge 0$ such that

$$\mathbb{P}(|e_t| \le \mathcal{E} \quad \text{for all } t\in[T]) \ge 1-\frac{\delta}{2}. \tag{76}$$

Define $\mathsf{OptError}(\eta,T,\delta) \coloneqq \frac{R^2}{2\eta T} + \frac{\eta G^2}{2} + \frac{2DG}{\sqrt{T}}\sqrt{2\log\frac{4}{\delta}}$. Set the switching threshold to $\varepsilon \coloneqq \mathsf{OptError}(\eta,T,\delta) + \mathcal{E}$. Then, with probability at least $1-\delta$, the selected set $\mathcal{A}$ is non-empty, and the active-phase selected average $\bar{w} \coloneqq \frac{1}{|\mathcal{A}|}\sum_{t\in\mathcal{A}} w_t$ satisfies

$$f(\bar{w}) - f(w^\star) \le \varepsilon, \qquad h(\bar{w}) \le \varepsilon + \mathcal{E},$$

where $w^\star$ is any optimal feasible solution.

In all family and parameter regimes considered below, the envelope $\mathcal{E}$ in (76) is verified by specializing Theorem 4.2 and bounding the resulting estimator error uniformly over $t\in[T]$.

Proof. We work on the intersection of the following two high-probability events:

- the tail-envelope event from (76),
- the martingale concentration event established below.

By a union bound, this intersection of the two high-probability events has probability at least $1-\delta$.

Step 1: Mirror-descent inequality on the active phase. By the first-order optimality condition (see Remark A.6 of First-Order Optimality Condition) for the mirror-descent update such that $\langle \eta_t U_t + \nabla\Phi(w_{t+1}) - \nabla\Phi(w_t),\, w^\star - w_{t+1}\rangle \ge 0$, and using the three-point identity for Bregman divergences, we obtain

$$\eta_t\langle U_t, w_{t+1} - w^\star\rangle \le [D_\Phi(w^\star,w_t) - D_\Phi(w^\star,w_{t+1})] - D_\Phi(w_{t+1},w_t).$$

By the 1-strong convexity of Φ, we obtain $D_\Phi(w_{t+1},w_t) \ge \frac{1}{2}\|w_t - w_{t+1}\|^2 = \frac{\eta_t^2}{2}\|P(w_t,U_t,\eta_t)\|^2$, also noting that $\|U_t\|_*^2 + \|P(w_t,U_t,\eta_t)\|^2 \ge 2\|U_t\|_*\|P(w_t,U_t,\eta_t)\| \ge 2\langle U_t, P(w_t,U_t,\eta_t)\rangle$, hence,

$$\begin{aligned}
\langle U_t, w_t - w^\star\rangle &= \langle U_t, w_{t+1} - w^\star\rangle + \langle U_t, w_t - w_{t+1}\rangle \\
&\le \frac{1}{\eta_t}[D_\Phi(w^\star,w_t) - D_\Phi(w^\star,w_{t+1})] - \frac{\eta_t}{2}\|P(w_t,U_t,\eta_t)\|^2 + \eta_t\langle U_t, P(w_t,U_t,\eta_t)\rangle \\
&\le \frac{1}{\eta_t}[D_\Phi(w^\star,w_t) - D_\Phi(w^\star,w_{t+1})] + \frac{\eta_t}{2}\|U_t\|_*^2
\end{aligned}$$

Summing over $t \in [T]$ and using $\eta_t = \eta$ yields

$$\sum_{t\in[T]} \langle U_t, w_t - w^\star \rangle \le \frac{D_\Phi(w^\star, w_1) - D_\Phi(w^\star, w_{T+1})}{\eta} + \frac{\eta}{2} \sum_{t\in[T]} \|U_t\|_*^2.$$

Since $\sup_{u,v\in\mathcal{W}} D_\Phi(u,v) \le R^2/2$ and $\|U_t\|_* \le G$ almost surely, and $D_\Phi(w^\star, w_{T+1}) \ge 0$,

$$\sum_{t\in[T]} \langle U_t, w_t - w^\star \rangle \le \frac{R^2}{2\eta} + \frac{\eta T G^2}{2}. \tag{77}$$

Step 2: Convexity decomposition and martingale concentration.

Let $\mathcal{G}_t := \sigma(\mathcal{F}_{t-1}, \xi_t^B \mid_{b_t=1}, \xi_t \mid_{b_t=0}, b_t)$ denote the sigma-field generated by the history $\mathcal{F}_{t-1}$ and the estimator randomness up to the switching decision (which also depends on $\xi_t^B \mid_{b_t=1}, \xi_t \mid_{b_t=0}, b_t$) at iteration t. By construction, the indicator $I_t = \mathbb{I}_{\{v_t \le \varepsilon\}}$ is $\mathcal{G}_t$-measurable, $w_t - w^\star$ is $\mathcal{G}_t$-measurable, while ζ_t is independent of $\mathcal{G}_t$. Define

$$g_t^f := \mathbb{E}[F'(w_t, \zeta_t) \mid \mathcal{G}_t] \in \partial f(w_t), \qquad g_t^h := \mathbb{E}[H'(w_t, \zeta_t) \mid \mathcal{G}_t] \in \partial h(w_t).$$

Also define

$$M_t := I_t \langle F'(w_t, \zeta_t) - g_t^f,\ w_t - w^\star \rangle + (1 - I_t) \langle H'(w_t, \zeta_t) - g_t^h,\ w_t - w^\star \rangle.$$

By the conditional unbiasedness assumption, $\mathbb{E}[M_t \mid \mathcal{G}_t] = 0$, so $\{M_t\}_{t\in[T]}$ is a martingale difference sequence. Moreover,

$$|M_t| \le 2G\|w_t - w^\star\| \le 2DG.$$

Hence, Azuma–Hoeffding inequality implies that with probability at least $1 - \delta/2$,

$$-\sum_{t\in[T]} M_t \le 2DG\sqrt{2T \log \frac{4}{\delta}}.$$

By convexity of f and h, and $w^\star$ is feasible, $h(w^\star) \le 0$, hence

$$\langle g_t^f, w_t - w^\star \rangle \ge f(w_t) - f(w^\star), \qquad \langle g_t^h, w_t - w^\star \rangle \ge h(w_t) - h(w^\star) \ge h(w_t).$$

Therefore,

$$\sum_{t\in[T]} \langle U_t, w_t - w^\star \rangle \ge \sum_{t\in\mathcal{A}} \big(f(w_t) - f(w^\star)\big) + \sum_{t\in\mathcal{B}} h(w_t) + \sum_{t\in[T]} M_t.$$

Combining this with (77), dividing by T, and using the Azuma bound gives

$$\frac{1}{T} \sum_{t\in\mathcal{A}} \big(f(w_t) - f(w^\star)\big) + \frac{1}{T} \sum_{t\in\mathcal{B}} h(w_t) \le \frac{R^2}{2\eta T} + \frac{\eta G^2}{2} + \frac{2DG}{\sqrt{T}} \sqrt{2 \log \frac{4}{\delta}} =: \mathsf{OptError}(\eta, T, \delta). \tag{78}$$

Now use $h(w_t) = v_t - e_t \ge v_t - |e_t|$ on $\mathcal{B}$:

$$\frac{1}{T} \sum_{t\in\mathcal{A}} \big(f(w_t) - f(w^\star)\big) + \frac{1}{T} \sum_{t\in\mathcal{B}} v_t \le \mathsf{OptError}(\eta, T, \delta) + \frac{1}{T} \sum_{t\in\mathcal{B}} |e_t|.$$

On the tail-envelope event, $|e_t| \le \mathcal{E}$ for all $t \in [T]$, hence by letting $\varepsilon = \mathsf{OptError}(\eta, T, \delta) + \mathcal{E}$,

$$\frac{1}{T} \sum_{t\in\mathcal{A}} \big(f(w_t) - f(w^\star)\big) + \frac{1}{T} \sum_{t\in\mathcal{B}} v_t \le \mathsf{OptError}(\eta, T, \delta) + \mathcal{E} = \varepsilon.$$

Thus

$$\frac{1}{T} \sum_{t\in\mathcal{A}} \big(f(w_t) - f(w^\star)\big) + \frac{1}{T} \sum_{t\in\mathcal{B}} v_t \le \varepsilon. \tag{79}$$

Step 3: Non-emptiness of the selected set. Assume for contradiction that $\mathcal{A} = \varnothing$. Then $\mathcal{B} = [T]$, and by definition of $\mathcal{B}$ we have $v_t > \varepsilon$ for every $t \in \mathcal{B}$. Therefore $\frac{1}{T} \sum_{t\in\mathcal{B}} v_t > \varepsilon$, which contradicts (79). Hence $\mathcal{A} \neq \varnothing$.

Table 7: Cross-reference of results for Application 2 (Section 6.2 and Appendix F). For each parameter regime (case), the table lists the tuned-bound theorem and complexity corollary, and the dominant ε-dependence of the oracle complexity for each family. All entries use $B{=}1$ for Case 1, $B{=}1/p$ for Case 2, and $B{=}E$ for Case 3, yielding $N{<}2T$.

Case	Reference		Dominant ε-rate		
	Bound	Complexity	Fam. 1	Fam. 2	Fam. 3
Case 1: $\beta_t{=}\beta{<}1,\ p_t{=}0,\ B{=}1$					
Mom. / STORM / SO Mom.	Thm. F.3	Cor. F.4	ε^{-4}	ε^{-3}	ε^{-3} $(\varepsilon^{-\frac{5}{2}})$
Case 2: $\beta_t{=}1,\ p_t{=}p{>}0,\ B{=}1/p$					
Prob. Mom. / PAGE / SO PAGE	Thm. F.5	Cor. F.6	ε^{-4}	ε^{-3}	ε^{-3} $(\varepsilon^{-\frac{5}{2}})$
Case 3: $\beta_t{=}1,\ p_t{=}\mathbb{I}_{\{t \bmod E=0\}},\ B{=}E$					
Per. Mom. / SPIDER / SO SPIDER	Thm. F.7	Cor. F.8	ε^{-4}	ε^{-3}	ε^{-3} $(\varepsilon^{-\frac{5}{2}})$

The notation ε^{-3} $(\varepsilon^{-5/2})$ for Family 3 indicates a dominant ε^{-3} term (from the γ-dependent variance) with a subdominant $\varepsilon^{-5/2}$ contribution (from the α-dependent second-order bias). The base SGM reduction used by all entries is Theorem 6.3 / Theorem F.1.

