ON STOCHASTIC PROXIMAL LINESEARCH METHODS IN ℓ_p GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

In this work we focus on stochastic nonconvex composite optimization and stochastic linesearch frameworks with each backtracking search reusing a fixed batch of samples (common random number-enabled). Prevailing stochastic linesearch methods of this type are formulated in Euclidean geometry and afflicted by restrictive assumptions or prior knowledge of problem dependent parameters. Motivated by these observations, we develop stochastic Armijo linesearch methods with a non-Euclidean proximal term (ℓ_p-norm square with $1 < p \leq 2$), under unbiased stochastic zeroth- and first-order oracles. They allow the algorithm inputs such as stepsizes and batchsizes to be selected without knowledge of problem dependent parameters, and the analysis is free of restrictive assumptions such as strong growth conditions and bounded gradient. Both minibatch and variance-reduced variants are considered and corresponding complexity guarantees are derived with first-order stationarity residual measured in the dual norm. The results include the Euclidean case and a near-ℓ_1/ℓ_∞ specialization, extend to controlled inexact subproblem solves, and are supported by numerical comparisons with stochastic linesearch and variance-reduced baselines.

1 INTRODUCTION

We consider the composite stochastic optimization problem

$$\min_{x \in \mathbb{R}^d} \Phi(x) := f(x) + r(x), \tag{1}$$

where $f(x) := \mathbb{E}[F(x, \xi)]$ is smooth and possibly nonconvex, $\xi : \Omega \to \Xi \subseteq \mathbb{R}^m$ denotes a random variable, $\Xi \triangleq \{\xi(\omega) \mid \omega \in \Omega\}$, $F : \mathbb{R}^d \times \Xi \to \mathbb{R}$ is a real-valued function, $(\Omega, \mathcal{F}, \mathbb{P})$ denotes the associated probability space. $r : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is proper, closed, and convex. Proximal-gradient methods handle nonsmooth regularization and constraints, but their prescribed parameters often use unknown smoothness or noise bounds (Ghadimi & Lan, 2013; Ghadimi et al., 2016). Adaptive gradients (Duchi et al., 2011), learning-rate-free methods (Defazio & Mishchenko, 2023), and auto-conditioned methods (Lan et al., 2024) address the related parameter-selection difficulties. Recent deterministic methods adapt backtracking contractions (Cavalcanti et al., 2025) or combine local-curvature adaptation with acceleration in convex optimization (Borodich & Kovalev, 2026).

Stochastic linesearch methods test decrease using noisy observations. Existing algorithms include step search with probability oracles where independent samples are used across the oracle calls rather than reusing a fixed set of realizations (see Cartis & Scheinberg (2018); Paquette & Scheinberg (2020); Berahas et al. (2021); Jin et al. (2021)), and common random number (CRN) enabled schemes where a fixed batch is used for all trial tests (Vaswani et al. (2019); Vaswani & Babanezhad Harikandeh (2025); Tsukada & Iiduka (2025); Wang et al. (2025)). The former is more applicable to black-box settings such as in privacy-preserving federated learning but it requires more samples and can be less efficient. The latter, although more specific, is readily available across a host of stochastic optimization and machine learning pipelines, since a batch can be cached and reused during the linesearch at negligible cost. In this work we focus on the latter. Note that existing CRN enabled stochastic linesearch design and analysis necessitate either assumptions such as growth conditions/interpolation (Vaswani et al., 2019; Vaswani & Babanezhad Harikandeh, 2025), or knowledge of problem-dependent parameters (Vaswani et al., 2019; Tsukada & Iiduka, 2025). SLAM

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

obtains nonconvex stationarity guarantees using periodic stepsize resets, without problem constants as inputs (hence problem-parameter-free), but the analysis requires bounded true gradients (Wang et al., 2025). Moreover, these developments largely employ Euclidean geometry. Non-Euclidean proximal methods instead accommodate smoothness and stationarity in a primal/dual norm pair (Beck & Teboulle, 2003; Fatkhullin & He, 2024), with suitable geometries reducing dimension dependence in high-dimensional regimes (Xie et al., 2025).

The stochastic-oracle/sample complexity of stochastic first-order methods can be further improved via variance reduction. Stochastic first-order methods with single or minibatch sampling typically achieve sample complexity of $\mathcal{O}(\epsilon^{-4})$ (up to logarithmic factors) to obtain an ϵ stationary point under appropriate Lipschitz smoothness and bounded variance conditions. Among the variance reduced methods, SVRG and SARAH use control-variate and recursive estimators (Johnson & Zhang, 2013; Nguyen et al., 2017); SPIDER, SpiderBoost, and ProxSARAH use recursive estimators to obtain state-of-the-art nonconvex guarantees of $\mathcal{O}(\epsilon^{-3})$ (up to logarithmic factors) (Fang et al., 2018; Wang et al., 2019; Pham et al., 2020). Variance reduced methods with problem-parameter-free designs include AdaSPIDER, which adapts stepsizes in finite-sum optimization (Kavis et al., 2022), while AC-VR-SPG designs auto-conditioned stepsizes for optimizing expected-value functions under general distributions (Lan et al. (2024)). However, a problem-parameter-free stochastic linesearch method with variance reduction is still missing in the literature, especially in an ℓ_p geometry setting.