Step 4: Optimality of the selected average. If $\sum_{t\in\mathcal{A}}\big(f(w_t)-f(w^\star)\big)\le 0$, then convexity of f implies $f(\bar{w})-f(w^\star)\le 0\le\varepsilon$. Otherwise, the sum over $\mathcal{A}$ is positive. Since $v_t>\varepsilon$ for all $t\in\mathcal{B}$, (79) implies

$$\frac{1}{T}\sum_{t\in\mathcal{A}}\big(f(w_t)-f(w^\star)\big)<\varepsilon-\frac{|\mathcal{B}|}{T}\varepsilon=\frac{|\mathcal{A}|}{T}\varepsilon.$$

Dividing by $|\mathcal{A}|/T$ and using Jensen's inequality yields $f(\bar{w})-f(w^\star)\le\frac{1}{|\mathcal{A}|}\sum_{t\in\mathcal{A}}\big(f(w_t)-f(w^\star)\big)<\varepsilon$. Hence $f(\bar{w})-f(w^\star)\le\varepsilon$.

Step 5: Feasibility of the selected average. By the convexity of h,

$$h(\bar{w})\le\frac{1}{|\mathcal{A}|}\sum_{t\in\mathcal{A}}h(w_t)=\frac{1}{|\mathcal{A}|}\sum_{t\in\mathcal{A}}\big(v_t-e_t\big)\le\frac{1}{|\mathcal{A}|}\sum_{t\in\mathcal{A}}\big(v_t+|e_t|\big).$$

Since $v_t\le\varepsilon$ on $\mathcal{A}$ and $|e_t|\le\mathcal{E}$ on the entire active phase,

$$h(\bar{w})\le\frac{1}{|\mathcal{A}|}\sum_{t\in\mathcal{A}}\big(\varepsilon+\mathcal{E}\big)=\varepsilon+\mathcal{E}.$$

This completes the proof. □

Remark F.2 (Constrained SGM on a Reusable Analysis Window)**.** We can extend the above analysis to a selected set such that for $t_0<T$ and $T_a:=T-t_0$

$$\mathcal{A}:=\{t\in[T]:t>t_0,v_t\le\varepsilon\}$$

By letting the threshold be

$$\varepsilon:=\frac{R^2}{2\eta T_a}+\frac{\eta G^2}{2}+\frac{2DG}{\sqrt{T_a}}\sqrt{2\log\frac{4}{\delta}}+\mathcal{E},$$

and $\mathbb{P}(\sup_{t\in[t_0+1:T]}|e_t|>\mathcal{E})\le\delta/2$, then, with probability at least $1-\delta$, the selected set $\mathcal{A}$ is non-empty, and the uniform average $\bar{w}=\frac{1}{|\mathcal{A}|}\sum_{t\in\mathcal{A}}w_t$ satisfies both $f(\bar{w})-f\left(w^*\right)\le\varepsilon$ and $h(\bar{w})\le\varepsilon+\mathcal{E}$.

F.1 Case 1 for Families 1–3 in the constrained setting

We instantiate Theorem F.1 for the three Case 1 estimators, and note that $\kappa=1$ for $(\mathbb{R},|\cdot|)$. Throughout this subsection, $1+\sqrt{3}<2\sqrt{2}$, $\log\frac{4T}{\delta}\ge\log 4>1=\kappa$, then

$$C\left(\frac{\delta}{4T},\kappa\right)=\sqrt{\kappa}+\sqrt{3\log\frac{4T}{\delta}}\le(1+\sqrt{3})\sqrt{\kappa\vee\log\frac{4T}{\delta}}\le 2\sqrt{2[\kappa\vee\log\frac{4T}{\delta}]}=2\sqrt{2\log\frac{4T}{\delta}}.$$

Theorem F.3 (Constrained Case 1 Bounds for Families 1–3)**.** For each family below, set $\varepsilon := \mathsf{OptError}(\eta, T_a, \delta) + \mathcal{E}$ with $T_a = T - t_0, t_0 = \lfloor T/2 \rfloor$, select the batch size $B = 1$, and let the selected set be $\mathcal{A} = \{t \in [T] : t > t_0, v_t \le \varepsilon\}$, $\Gamma_T := 2 \log \frac{4T}{\delta}$.

1. **Family 1.** Choose step size $\eta = \min\left\{\sqrt{2}\frac{R}{G}T^{-\frac{1}{2}}, \left(\frac{R^6}{32\sigma^2 LG\cdot\Gamma_T}\right)^{\frac{1}{4}} T^{-\frac{3}{4}}\right\}$, and momentum parameter $1-\beta = \max\left\{T^{-\frac{2}{3}}, \left(\frac{R^2 LG}{8\sigma^2\cdot\Gamma_T}\right)^{\frac{1}{2}} T^{-\frac{1}{2}}\right\}$. Assume moreover that $T\Gamma_T \ge \frac{R^2 LG}{8\sigma^2}$. The selected set $\mathcal{A}$ is nonempty with probability at least $1-\delta$, and the active-phase selected average $\bar{w} = \frac{1}{|\mathcal{A}|}\sum_{t\in\mathcal{A}} w_t$ satisfies $f(\bar{w}) - f(w^\star) \le Q_T^{(1,1)}$, $h(\bar{w}) \le Q_T^{(1,1)}$, where

$$Q_T^{(1,1)} = \sqrt{2}RGT^{-\frac{1}{2}} + 8\sigma\left(\frac{\Gamma_T^{\frac{3}{2}}}{T}\right)^{\frac{1}{3}} + 3\left(32R^2LG\sigma^2\right)^{\frac{1}{4}}\left(\frac{\Gamma_T}{T}\right)^{\frac{1}{4}} + 4DG\left(\frac{\log\frac{4}{\delta}}{T}\right)^{\frac{1}{2}}.$$

2. **Family 2.** Choose $\eta = \min\left\{\frac{\sqrt{2}R}{G}T^{-\frac{1}{2}}, \left(\frac{R^4}{32\ell G\sigma\Gamma_T}\right)^{\frac{1}{3}} T^{-\frac{2}{3}}\right\}$, $1-\beta = \max\left\{2^{-\frac{1}{3}}T^{-\frac{2}{3}}, \left(\frac{R^4\ell^2G^2/\sigma^4}{32\Gamma_T}\right)^{\frac{1}{3}} T^{-\frac{2}{3}}\right\}$. Assume moreover that $T\Gamma_T^{1/2} > \frac{R^2\ell G}{4\sqrt{2}\,\sigma^2}$. The selected set $\mathcal{A}$ is nonempty with probability at least $1-\delta$, and $f(\bar{w}) - f(w^\star) \le Q_T^{(2,1)}$, $h(\bar{w}) \le Q_T^{(2,1)}$, where

$$Q_T^{(2,1)} = \sqrt{2}RGT^{-\frac{1}{2}} + 8\sigma\left(\frac{2\Gamma_T^{\frac{3}{2}}}{T}\right)^{\frac{1}{3}} + 3(32R^2\ell G\sigma)^{\frac{1}{3}}\left(\frac{\Gamma_T}{T}\right)^{\frac{1}{3}} + 4DG\left(\frac{\log\frac{4}{\delta}}{T}\right)^{\frac{1}{2}}.$$

3. **Family 3.** Choose $\eta = \min\left\{\frac{\sqrt{2}R}{G}T^{-\frac{1}{2}}, \left(\frac{R^4}{32\gamma G\sigma\Gamma_T}\right)^{\frac{1}{3}} T^{-\frac{2}{3}}, \left(\frac{R^6}{32\alpha G^2\sigma^2\Gamma_T}\right)^{\frac{1}{5}} T^{-\frac{3}{5}}\right\}$, $1-\beta = \max\left\{2^{-\frac{1}{3}}T^{-\frac{2}{3}}, \left(\frac{R^4\gamma^2G^2/\sigma^4}{32\Gamma_T}\right)^{\frac{1}{3}} T^{-\frac{2}{3}}, \frac{1}{8}\left(\frac{R^8\alpha^2G^4/\sigma^6}{\Gamma_T^3}\right)^{\frac{1}{5}} T^{-\frac{4}{5}}\right\}$. Assume moreover that $T\Gamma_T^{\frac{1}{2}} \ge \frac{R^2\gamma G}{4\sqrt{2}\sigma^2}$ and $T\Gamma_T^{\frac{3}{4}} \ge 2^{-\frac{15}{4}}\left(\frac{R^2\sqrt{\alpha}G}{\sigma}\right)^{\frac{3}{2}}$. The selected set $\mathcal{A}$ is nonempty with probability at least $1-\delta$, and $f(\bar{w}) - f(w^\star) \le Q_T^{(3,1)}$, $h(\bar{w}) \le Q_T^{(3,1)}$, where

$$Q_T^{(3,1)} = \sqrt{2}RGT^{-\frac{1}{2}} + 8\sigma\left(\frac{2\Gamma_T^{\frac{3}{2}}}{T}\right)^{\frac{1}{3}} + 3(32R^2\gamma G\sigma)^{\frac{1}{3}}\left(\frac{\Gamma_T}{T}\right)^{\frac{1}{3}} + 6(R^4\alpha G^2\sigma^2)^{\frac{1}{5}}\left(\frac{\Gamma_T}{T^2}\right)^{1/5} + 4DG\left(\frac{\log\frac{4}{\delta}}{T}\right)^{\frac{1}{2}}.$$

Proof. For the analysis of case 1 in all families, we let $t_0 = \lfloor \frac{T}{2} \rfloor$, then $T_a = T - t_0 \ge T/2$, by noting that $x\exp(-x) \le 1/\mathrm{e} < 1/2, \forall x > 0$, then we have the following bound for $t \in [t_0 + 1 : T]$

$$\beta^t = (1-[1-\beta])^t \le (1-[1-\beta])^{T/2} \le \exp(-\frac{1-\beta}{2}T) \le \frac{1/2}{\frac{1-\beta}{2}T} = \frac{1}{(1-\beta)T}$$

The selected set is defined by $\mathcal{A} := \{t \in [T] : t > t_0, v_t \le \varepsilon\}$. Applying the established Theorem 6.3 and using its remark.

$$\max\{f(\bar{w}) - f(w^*), h(\bar{w})\} \le \varepsilon + \mathcal{E} = (\mathsf{OptError}(\eta, T_a, \delta) + \mathcal{E}) + \mathcal{E} \le \mathsf{OptError}(\eta, T/2, \delta) + 2\mathcal{E}$$

Family 1 By substituting the high-probability upperbound $\mathcal{E}$ of $|e_t|$ from the instantiations of Theorem 4.2 (replacing δ with $\delta/2$ and selecting $B = 1$ in the bound of Table 2) into the above inequality, using $\beta^t \le \frac{1}{(1-\beta)T}$ for $t \in [t_0 + 1 : T]$ and $C(\delta/(4T), \kappa) \le 2\sqrt{2\log(4T/\delta)} =: 2\sqrt{\Gamma_T}$,

we have

$$\begin{aligned}
&\mathsf{OptError}(\eta, T/2, \delta) + 2\mathcal{E} \leq \left(\frac{R^2}{\eta T} + \frac{\eta G^2}{2} + 4DG\sqrt{\frac{\log\frac{4}{\delta}}{T}}\right) \\
&+ 2\left(\frac{C\left(\frac{\delta}{4T}, \kappa\right)\sigma}{(1-\beta)T} + \frac{LG\eta}{1-\beta} + C\left(\frac{\delta}{4T}, \kappa\right)\sigma\sqrt{1-\beta}\right) \\
&\leq 4DG\sqrt{\frac{\log\frac{4}{\delta}}{T}} + \left[\frac{R^2}{\eta T} + \frac{G^2}{2}\cdot\eta + 4\sqrt{\Gamma_T}\sigma\cdot\frac{1}{(1-\beta)T} + 4\sqrt{\Gamma_T}\sigma\cdot\sqrt{1-\beta} + 2LG\cdot\frac{\eta}{1-\beta}\right]
\end{aligned}$$

By letting $C_1 = R^2, C_2 = \frac{G^2}{2}, C_3 = 4\sqrt{\Gamma_T}\sigma, C_4 = 4\sqrt{\Gamma_T}\sigma, C_5 = 2LG$ in Lemma D.2, selecting the step size and momentum parameter as $\eta = \min\left\{\left(\frac{C_1}{C_2 T}\right)^{\frac{1}{2}}, \left(\frac{C_1^3}{C_4^2 C_5 T^3}\right)^{\frac{1}{4}}\right\} = \min\left\{\sqrt{2}\frac{R}{G}T^{-\frac{1}{2}}, \left(\frac{R^6}{32\sigma^2 LG\cdot\Gamma_T}\right)^{\frac{1}{4}}T^{-\frac{3}{4}}\right\}$, $1-\beta = \max\left\{\left(\frac{C_3}{C_4 T}\right)^{\frac{2}{3}}, \left(\frac{C_1 C_5}{C_4^2 T}\right)^{\frac{1}{2}}\right\} = \max\left\{T^{-\frac{2}{3}}, \left(\frac{R^2 LG}{8\sigma^2\cdot\Gamma_T}\right)^{\frac{1}{2}}T^{-\frac{1}{2}}\right\}$ when $T\Gamma_T \geq \frac{R^2 LG}{8\sigma^2}$, we have:

$$\begin{aligned}
&\max\{f(\bar{w}) - f(w^*), h(\bar{w})\} \\
\leq\ & 2\left(\frac{C_1 C_2}{T}\right)^{\frac{1}{2}} + 2\left(\frac{C_3 C_4^2}{T}\right)^{\frac{1}{3}} + 3\left(\frac{C_1 C_4^2 C_5}{T}\right)^{\frac{1}{4}} + 4DG\left(\frac{\log\frac{4}{\delta}}{T}\right)^{\frac{1}{2}} \\
\leq\ & \sqrt{2}RGT^{-\frac{1}{2}} + 8\sigma\left(\frac{\Gamma_T^{\frac{3}{2}}}{T}\right)^{\frac{1}{3}} + 3\left(32R^2LG\sigma^2\right)^{\frac{1}{4}}\left(\frac{\Gamma_T}{T}\right)^{\frac{1}{4}} + 4DG\left(\frac{\log\frac{4}{\delta}}{T}\right)^{\frac{1}{2}}
\end{aligned}$$

Family 2 By substituting the high probability upperbound $\mathcal{E}$ of $|e_t|$ from the instantiations of Theorem 4.2 (replacing δ with $\delta/2$ and selecting $B = 1$ in the bound of Table 2) into the above inequality, and using $\beta^t \leq \frac{1}{(1-\beta)T}$ for $t \in [t_0 + 1 : T]$ and $C(\delta/(4T), \kappa) \leq 2\sqrt{2\log(4T/\delta)} =: 2\sqrt{\Gamma_T}$, noting that $\frac{(1-\beta)^2}{1-\beta^2} \leq 1-\beta, \frac{\beta^2}{1-\beta^2} \leq \frac{1}{1-\beta}$ for $\beta \in (0,1)$ and the elementary inequality $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}, \forall x, y \geq 0$, we show the following upper bound for $\mathcal{E}$:

$$\mathcal{E} \leq \frac{C\left(\frac{\delta}{4T}, \kappa\right)\sigma}{(1-\beta)T} + C\left(\frac{\delta}{4T}, \kappa\right)\left[\sigma\sqrt{2(1-\beta)} + \ell G\eta\sqrt{\frac{2}{1-\beta}}\right],$$

then we have

$$\begin{aligned}
&\mathsf{OptError}(\eta, T/2, \delta) + 2\mathcal{E} \leq \left(\frac{R^2}{\eta T} + \frac{\eta G^2}{2} + 4DG\sqrt{\frac{\log\frac{4}{\delta}}{T}}\right) \\
&+ 2\left(\frac{C\left(\frac{\delta}{4T}, \kappa\right)\sigma}{(1-\beta)T} + C\left(\frac{\delta}{4T}, \kappa\right)\left[\sigma\sqrt{2(1-\beta)} + \ell G\eta\sqrt{\frac{2}{1-\beta}}\right]\right) \\
&\leq 4DG\sqrt{\frac{\log\frac{4}{\delta}}{T}} + \left[\frac{R^2}{\eta T} + \frac{G^2}{2}\cdot\eta + 4\sqrt{\Gamma_T}\sigma\cdot\frac{1}{(1-\beta)T} + 4\sqrt{2\Gamma_T}\sigma\cdot\sqrt{1-\beta} + 4\ell G\sqrt{2\Gamma_T}\cdot\frac{\eta}{\sqrt{1-\beta}}\right]
\end{aligned}$$

By letting $C_1 = R^2, C_2 = \frac{G^2}{2}, C_3 = 4\sqrt{\Gamma_T}\sigma, C_4 = 4\sqrt{2\Gamma_T}\sigma, C_5 = 4\ell G\sqrt{2\Gamma_T}$ in Lemma D.2, selecting the step size and momentum parameter as $\eta = \min\left\{\left(\frac{C_1}{C_2}\right)^{\frac{1}{2}}T^{-\frac{1}{2}}, \left(\frac{C_1^2}{C_4 C_5}\right)^{\frac{1}{3}}T^{-\frac{2}{3}}\right\} = \min\left\{\frac{\sqrt{2}R}{G}T^{-\frac{1}{2}}, \left(\frac{R^4}{32\ell G\sigma\Gamma_T}\right)^{\frac{1}{3}}T^{-\frac{2}{3}}\right\}$, $1-\beta = \max\left\{\left(\frac{C_3}{C_4 T}\right)^{\frac{2}{3}}, \left(\frac{C_1 C_5}{C_4^2 T}\right)^{\frac{2}{3}}\right\} = \max\left\{2^{-\frac{1}{3}}T^{-\frac{2}{3}},\right.$

$\left(\frac{R^4\ell^2G^2/\sigma^4}{32\Gamma_T}\right)^{\frac{1}{3}}T^{-\frac{2}{3}}\Big\}$ then when $T\Gamma_T^{\frac{1}{2}} \geq \frac{R^2\ell G}{4\sqrt{2}\sigma^2}$, we have:

$$\begin{aligned}
& \max\{f(\bar{w}) - f(w^*), h(\bar{w})\} \\
\leq\ & 2\left(\frac{C_1C_2}{T}\right)^{\frac{1}{2}} + 2\left(\frac{C_3C_4^2}{T}\right)^{\frac{1}{3}} + 3\left(\frac{C_1C_4C_5}{T}\right)^{\frac{1}{3}} + 4DG\left(\frac{\log\frac{4}{\delta}}{T}\right)^{\frac{1}{2}} \\
\leq\ & \sqrt{2}RGT^{-\frac{1}{2}} + 8\sigma\left(\frac{2\Gamma_T^{\frac{3}{2}}}{T}\right)^{\frac{1}{3}} + 3(32R^2\ell G\sigma)^{\frac{1}{3}}\left(\frac{\Gamma_T}{T}\right)^{\frac{1}{3}} + 4DG\left(\frac{\log\frac{4}{\delta}}{T}\right)^{\frac{1}{2}}
\end{aligned}$$

Family 3 By substituting the high probability upperbound $\mathcal{E}$ of $|e_t|$ from the instantiations of Theorem 4.2 (replacing δ with $\delta/2$ and selecting $B = 1$ in the bound of Table 2) into the above inequality, and using $\beta^t \leq \frac{1}{(1-\beta)T}$ for $t \in [t_0 + 1 : T]$ and $C(\delta/(4T), \kappa) \leq 2\sqrt{2\log(4T/\delta)} =: 2\sqrt{\Gamma_T}$, noting that $\frac{(1-\beta)^2}{1-\beta^2} \leq 1-\beta, \frac{\beta^2}{1-\beta^2} \leq \frac{1}{1-\beta}$ for $\beta \in (0,1)$ and the elementary inequality $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}, \forall x, y \geq 0$, we show the following upper bound for $\mathcal{E}$:

$$\mathcal{E} \leq \frac{C\left(\frac{\delta}{4T}, \kappa\right)\sigma}{(1-\beta)T} + \frac{\alpha\beta\eta^2G^2}{2(1-\beta)} + C\left(\frac{\delta}{4T}, \kappa\right)\left[\sigma\sqrt{2(1-\beta)} + \gamma G\eta\sqrt{\frac{2}{1-\beta}}\right],$$

then we have

$$\begin{aligned}
&\mathsf{OptError}(\eta, T/2, \delta) + 2\mathcal{E} \leq \left(\frac{R^2}{\eta T} + \frac{\eta G^2}{2} + 4DG\sqrt{\frac{\log\frac{4}{\delta}}{T}}\right) \\
&+ 2\left(\frac{C\left(\frac{\delta}{4T}, \kappa\right)\sigma}{(1-\beta)T} + \frac{\alpha\beta\eta^2G^2}{2(1-\beta)} + C\left(\frac{\delta}{4T}, \kappa\right)\left[\sigma\sqrt{2(1-\beta)} + \gamma G\eta\sqrt{\frac{2}{1-\beta}}\right]\right) \\
&\leq \left[\frac{R^2}{\eta T} + \frac{G^2}{2}\cdot\eta + 4\sqrt{\Gamma_T}\sigma\cdot\frac{1}{(1-\beta)T} + 4\sqrt{2\Gamma_T}\sigma\cdot\sqrt{1-\beta} + 4\gamma G\sqrt{2\Gamma_T}\cdot\frac{\eta}{\sqrt{1-\beta}} + \alpha G^2\cdot\frac{\eta^2}{1-\beta}\right] \\
&+ 4DG\sqrt{\frac{\log\frac{4}{\delta}}{T}}
\end{aligned}$$

By letting $C_1 = R^2, C_2 = \frac{G^2}{2}, C_3 = 4\sqrt{\Gamma_T}\sigma, C_4 = 4\sqrt{2\Gamma_T}\sigma, C_5 = 4\gamma G\sqrt{2\Gamma_T}, C_6 = \alpha G^2$ in Lemma D.2, selecting the step size and momentum parameter such that $\eta = \min\Big\{\left(\frac{C_1}{C_2}\right)^{\frac{1}{2}}T^{-\frac{1}{2}}, \left(\frac{C_1^2}{C_4C_5}\right)^{\frac{1}{3}}T^{-\frac{2}{3}}, \left(\frac{C_1^3}{C_4^2C_6}\right)^{\frac{1}{5}}T^{-\frac{3}{5}}\Big\} = \min\Big\{\frac{\sqrt{2}R}{G}T^{-\frac{1}{2}}, \left(\frac{R^4}{32\gamma G\sigma\Gamma_T}\right)^{\frac{1}{3}}T^{-\frac{2}{3}}, \left(\frac{R^6}{32\alpha G^2\sigma^2\Gamma_T}\right)^{\frac{1}{5}}T^{-\frac{3}{5}}\Big\}$, the momentum parameter β as $1-\beta = \max\Big\{\left(\frac{C_3}{C_4}\right)^{\frac{2}{3}}T^{-\frac{2}{3}}, \left(\frac{C_1C_5}{C_4^2}\right)^{\frac{2}{3}}T^{-\frac{2}{3}}, \left(\frac{C_1^2C_6}{C_4^3}\right)^{\frac{2}{5}}T^{-\frac{4}{5}}\Big\} = \max\Big\{2^{-\frac{1}{3}}T^{-\frac{2}{3}}, \left(\frac{R^4\gamma^2G^2/\sigma^4}{32\Gamma_T}\right)^{\frac{1}{3}}T^{-\frac{2}{3}}, \frac{1}{8}\left(\frac{R^8\alpha^2G^4/\sigma^6}{\Gamma_T^3}\right)^{\frac{1}{5}}T^{-\frac{4}{5}}\Big\}$, then when $T\Gamma_T^{\frac{1}{2}} \geq \frac{R^2\gamma G}{4\sqrt{2}\sigma^2}$ and $T\Gamma_T^{\frac{3}{4}} \geq 2^{-\frac{15}{4}}\left(\frac{R^2\sqrt{\alpha}G}{\sigma}\right)^{\frac{3}{2}}$, we have:

$$\begin{aligned}
& \max\{f(\bar{w}) - f(w^*), h(\bar{w})\} \\
\leq\ & 2\left(\frac{C_1C_2}{T}\right)^{\frac{1}{2}} + 2\left(\frac{C_3C_4^2}{T}\right)^{\frac{1}{3}} + 3\left(\frac{C_1C_4C_5}{T}\right)^{\frac{1}{3}} + 3\left(\frac{C_1^2C_4^2C_6}{T^2}\right)^{\frac{1}{5}} + 4DG\left(\frac{\log\frac{4}{\delta}}{T}\right)^{\frac{1}{2}} \\
\leq\ & \sqrt{2}RGT^{-\frac{1}{2}} + 8\sigma\left(\frac{2\Gamma_T^{\frac{3}{2}}}{T}\right)^{\frac{1}{3}} + 3(32R^2\gamma G\sigma)^{\frac{1}{3}}\left(\frac{\Gamma_T}{T}\right)^{\frac{1}{3}} + 6(R^4\alpha G^2\sigma^2)^{\frac{1}{5}}\left(\frac{\Gamma_T}{T^2}\right)^{1/5} \\
& + 4DG\left(\frac{\log\frac{4}{\delta}}{T}\right)^{\frac{1}{2}}
\end{aligned}$$

□

For later use in the complexity bounds, define $\mathcal{L}_q(u,\delta;c) \coloneqq \frac{\mathrm{e}^{q^2}}{u}\log^q\left(\mathrm{e}\vee\frac{c}{u\delta}\right),\ q>0,\ u>0,\ c>0.$ By Lemma D.4,

$$T \geq \mathcal{L}_q(u,\delta;c) \quad \Longrightarrow \quad \frac{\log^q(cT/\delta)}{T} \leq u.$$

Corollary F.4 (Constrained Case 1 Oracle Complexity for Families 1–3)**.** Under the assumptions of Theorem F.3, let $\varepsilon>0$, the iteration / oracle complexities to guarantee $f(\bar{w})-f(w^\star)\leq\varepsilon, h(\bar{w})\leq\varepsilon$ for Families 1-3 are listed as follows, since $B=1$ is selected as the batch size, the oracle complexity has the same order of the iteration complexity.

1. **Family 1.**

$$N,\,T=\mathcal{O}\left(\frac{R^2LG\sigma^2}{\varepsilon^4}\log\frac{R^2LG\sigma^2}{\varepsilon^4\delta}\vee\frac{\sigma^3}{\varepsilon^3}\log^{3/2}\frac{\sigma^3}{\varepsilon^3\delta}\vee\frac{R^2G^2+D^2G^2\log(1/\delta)}{\varepsilon^2}\right).$$

2. **Family 2.**

$$N,\,T=\mathcal{O}\left(\frac{\sigma^3}{\varepsilon^3}\log^{3/2}\frac{\sigma^3}{\varepsilon^3\delta}\vee\frac{R^2\ell G\sigma}{\varepsilon^3}\log\frac{R^2\ell G\sigma}{\varepsilon^3\delta}\vee\frac{R^2G^2+D^2G^2\log(1/\delta)}{\varepsilon^2}\right).$$

3. **Family 3.**

$$N,\,T=\mathcal{O}\left(\frac{\sigma^3}{\varepsilon^3}\log^{3/2}\frac{\sigma^3}{\varepsilon^3\delta}\vee\frac{R^2\gamma G\sigma}{\varepsilon^3}\log\frac{R^2\gamma G\sigma}{\varepsilon^3\delta}\vee\sqrt{\frac{R^4\alpha G^2\sigma^2}{\varepsilon^5}}\log^{1/2}\left(\frac{1}{\delta}\sqrt{\frac{R^4\alpha G^2\sigma^2}{\varepsilon^5}}\right)\vee\frac{R^2G^2+D^2G^2\log(1/\delta)}{\varepsilon^2}\right).$$

Proof. **Family 1**. To guarantee

$$\begin{aligned}&\max\{f(\bar{w})-f(w^\star),h(\bar{w})\}\\ \leq&Q_T^{(1,1)}\equiv\left(\sqrt{2}RG+4DG\sqrt{\log\frac{4}{\delta}}\right)T^{-1/2}+8\sigma\left(\frac{\Gamma_T^{\frac{3}{2}}}{T}\right)^{\frac{1}{3}}+3\left(32R^2LG\sigma^2\right)^{\frac{1}{4}}\left(\frac{\Gamma_T}{T}\right)^{\frac{1}{4}}\\ \leq&\varepsilon/2+\varepsilon/4+\varepsilon/4=\varepsilon\end{aligned}$$

with sufficient large T, where $\Gamma_T \coloneqq 2\log\frac{4T}{\delta}$ satisfies $T\Gamma_T\geq T\geq\frac{R^2LG}{8\sigma^2}$. By noting $T\geq\mathcal{L}_q(u,\delta;c)=\frac{\mathrm{e}^{q^2}}{u}\log^q(\mathrm{e}\vee\frac{c}{u\delta})$ is sufficient to show $\frac{\log^q\frac{cT}{\delta}}{T}\leq u$ (see Lemma D.4), we can choose T as follow to ensure $\max\{f(\bar{w})-f(w^\star),h(\bar{w})\}\leq\varepsilon$.

$$\begin{aligned}T\geq&\frac{R^2LG}{8\sigma^2}\vee(\varepsilon/2)^{-2}\left(\sqrt{2}RG+4DG\sqrt{\log\frac{4}{\delta}}\right)^2\\ &\vee\frac{(32\sqrt{2}\sigma)^3}{\varepsilon^3}\vee\mathrm{e}^{9/4}\log^{3/2}\left[\mathrm{e}\vee\frac{(32\sqrt{2}\sigma)^3}{\varepsilon^3}\cdot\frac{4}{\delta}\right]\\ &\vee\frac{64R^2LG\sigma^2}{(\varepsilon/12)^4}\mathrm{e}\log\left[\mathrm{e}\vee\frac{64R^2LG\sigma^2}{(\varepsilon/12)^4}\cdot\frac{4}{\delta}\right]\end{aligned}$$

Family 2. To guarantee

$$\begin{aligned}&\max\{f(\bar{w})-f(w^\star),h(\bar{w})\}\\ \leq&Q_T^{(2,1)}\equiv\left(\sqrt{2}RG+4DG\sqrt{\log\frac{4}{\delta}}\right)T^{-1/2}+8\sigma\left(\frac{2\Gamma_T^{\frac{3}{2}}}{T}\right)^{\frac{1}{3}}+3\left(32R^2\ell G\sigma\right)^{\frac{1}{3}}\left(\frac{\Gamma_T}{T}\right)^{\frac{1}{3}}\\ \leq&\varepsilon/2+\varepsilon/4+\varepsilon/4=\varepsilon\end{aligned}$$

with sufficient large T, where $\Gamma_T := 2\log\frac{4T}{\delta}$ satisfies $T\Gamma_T \geq T \geq \frac{R^2\ell G}{4\sqrt{2}\sigma^2}$. By noting $T \geq \mathcal{L}_q(u,\delta;c) = \frac{\mathrm{e}^{q^2}}{u}\log^q(\mathrm{e}\vee\frac{c}{u\delta})$ is sufficient to show $\frac{\log^q\frac{cT}{\delta}}{T} \leq u$ (see Lemma D.4), we can choose T as follow to ensure $\max\{f(\bar{w})-f(w^\star), h(\bar{w})\} \leq \varepsilon$.

$$
\begin{aligned}
T \geq & \frac{R^2\ell G}{4\sqrt{2}\sigma^2} \vee (\varepsilon/2)^{-2}\left(\sqrt{2}RG + 4DG\sqrt{\log\frac{4}{\delta}}\right)^2 \\
& \vee \frac{2(32\sqrt{2}\sigma)^3}{\varepsilon^3} \vee \mathrm{e}^{9/4}\log^{3/2}\left[\mathrm{e}\vee\frac{2(32\sqrt{2}\sigma)^3}{\varepsilon^3}\cdot\frac{4}{\delta}\right] \\
& \vee \frac{64R^2\ell G\sigma}{(\varepsilon/12)^3}\mathrm{e}\log\left[\mathrm{e}\vee\frac{64R^2\ell G\sigma}{(\varepsilon/12)^3}\cdot\frac{4}{\delta}\right]
\end{aligned}
$$

Family 3. To guarantee

$$
\begin{aligned}
& \max\{f(\bar{w})-f(w^\star), h(\bar{w})\} \\
\leq & Q_T^{(3,1)} \equiv \left(\sqrt{2}RG + 4DG\sqrt{\log\frac{4}{\delta}}\right)T^{-1/2} + 8\sigma\left(\frac{2\Gamma_T^{\frac{3}{2}}}{T}\right)^{\frac{1}{3}} + 3\left(32R^2\gamma G\sigma\right)^{\frac{1}{3}}\left(\frac{\Gamma_T}{T}\right)^{\frac{1}{3}} \\
& + 6(R^4\alpha G^2\sigma^2)^{\frac{1}{5}}\left(\frac{\Gamma_T}{T^2}\right)^{1/5} \\
\leq & \varepsilon/4+\varepsilon/4+\varepsilon/4+\varepsilon/4 = \varepsilon
\end{aligned}
$$

with sufficient large T, where $\Gamma_T := 2\log\frac{4T}{\delta}$ satisfies $T\Gamma_T \geq T \geq \frac{R^2\gamma G}{4\sqrt{2}\sigma^2}$ and $T\Gamma_T^{3/4} \geq 2^{-15/4}\left(\frac{R^2\sqrt{\alpha}G}{\sigma}\right)^{3/2}$. By noting $T \geq \mathcal{L}_q(u,\delta;c) = \frac{\mathrm{e}^{q^2}}{u}\log^q(\mathrm{e}\vee\frac{c}{u\delta})$ is sufficient to show $\frac{\log^q\frac{cT}{\delta}}{T} \leq u$ (see Lemma D.4), we can choose T as follow to ensure $\max\{f(\bar{w})-f(w^\star), h(\bar{w})\} \leq \varepsilon$.

$$
\begin{aligned}
T \geq & \frac{R^2\gamma G}{4\sqrt{2}\sigma^2} \vee 2^{-15/4}\left(\frac{R^2\sqrt{\alpha}G}{\sigma}\right)^{3/2} \vee (\varepsilon/4)^{-2}\left(\sqrt{2}RG + 4DG\sqrt{\log\frac{4}{\delta}}\right)^2 \\
& \vee \frac{2(32\sqrt{2}\sigma)^3}{\varepsilon^3}\mathrm{e}^{9/4}\log^{3/2}\left[\mathrm{e}\vee\frac{2(32\sqrt{2}\sigma)^3}{\varepsilon^3}\cdot\frac{4}{\delta}\right] \vee \frac{64R^2\gamma G\sigma}{(\varepsilon/12)^3}\mathrm{e}\log\left[\mathrm{e}\vee\frac{64R^2\gamma G\sigma}{(\varepsilon/12)^3}\cdot\frac{4}{\delta}\right] \\
& \vee \sqrt{\frac{2R^4\alpha G^2\sigma^2}{(\varepsilon/24)^5}}\mathrm{e}^{1/4}\log^{1/2}\left[\mathrm{e}\vee\sqrt{\frac{2R^4\alpha G^2\sigma^2}{(\varepsilon/24)^5}}\cdot\frac{4}{\delta}\right]
\end{aligned}
$$

□

F.2 Case 2 for Families 1–3 in the constrained setting

We now specialize Theorem F.1 to the probabilistic-reset setting with $p = \frac{1}{B} \in (0, 1]$. Since Application 2 estimates the scalar constraint value, we have $(\mathcal{G}, \|\cdot\|_{\mathcal{G}}) = (\mathbb{R}, |\cdot|)$, and hence $\kappa = 1$. Let $\Lambda_T := \log \frac{8T}{\delta}$, $C_T := C\big(\frac{\delta}{8T}, 1\big)$, and $\Gamma_T := 2\Lambda_T$. Then, for $T \geq 1$, we have $\Lambda_T > 1$, and $C_T = 1 + \sqrt{3\Lambda_T} \leq 2\sqrt{\Gamma_T}$.

Theorem F.5 (Constrained Case 2 Bounds for Families 1–3)**.** Assume the hypotheses of Theorem F.1. For each family below, let $\mathcal{A} = \{t \in [T] : v_t \leq \varepsilon\}$, $\varepsilon = \mathsf{OptError}(\eta, T, \delta) + \mathcal{E}$, and $\bar{w} = \frac{1}{|\mathcal{A}|} \sum_{t \in \mathcal{A}} w_t$.

1. **Family 1.** Assume that, with probability at least $1 - \delta/2$,

$$|e_t| \leq C_T \frac{\sigma}{\sqrt{B}} + \frac{GL\eta}{p} \Lambda_T, \qquad \forall t \in [T].$$

Choose $\eta = \min\left\{\frac{R}{G\sqrt{T}}, \left(\frac{R^6}{256GL\sigma^2\Lambda_T^2 T^3}\right)^{1/4}\right\}$, $B = \left\lceil \left(\frac{4\sigma^2 T}{R^2 GL}\right)^{1/2} \right\rceil$, $p = \frac{1}{B}$. Assume moreover that $T \geq \frac{R^2 GL}{4\sigma^2}$. Then, with probability at least $1 - \delta$, $f(\bar{w}) - f(w^\star) \leq Q_T^{(1,2)}$, $h(\bar{w}) \leq Q_T^{(1,2)}$, where

$$Q_T^{(1,2)} = \frac{RG + 2\sqrt{2}\, DG\sqrt{\log(4/\delta)}}{\sqrt{T}} + 4\left(\frac{16R^2 GL\sigma^2\Lambda_T^2}{T}\right)^{1/4}.$$

2. **Family 2.** Assume that, with probability at least $1 - \delta/2$,

$$|e_t| \leq C_T \frac{\sigma}{\sqrt{B}} + C_T \ell\eta G\sqrt{\frac{2\Lambda_T}{p}}, \qquad \forall t \in [T].$$

Choose $\eta = \min\left\{\frac{R}{G\sqrt{T}}, \left(\frac{R^4}{256\sigma\ell G\Lambda_T^{3/2} T^2}\right)^{1/3}\right\}$, $B = \left\lceil \left(\frac{32\sigma^4 T^2}{R^4\ell^2 G^2}\right)^{1/3} \right\rceil$, $p = \frac{1}{B}$. Assume moreover that $T \geq \frac{R^2\ell G}{4\sqrt{2}\,\sigma^2}$. Then, with probability at least $1 - \delta$, $f(\bar{w}) - f(w^\star) \leq Q_T^{(2,2)}$, $h(\bar{w}) \leq Q_T^{(2,2)}$, where

$$Q_T^{(2,2)} = \frac{RG + 2\sqrt{2}\, DG\sqrt{\log(4/\delta)}}{\sqrt{T}} + 3\left(\frac{32R^2\sigma\ell G\Lambda_T^{3/2}}{T}\right)^{1/3}.$$

3. **Family 3.** Assume that, with probability at least $1 - \delta/2$,

$$|e_t| \leq C_T \frac{\sigma}{\sqrt{B}} + \frac{\alpha\eta^2 G^2}{2p}\Lambda_T + C_T\gamma\eta G\sqrt{\frac{2\Lambda_T}{p}}, \qquad \forall t \in [T].$$

Choose $\eta = \min\left\{\frac{R}{G\sqrt{T}}, \left(\frac{R^4}{256\sigma\gamma G\Lambda_T^{3/2} T^2}\right)^{1/3}, \left(\frac{R^6}{1024\alpha G^2\sigma^2\Lambda_T^2 T^3}\right)^{1/5}\right\}$, $B = \left\lceil \min\left\{\left(\frac{32\sigma^4 T^2}{R^4\gamma^2 G^2}\right)^{1/3}, \left(\frac{32768\sigma^6\Lambda_T T^4}{\alpha^2 R^8 G^4}\right)^{1/5}\right\}\right\rceil$, $p = \frac{1}{B}$. Assume moreover that $T \geq \frac{R^2\gamma G}{4\sqrt{2}\sigma^2}$, $2^{3/2} T^2 \Lambda_T^{1/2} \geq \frac{\alpha R^4 G^2}{64\sigma^3}$. Then, with probability at least $1 - \delta$, $f(\bar{w}) - f(w^\star) \leq Q_T^{(3,2)}$, $h(\bar{w}) \leq Q_T^{(3,2)}$, where

$$Q_T^{(3,2)} = \frac{RG + 2\sqrt{2}\, DG\sqrt{\log(4/\delta)}}{\sqrt{T}} + 3\left(\frac{32R^2\sigma\gamma G\Lambda_T^{3/2}}{T}\right)^{\frac{1}{3}} + 5\left(\frac{\alpha R^4\sigma^2 G^2\Lambda_T^2}{T^2}\right)^{\frac{1}{5}}.$$

Proof. Since each estimator bound holds uniformly for all $t \in [T]$ (substituting the high-probability upperbound $\mathcal{E}$ of $|e_t|$ from the instantiations of Theorem 4.2, replacing δ with $\delta/2$ and selecting $B = 1/p$ in the bound of Table 2), (76) in Theorem F.1 holds. Hence, with $\varepsilon = \mathsf{OptError}(\eta, T, \delta) + \mathcal{E}$, Theorem F.1 yields

$$f(\bar{w}) - f(w^\star) \le Q, \qquad h(\bar{w}) \le Q, \qquad Q := \mathsf{OptError}(\eta, T, \delta) + 2\mathcal{E}.$$

In every family,

$$\frac{2DG}{\sqrt{T}}\sqrt{2\log\frac{4}{\delta}} = \frac{2\sqrt{2}\,DG\sqrt{\log(4/\delta)}}{\sqrt{T}}.$$

Also,

$$2C_T\frac{\sigma}{\sqrt{B}} \le 4\sigma\sqrt{\frac{\Gamma_T}{B}} = 4\sigma\sqrt{p\Gamma_T} = 4\sigma\sqrt{2p\Lambda_T}.$$

For **Family 1**,

$$Q \le \frac{R^2}{2\eta T} + \frac{\eta G^2}{2} + \frac{2\sqrt{2}\,DG\sqrt{\log(4/\delta)}}{\sqrt{T}} + 4\sigma\sqrt{2p\Lambda_T} + \frac{2GL\Lambda_T}{p}\eta.$$

Set $C_1 = \frac{R^2}{2}, C_2 = \frac{G^2}{2}, C_4 = 4\sigma\sqrt{2\Lambda_T}, C_5 = 2GL\Lambda_T$. Then

$$Q \le \frac{C_1}{\eta T} + C_2\eta + C_4\sqrt{p} + \frac{C_5\eta}{p} + \frac{2\sqrt{2}\,DG\sqrt{\log(4/\delta)}}{\sqrt{T}}.$$

Lemma D.3 with $C_3 = 0$ gives the stated choice of (η, B, p) and yields

$$\frac{C_1}{\eta T} + C_2\eta + C_4\sqrt{p} + \frac{C_5\eta}{p} \le \frac{RG}{\sqrt{T}} + 4\left(\frac{16R^2GL\sigma^2\Lambda_T^2}{T}\right)^{1/4}.$$

For **Family 2**, using $C_T \le 2\sqrt{\Gamma_T} = 2\sqrt{2\Lambda_T}$, we get

$$Q \le \frac{R^2}{2\eta T} + \frac{\eta G^2}{2} + \frac{2\sqrt{2}\,DG\sqrt{\log(4/\delta)}}{\sqrt{T}} + 4\sigma\sqrt{2p\Lambda_T} + 8\ell G\Lambda_T\frac{\eta}{\sqrt{p}}.$$

Set $C_1 = \frac{R^2}{2}, C_2 = \frac{G^2}{2}, C_4 = 4\sigma\sqrt{2\Lambda_T}, C_5 = 8\ell G\Lambda_T$. Then

$$Q \le \frac{C_1}{\eta T} + C_2\eta + C_4\sqrt{p} + \frac{C_5\eta}{\sqrt{p}} + \frac{2\sqrt{2}\,DG\sqrt{\log(4/\delta)}}{\sqrt{T}}.$$

Lemma D.3 yields

$$\frac{C_1}{\eta T} + C_2\eta + C_4\sqrt{p} + \frac{C_5\eta}{\sqrt{p}} \le \frac{RG}{\sqrt{T}} + 3\left(\frac{32R^2\sigma\ell G\Lambda_T^{3/2}}{T}\right)^{1/3}.$$

For **Family 3**, the same estimate gives

$$Q \le \frac{R^2}{2\eta T} + \frac{\eta G^2}{2} + \frac{2\sqrt{2}\,DG\sqrt{\log(4/\delta)}}{\sqrt{T}} + 4\sigma\sqrt{2p\Lambda_T} + 8\gamma G\Lambda_T\frac{\eta}{\sqrt{p}} + \alpha G^2\Lambda_T\frac{\eta^2}{p}.$$

Set

$$C_1 = \frac{R^2}{2}, \quad C_2 = \frac{G^2}{2}, \quad C_4 = 4\sigma\sqrt{2\Lambda_T}, \quad C_5 = 8\gamma G\Lambda_T, \quad C_6 = \alpha G^2\Lambda_T.$$

Then

$$Q \le \frac{C_1}{\eta T} + C_2\eta + C_4\sqrt{p} + \frac{C_5\eta}{\sqrt{p}} + \frac{C_6\eta^2}{p} + \frac{2\sqrt{2}\,DG\sqrt{\log(4/\delta)}}{\sqrt{T}}.$$

Lemma D.3 yields

$$\frac{C_1}{\eta T} + C_2\eta + C_4\sqrt{p} + \frac{C_5\eta}{\sqrt{p}} + \frac{C_6\eta^2}{p} \le \frac{RG}{\sqrt{T}} + 3\left(\frac{32R^2\sigma\gamma G\Lambda_T^{3/2}}{T}\right)^{1/3} + 5\left(\frac{\alpha R^4\sigma^2G^2\Lambda_T^2}{T^2}\right)^{1/5}.$$

Combining these gives the three stated bounds. □

Corollary F.6 (Constrained Case 2 Iteration Complexity for Families 1–3)**.** Under the assumptions of Theorem F.5, let $\varepsilon > 0$.

1. **Family 1.** Define $A_1 := RG + 2\sqrt{2}\,DG\sqrt{\log(4/\delta)}$, $A_2 := 4(16R^2GL\sigma^2)^{1/4}$, $u_2 := \left(\frac{\varepsilon}{2A_2}\right)^4$. If $T \geq \frac{4A_1^2}{\varepsilon^2} \vee \mathcal{L}_2(u_2, \delta; 8) \vee \left\lceil \frac{R^2GL}{4\sigma^2} \right\rceil$, then $f(\bar{w}) - f(w^\star) \leq \varepsilon$, $h(\bar{w}) \leq \varepsilon$. Moreover,
$$T = \mathcal{O}\left(\frac{R^2GL\sigma^2}{\varepsilon^4} \log^2 \frac{R^2GL\sigma^2}{\varepsilon^4\delta} \vee \frac{R^2G^2 + D^2G^2\log(1/\delta)}{\varepsilon^2} \right).$$

2. **Family 2.** Define $A_1 := RG + 2\sqrt{2}\,DG\sqrt{\log(4/\delta)}$, $A_2 := 3(32R^2\sigma\ell G)^{1/3}$, $u_2 := \left(\frac{\varepsilon}{2A_2}\right)^3$. If $T \geq \frac{4A_1^2}{\varepsilon^2} \vee \mathcal{L}_{\frac{3}{2}}(u_2, \delta; 8) \vee \left\lceil \frac{R^2\ell G}{4\sqrt{2}\sigma^2} \right\rceil$, then $f(\bar{w}) - f(w^\star) \leq \varepsilon$, $h(\bar{w}) \leq \varepsilon$. Moreover,
$$T = \mathcal{O}\left(\frac{R^2\sigma\ell G}{\varepsilon^3} \log^{3/2} \frac{R^2\sigma\ell G}{\varepsilon^3\delta} \vee \frac{R^2G^2 + D^2G^2\log(1/\delta)}{\varepsilon^2} \right).$$

3. **Family 3.** Define $A_1 := RG + 2\sqrt{2}\,DG\sqrt{\log(4/\delta)}$, $A_2 := 3(32R^2\sigma\gamma G)^{1/3}$, $A_3 := 5\left(\alpha R^4\sigma^2G^2\right)^{1/5}$, $u_2 := \left(\frac{\varepsilon}{3A_2}\right)^3$, $u_3 := \left(\frac{\varepsilon}{3A_3}\right)^5$. If $T \geq \frac{9A_1^2}{\varepsilon^2} \vee \mathcal{L}_{\frac{3}{2}}(u_2, \delta; 8) \vee \mathcal{L}_1\left(\sqrt{u_3}, \delta; 8\right) \vee \left\lceil \frac{R^2\gamma G}{4\sqrt{2}\sigma^2} \right\rceil \vee \left\lceil \left(\frac{\alpha R^4G^2}{64\sigma^3}\right)^{1/2} \right\rceil$, then $f(\bar{w}) - f(w^\star) \leq \varepsilon$, $h(\bar{w}) \leq \varepsilon$. Moreover,
$$T = \mathcal{O}\left(\frac{R^2\sigma\gamma G}{\varepsilon^3} \log^{3/2} \frac{R^2\sigma\gamma G}{\varepsilon^3\delta} \vee \sqrt{\frac{\alpha R^4\sigma^2G^2}{\varepsilon^5}} \log\left(\frac{1}{\delta}\sqrt{\frac{\alpha R^4\sigma^2G^2}{\varepsilon^5}} \right) \vee \frac{R^2G^2 + D^2G^2\log(1/\delta)}{\varepsilon^2} \right).$$

In all three families, the expected oracle complexity is $N = [B\cdot p + 1\cdot(1-p)]T = (2 - 1/B)T < 2T$, hence is of the same order as T.

Proof. As in the previous corollary, it is enough to make every term of the corresponding Q-bound at most ε/r, where $r = 2$ for Families 1 and 2 and $r = 3$ for Family 3. Since $\Lambda_T = \log(8T/\delta)$, we use
$$\frac{\Lambda_T^2}{T} \leq u \Leftarrow T \geq \mathcal{L}_2(u, \delta; 8)\,,$$
$$\frac{\Lambda_T^{3/2}}{T} \leq u \Leftarrow T \geq \mathcal{L}_{\frac{3}{2}}(u, \delta; 8)\,,$$
and
$$\frac{\Lambda_T^2}{T^2} \leq u \Leftarrow T \geq \mathcal{L}_1\left(\sqrt{u}, \delta; 8\right).$$
Applying these implications to the three Q-bounds, together with the side conditions in the theorem, gives the stated lower bounds on T. The displayed asymptotic orders follow by collecting the largest ε-dependences. Finally, $N = (2 - 1/B)T < 2T$. □

F.3 Case 3 for Families 1–3 in the constrained setting

In Case 3, the reset period is deterministic. We write $E \in \mathbb{Z}_+$, $p_t = \mathbb{I}_{\{t \bmod E = 0\}}$, $B = E$. Since Application 2 estimates the scalar constraint value, we have $(\mathcal{G}, \|\cdot\|_{\mathcal{G}}) = (\mathbb{R}, |\cdot|)$, and hence $\kappa = 1$. Let $\Lambda_T := \log \frac{8T}{\delta}$, $C_T := C\big(\frac{\delta}{8T}, 1\big)$, and $\Gamma_T := 2\Lambda_T$. Then, for $T \geq 1$, we have $\Lambda_T > 1$, and $C_T = 1 + \sqrt{3\Lambda_T} \leq 2\sqrt{\Gamma_T}$.

Theorem F.7 (Constrained Case 3 Tuned Bounds for Families 1–3)**.** Assume the hypotheses of Theorem F.1. For each family below, let $\varepsilon := \mathsf{OptError}(\eta, T, \delta) + \mathcal{E}$, $\mathcal{A} = \{t \in [T] : v_t \leq \varepsilon\}$, and $\bar{w} = \frac{1}{|\mathcal{A}|} \sum_{t \in \mathcal{A}} w_t$.

1. **Family 1.** Assume that, with probability at least $1 - \delta/2$,

$$|e_t| \leq C_T \frac{\sigma}{\sqrt{E}} + GL\eta E, \qquad \forall t \in [T].$$

Choose $\eta = \min\left\{\frac{R}{G\sqrt{T}}, \left(\frac{R^6}{256GL\sigma^2\Lambda_T T^3}\right)^{1/4}\right\}$, $E = B = \left\lceil \left(\frac{4\sigma^2\Lambda_T T}{R^2 GL}\right)^{1/2} \right\rceil$. Assume moreover that $T\Lambda_T \geq \frac{R^2 GL}{4\sigma^2}$. Then, with probability at least $1 - \delta$, $f(\bar{w}) - f(w^\star) \leq Q_T^{(1,3)}$, $h(\bar{w}) \leq Q_T^{(1,3)}$, where

$$Q_T^{(1,3)} = \frac{RG + 2\sqrt{2}\, DG\sqrt{\log(4/\delta)}}{\sqrt{T}} + 4\left(\frac{16R^2 GL\sigma^2\Lambda_T}{T}\right)^{1/4}.$$

2. **Family 2.** Assume that, with probability at least $1 - \delta/2$,

$$|e_t| \leq C_T \frac{\sigma}{\sqrt{E}} + C_T \ell\eta G\sqrt{2E}, \qquad \forall t \in [T].$$

Choose $\eta = \min\left\{\frac{R}{G\sqrt{T}}, \left(\frac{R^4}{256\sigma\ell G\Lambda_T T^2}\right)^{1/3}\right\}$, $E = B = \left\lceil \left(\frac{32\sigma^4\Lambda_T T^2}{R^4\ell^2 G^2}\right)^{1/3} \right\rceil$. Assume moreover that $T\sqrt{2\Lambda_T} \geq \frac{R^2\ell G}{4\sigma^2}$. Then, with probability at least $1 - \delta$, $f(\bar{w}) - f(w^\star) \leq Q_T^{(2,3)}$, $h(\bar{w}) \leq Q_T^{(2,3)}$, where

$$Q_T^{(2,3)} = \frac{RG + 2\sqrt{2}\, DG\sqrt{\log(4/\delta)}}{\sqrt{T}} + 3\left(\frac{32R^2\sigma\ell G\Lambda_T}{T}\right)^{1/3}.$$

3. **Family 3.** Assume that, with probability at least $1 - \delta/2$,

$$|e_t| \leq C_T \frac{\sigma}{\sqrt{E}} + \frac{\alpha\eta^2 G^2 E}{2} + C_T\gamma\eta G\sqrt{2E}, \qquad \forall t \in [T].$$

Choose $\eta = \min\left\{\frac{R}{G\sqrt{T}}, \left(\frac{R^4}{256\sigma\gamma G\Lambda_T T^2}\right)^{1/3}, \left(\frac{R^6}{1024\alpha G^2\sigma^2\Lambda_T T^3}\right)^{1/5}\right\}$, $E = B = \left\lceil \min\left\{\left(\frac{32\sigma^4\Lambda_T T^2}{R^4\gamma^2 G^2}\right)^{1/3}, \left(\frac{32768\sigma^6\Lambda_T^3 T^4}{\alpha^2 R^8 G^4}\right)^{1/5}\right\}\right\rceil$. Assume moreover that $T\sqrt{2\Lambda_T} \geq \frac{R^2\gamma G}{4\sigma^2}$, $2^{3/2}T^2\Lambda_T^{3/2} \geq \frac{\alpha R^4 G^2}{64\sigma^3}$. Then, with probability at least $1 - \delta$, $f(\bar{w}) - f(w^\star) \leq Q_T^{(3,3)}$, $h(\bar{w}) \leq Q_T^{(3,3)}$, where

$$Q_T^{(3,3)} = \frac{RG + 2\sqrt{2}\, DG\sqrt{\log(4/\delta)}}{\sqrt{T}} + 3\left(\frac{32R^2\sigma\gamma G\Lambda_T}{T}\right)^{1/3} + 5\left(\frac{\alpha R^4\sigma^2 G^2\Lambda_T}{T^2}\right)^{1/5}.$$

Proof. As in Case 2, the uniform estimator bound (substituting the high-probability upperbound $\mathcal{E}$ of $|e_t|$ from the instantiations of Theorem 4.2, replacing δ with $\delta/2$ and selecting $B = E$ in the bound of Table 2) implies that Theorem F.1 applies and $Q := \mathsf{OptError}(\eta, T, \delta) + 2\mathcal{E}$. Since $C_T \le 2\sqrt{\Gamma_T}$, $2C_T\sigma/\sqrt{E} \le 4\sigma\sqrt{\Gamma_T}/\sqrt{E} = 4\sigma\sqrt{2\Lambda_T}/\sqrt{E}$.

For **Family 1**,

$$Q \le \frac{R^2}{2\eta T} + \frac{\eta G^2}{2} + \frac{2\sqrt{2}\,DG\sqrt{\log(4/\delta)}}{\sqrt{T}} + 4\sigma\frac{\sqrt{2\Lambda_T}}{\sqrt{E}} + 2GL\eta E.$$

Set $C_1 = \frac{R^2}{2}$, $C_2 = \frac{G^2}{2}$, $C_4 = 4\sigma\sqrt{2\Lambda_T}$, $C_5 = 2GL$. Introducing $p = 1/E$, this becomes

$$Q \le \frac{C_1}{\eta T} + C_2\eta + C_4\sqrt{p} + \frac{C_5\eta}{p} + \frac{2\sqrt{2}\,DG\sqrt{\log(4/\delta)}}{\sqrt{T}}.$$

Lemma D.3 yields $\frac{C_1}{\eta T} + C_2\eta + C_4\sqrt{p} + \frac{C_5\eta}{p} \le \frac{RG}{\sqrt{T}} + 4\left(\frac{16R^2GL\sigma^2\Lambda_T}{T}\right)^{1/4}$.

For **Family 2**, using $C_T \le 2\sqrt{\Gamma_T} = 2\sqrt{2\Lambda_T}$, we get

$$Q \le \frac{R^2}{2\eta T} + \frac{\eta G^2}{2} + \frac{2\sqrt{2}\,DG\sqrt{\log(4/\delta)}}{\sqrt{T}} + 4\sigma\frac{\sqrt{2\Lambda_T}}{\sqrt{E}} + 8\ell G\sqrt{\Lambda_T}\,\eta\sqrt{E}.$$

Set $C_1 = \frac{R^2}{2}$, $C_2 = \frac{G^2}{2}$, $C_4 = 4\sigma\sqrt{2\Lambda_T}$, $C_5 = 8\ell G\sqrt{\Lambda_T}$. After setting $p = 1/E$,

$$Q \le \frac{C_1}{\eta T} + C_2\eta + C_4\sqrt{p} + \frac{C_5\eta}{\sqrt{p}} + \frac{2\sqrt{2}\,DG\sqrt{\log(4/\delta)}}{\sqrt{T}}.$$

Lemma D.3 yields $\frac{C_1}{\eta T} + C_2\eta + C_4\sqrt{p} + \frac{C_5\eta}{\sqrt{p}} \le \frac{RG}{\sqrt{T}} + 3\left(\frac{32R^2\sigma\ell G\Lambda_T}{T}\right)^{1/3}$.

For **Family 3**, the same estimate gives

$$Q \le \frac{R^2}{2\eta T} + \frac{\eta G^2}{2} + \frac{2\sqrt{2}\,DG\sqrt{\log(4/\delta)}}{\sqrt{T}} + 4\sigma\frac{\sqrt{2\Lambda_T}}{\sqrt{E}} + 8\gamma G\sqrt{\Lambda_T}\,\eta\sqrt{E} + \alpha G^2\eta^2 E.$$

Set $C_1 = \frac{R^2}{2}$, $C_2 = \frac{G^2}{2}$, $C_4 = 4\sigma\sqrt{2\Lambda_T}$, $C_5 = 8\gamma G\sqrt{\Lambda_T}$, $C_6 = \alpha G^2$. After setting $p = 1/E$,

$$Q \le \frac{C_1}{\eta T} + C_2\eta + C_4\sqrt{p} + \frac{C_5\eta}{\sqrt{p}} + \frac{C_6\eta^2}{p} + \frac{2\sqrt{2}\,DG\sqrt{\log(4/\delta)}}{\sqrt{T}}.$$

Lemma D.3 yields

$$\frac{C_1}{\eta T} + C_2\eta + C_4\sqrt{p} + \frac{C_5\eta}{\sqrt{p}} + \frac{C_6\eta^2}{p} \le \frac{RG}{\sqrt{T}} + 3\left(\frac{32R^2\sigma\gamma G\Lambda_T}{T}\right)^{1/3} + 5\left(\frac{\alpha R^4\sigma^2G^2\Lambda_T}{T^2}\right)^{1/5}.$$

Combining these gives the bounds. □

Corollary F.8 (Constrained Case 3 Iteration Complexity for Families 1–3)**.** Under the assumptions of Theorem F.7, let $\varepsilon > 0$.

1. **Family 1.** Define $A_1 \coloneqq RG + 2\sqrt{2}\, DG\sqrt{\log(4/\delta)}$, $A_2 \coloneqq 4(16R^2GL\sigma^2)^{1/4}$, $u_2 \coloneqq \left(\frac{\varepsilon}{2A_2}\right)^4$. If $T \geq \frac{4A_1^2}{\varepsilon^2} \vee \mathcal{L}_1\left(u_2, \delta; 8\right) \vee \left\lceil \frac{R^2GL}{4\sigma^2} \right\rceil$, then $f(\bar{w}) - f(w^\star) \leq \varepsilon$, $h(\bar{w}) \leq \varepsilon$. Moreover,
$$T = \mathcal{O}\left(\frac{R^2GL\sigma^2}{\varepsilon^4} \log \frac{R^2GL\sigma^2}{\varepsilon^4\delta} \vee \frac{R^2G^2 + D^2G^2\log(1/\delta)}{\varepsilon^2}\right).$$
2. **Family 2.** Define $A_1 \coloneqq RG + 2\sqrt{2}\, DG\sqrt{\log(4/\delta)}$, $A_2 \coloneqq 3(32R^2\sigma\ell G)^{1/3}$, $u_2 \coloneqq \left(\frac{\varepsilon}{2A_2}\right)^3$. If $T \geq \frac{4A_1^2}{\varepsilon^2} \vee \mathcal{L}_1\left(u_2, \delta; 8\right) \vee \left\lceil \frac{R^2\ell G}{4\sqrt{2}\sigma^2} \right\rceil$, then $f(\bar{w}) - f(w^\star) \leq \varepsilon$, $h(\bar{w}) \leq \varepsilon$. Moreover,
$$T = \mathcal{O}\left(\frac{R^2\sigma\ell G}{\varepsilon^3} \log \frac{R^2\sigma\ell G}{\varepsilon^3\delta} \vee \frac{R^2G^2 + D^2G^2\log(1/\delta)}{\varepsilon^2}\right).$$
3. **Family 3.** Define $A_1 \coloneqq RG + 2\sqrt{2}\, DG\sqrt{\log(4/\delta)}$, $A_2 \coloneqq 3(32R^2\sigma\gamma G)^{1/3}$, $A_3 \coloneqq 5\left(\alpha R^4\sigma^2G^2\right)^{1/5}$, $u_2 \coloneqq \left(\frac{\varepsilon}{3A_2}\right)^3$, $u_3 \coloneqq \left(\frac{\varepsilon}{3A_3}\right)^5$. If $T \geq \frac{9A_1^2}{\varepsilon^2} \vee \mathcal{L}_1\left(u_2, \delta; 8\right) \vee \mathcal{L}_{1/2}\left(\sqrt{u_3}, \delta; 8\right) \vee \left\lceil \frac{R^2\gamma G}{4\sqrt{2}\sigma^2} \right\rceil \vee \left\lceil 2^{-3/4}\left(\frac{\alpha R^4G^2}{64\sigma^3}\right)^{1/2} \right\rceil$, then $f(\bar{w}) - f(w^\star) \leq \varepsilon$, $h(\bar{w}) \leq \varepsilon$. Moreover,
$$T = \mathcal{O}\left(\frac{R^2\sigma\gamma G}{\varepsilon^3} \log \frac{R^2\sigma\gamma G}{\varepsilon^3\delta} \vee \sqrt{\frac{\alpha R^4\sigma^2G^2}{\varepsilon^5}} \log^{1/2}\left(\frac{1}{\delta}\sqrt{\frac{\alpha R^4\sigma^2G^2}{\varepsilon^5}}\right) \vee \frac{R^2G^2 + D^2G^2\log(1/\delta)}{\varepsilon^2}\right).$$

In all three families, the deterministic oracle complexity satisfies $N = B\left\lfloor \frac{T}{E} \right\rfloor + \left(T - \left\lfloor \frac{T}{E} \right\rfloor\right) = T + (E-1)\left\lfloor \frac{T}{E} \right\rfloor < 2T$.

Proof. As before, it is enough to make each term in the relevant bound $Q_T^{(i,3)}$ at most ε/r, with $r = 2$ for Families 1 and 2 and $r = 3$ for Family 3. Since $\Lambda_T = \log(8T/\delta)$, we use
$$\frac{\Lambda_T}{T} \leq u \Leftarrow T \geq \mathcal{L}_1\left(u, \delta; 8\right),$$
and
$$\frac{\Lambda_T}{T^2} \leq u \Leftarrow T \geq \mathcal{L}_{1/2}\left(\sqrt{u}, \delta; 8\right).$$
Applying these implications to $Q_T^{(1,3)}, Q_T^{(2,3)}, Q_T^{(3,3)}$, together with the side conditions in the theorem, gives the stated lower bounds on T. The asymptotic statements follow by collecting the dominant ε-terms. The oracle bound is the displayed identity together with $\lfloor T/E \rfloor \leq T/E$. □

G Related Work

Variance-reduced stochastic estimation. The mathematical foundations of stochastic first-order methods have been extensively developed since the introduction of stochastic approximation, traditionally relying on uniformly bounded stochastic subgradients or globally bounded variance assumptions to establish convergence [116, 16, 48, 112, 103, 9, 113, 121, 19, 74, 119, 55]. In smooth nonconvex optimization, vanilla SGD attains an $\mathcal{O}(\varepsilon^{-4})$ oracle complexity for finding an ε-stationary point [46]. Variance reduction improves this benchmark through two broad routes: finite-sum control-variate methods such as SAG, SAGA, and SVRG [117, 76, 36, 18], which use periodic reference-gradient computation, and recursive estimators such as SARAH, SPIDER, STORM, and PAGE [106, 43, 33, 131, 87, 85, 27, 34, 60], which update an online estimator via stochastic differences. Under mean-squared Lipschitz gradients, recursive estimators achieve the optimal $\mathcal{O}(\varepsilon^{-3})$ complexity, matching lower-bound theory [7, 39]. Second-order variants further improve the dependence on problem constants using Hessian-vector products while preserving the $\mathcal{O}(\varepsilon^{-3})$ rate [6], and higher-order methods under Lipschitz Hessians can reach rates of $\mathcal{O}(\varepsilon^{-3.5})$ or better through cubic regularization or perturbed acceleration [140, 4, 132, 44]. More recent works have also pursued broader viewpoints, including filtering-based perspectives [141], manifold extensions such as SERENA [89], and unified adaptive theories that relax unbiasedness [122]. On the practical side, MARS shows that recursive variance reduction can substantially accelerate large-model training when combined with preconditioned updates [142]. Our framework differs from these lines by making the *correction term* explicit, thereby isolating the mechanism that aligns stale information from w_{t-1} with the current iterate w_t and yielding a unified analysis that also suggests new second-order variants.

High-probability guarantees. Most guarantees in stochastic optimization are stated only in expectation, which can mask rare but consequential failures. High-probability results are therefore especially important in nonconvex regimes where noise can substantially degrade the signal-to-noise ratio. For non-variance-reduced methods, such guarantees are available for convex SGD [59], normalized SGD with momentum achieving $\mathcal{O}(\varepsilon^{-4})$ complexity [31], and heavy-tailed stochastic optimization in arbitrary smooth norms [32]. For recursive variance-reduced methods, however, high-probability analyses remain scarce and largely estimator-specific: existing results for PAGE [85] and STORM-type methods [139] do not directly transfer across estimator families. A unified high-probability treatment for multiple recursive estimators has been missing. Related efforts under relaxed noise assumptions include bounds under bounded α-th central moments [120] and various heavy-tailed analyses [49, 51, 107, 54]. These works address the noise-model axis, whereas our focus is on the estimator-design axis. Our main technical tool is a novel vector-valued Freedman inequality for κ-smooth Banach spaces, building on [110, 77]; crucially, it accommodates the random predictable variance proxies and stopping-time intervals induced by algorithmic reset events, enabling a single, unified high-probability guarantee across all estimator families.

Relaxed variance assumptions. To move beyond globally bounded variance, recent work has introduced structural relaxations such as expected smoothness and ABC-type conditions [50, 79, 80, 68, 52], which control the variance by a constant plus terms depending on the gradient norm and suboptimality. The BG-0 condition is weaker still and captures realistic optimization trajectories without uniform noise bounds [2, 135, 30, 8, 70, 127, 38, 105]. Recent lower bounds show that finding an ε-stationary point under BG-0 requires at least $\Omega(\varepsilon^{-6})$ and $\Omega(\varepsilon^{-4})$ queries in smooth and mean-square smooth settings, respectively [45]. Our sub-Gaussian oracle model is stronger than these relaxations, and extending the present unified high-probability framework to the ABC or BG-0 regime remains an important open problem.

Mirror descent and non-Euclidean optimization. Mirror descent [104, 14] generalizes projected gradient descent to non-Euclidean geometries through Bregman divergences and is particularly natural on domains such as the probability simplex. High-probability convergence for stochastic mirror descent without variance reduction has been established in convex settings under sub-Gaussian noise [91] and extended to heavy-tailed settings via clipping [107]. In nonconvex problems, [32] obtain high-probability guarantees for normalized SGD with momentum in smooth norms. However, these analyses do not incorporate recursive variance-reduced estimators within the mirror-descent framework. By formulating our estimator-level analysis directly in a κ-smooth normed space, we decouple the concentration argument from the specific update rule and obtain guarantees that apply

when the estimator lives in the dual space, and iterates are updated by mirror descent. The parameter κ captures the geometric regularity, including $\kappa = 1$ in Hilbert spaces, $\kappa = q - 1$ in ℓ_q for $q \geq 2$, and $\kappa = \mathcal{O}(\log d)$ for smooth approximations of ℓ_∞ (see Appendix A for detailed discussion).

Constrained stochastic optimization. Stochastic optimization with expectation constraints (‡) requires estimating both objective and feasibility quantities from noisy observations. Standard approaches include primal-dual, saddle-point, and augmented-Lagrangian methods [102, 24, 15, 56, 143, 64, 17], which can be sensitive to dual tuning and bounded-domain projections. A primal-only alternative is the switching gradient method (SGM), originating in [111], which alternates between objective decrease and constraint reduction. Its mirror-descent variants address non-smoothness and non-Euclidean geometry [129, 123, 3, 124, 125, 126, 128, 13]. In the stochastic setting, [82] establish an $\mathcal{O}(\varepsilon^{-4})$ oracle complexity for convex expectation-constrained problems using batch averaging for constraint estimation, and related weakly convex analyses retain the same order [65, 88, 97, 75, 35, 40, 98]. For nonconvex problems, [86] combine momentum-based variance reduction with an inexact augmented-Lagrangian framework and obtain an $\mathcal{O}(\varepsilon^{-5})$ complexity in expectation; however, the method is a double-loop scheme, using PStorm as an inner subroutine. Our contribution is to show that replacing batch averaging in the SGM-style setting of [82] with recursive variance-reduced estimators from our unified framework yields improved high-probability oracle complexity bounds of order $\mathcal{O}(\varepsilon^{-3})$ in the settings studied here.

Complementary variance reduction and gradient estimation strategies. Beyond recursive estimators, variance can also be reduced through negative dependence. Antithetic variates are a classical example [58, 78, 118, 57, 138]. More recent work strengthens these guarantees through strongly isotonic assumptions and optimal-transport connections [62, 63] (cf. [5, 108, 20]). The RAMPAGE framework [61] extends this viewpoint to variational inequalities through geometric path integration that removes first-order bias. Separately, exploiting structural properties for estimation, such as low rank, data smoothness, and interpolation, leads to improved rates in canonical settings [84, 73, 29, 71, 28]. These directions are complementary to our estimator design and suggest promising ways to enrich the correction term $\mathcal{T}_t$ in future work.

Decentralized and federated optimization. Distributed and federated optimization must balance local computation, communication, and statistical heterogeneity [99, 115, 109, 136, 10, 137, 25, 72]. Gradient-tracking approaches such as Global Update Tracking (GUT) mitigate local drift without additional communication [1], while switching-based and partially participating schemes address failures caused by heterogeneity [69, 133, 26, 53, 67, 114, 66, 134, 96]. Incorporating switching-based mechanics with partial client participation mitigates systemic failures induced by statistical heterogeneity, appropriately shifting the algorithmic focus from maximum likelihood estimation [93, 94, 95] to worst-case distributionally robust optimization. Extending this robust methodology to fully adversarial environments, online learning in MDPs has likewise benefited from optimistic estimators with sublinear regret [101]. Since client drift compounds the temporal iterate movement that drives estimation error in our framework, extending the correction-term viewpoint to communication-constrained and decentralized settings may yield sharper communication-complexity guarantees.

Future directions. Several directions remain open. Extending the unified high-probability theory to relaxed noise models such as BG-0 [2, 45], and combining it with adaptive or hyperparameter-free step-size rules [122, 41, 81, 37, 92, 100, 23], are especially natural next steps. Another promising direction is to transport the estimator-level perspective developed here to decentralized and federated settings, where client drift and communication delays create additional sources of tracking error.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: As claimed, we provide a unified variance-reduction framework and a novel dimension-free Freedman inequality that together establish sharp logarithmic confidence dependence for unconstrained optimization and improve oracle complexity for expectation-constrained settings.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Throughout the paper and in the related work section, we clarify that while our unified framework structurally accommodates adaptive step sizes, our derived oracle complexities are currently limited to the constant step-size regime. Furthermore, the lack of relaxed noise assumptions remains a limitation, we clearly identify the extension to relaxed variance conditions (like BG-0) and decentralized settings as critical open problems for future research.

   Guidelines:

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

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: All assumptions are clearly stated and all theoretical results are rigorously proved in the appendix.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

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

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

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

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

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

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

8. **Experiments compute resources**

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

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The work conforms with the code of ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: The contributions are purely theoretical, and we do not believe the work has any immediate societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper poses no such risks

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [N/A]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

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