We fill these gaps by developing stochastic proximal linesearch (SPL) methods in ℓ_p geometry, $1 < p \leq 2$, with both minibatch (SPL-MB) and variance-reduced (SPL-VR) variants. The methods use unbiased stochastic zeroth- and first-order oracles, and allow the algorithm inputs such as stepsizes and batchsizes to be selected without knowledge of problem dependent parameters. Both methods are common random number enabled, reusing the same batch for building the direction and all trial tests within the same iteration, and the variance-reduced method uses the recursive gradient estimation. The analysis is free of assumptions such as growth conditions and bounded gradient. In addition, we make contributions in the following aspects.

- **Sample complexity.** With the normalized squared-ℓ_p proximal term, $1 < p \leq 2$, we derive sample complexity bounds for expected squared ℓ_{p^*} stationarity. For SPL-MB, horizon-proportional minibatches give $\mathcal{O}(\varepsilon^{-4})$ stochastic-oracle complexity; increasing minibatches retain this order up to logarithms. For SPL-VR, a fixed-budget SPIDER type schedule gives $\mathcal{O}(\varepsilon^{-3}\sqrt{\log(\mathrm{e}+1/\varepsilon)})$.
- **Geometry and computation.** The results include Euclidean and near-ℓ_1/ℓ_∞ cases, the latter with logarithmic dimension dependence under uniform control of problem data. Closed-form unconstrained steps, scalar box-constrained solves, and controlled inexact variants make the proximal computations explicit. Rosenbrock and logistic-regression experiments compare the methods with stochastic linesearch and variance-reduced baselines.

Organization. In Section 2 we discuss the basic assumptions, notation, and proximal geometry; Sections 3–4 present the two methods and their analysis; Sections 5–6 discuss specializations and experiments. Proofs and implementation details are in the appendices.

2 PRELIMINARIES

Fix $p \in (1, 2]$, let $p^* = p/(p-1)$, and write $\|\cdot\| := \|\cdot\|_p$ and $\|\cdot\|_* := \|\cdot\|_{p^*}$. For a set $S \subseteq \mathbb{R}^d$, define $\mathrm{dist}_{p^*}(0, S) := \inf_{v \in S} \|v\|_*$. Throughout, ∂ denotes the limiting subdifferential; it agrees with the convex subdifferential on r, and $\partial\Phi(x) = \nabla f(x) + \partial r(x)$ (Rockafellar & Wets, 1998).

Throughout the oracle-complexity statements, we aim to find an ϵ-stationary solution/point x_{out} such that

$$\mathbb{E}\,\mathrm{dist}^2_{p^*}(0, \partial\Phi(x_{\mathrm{out}})) \leq \varepsilon^2.$$

The same convention applies to the specialization discussed later in Section 5.1 ($p = 1$, $p^* = +\infty$). The analysis in this work may necessitate the following assumptions.

Assumption 1 (Unbiased stochastic oracles, bounded variance and Lipschitz smoothness)**.** *There is an open set $U \supseteq \mathrm{dom}(r)$ such that f is continuously differentiable on U and almost surely, $F(\cdot, \xi)$*

is continuously differentiable on U with gradient $G(\cdot,\xi)$. Given any pair $(x,\xi)\in\mathbb{R}^d\times\Xi$, we have access to stochastic oracles $F(\cdot,\xi)$ and $G(\cdot,\xi)$ such that

$$f(x)=\mathbb{E}[F(x,\xi)],\qquad \mathbb{E}[G(x,\xi)]=\nabla f(x). \tag{2}$$

There are finite constants $L>0$ and $\sigma_*\geq 0$ such that the following smoothness bound holds almost surely simultaneously for all $x,y\in\mathrm{dom}(r)$, and the second-moment bound holds for every $x\in\mathrm{dom}(r)$:

$$\|G(x,\xi)-G(y,\xi)\|_*\leq L\|x-y\|, \tag{3}$$

$$\mathbb{E}\|G(x,\xi)-\nabla f(x)\|_*^2\leq\sigma_*^2. \tag{4}$$

Assumption 2 (Lower boundedness)**.** $\Phi_{\inf}:=\inf_x\Phi(x)>-\infty$.

The above assumptions are all common assumptions in stochastic optimization literature, especially the ones regarding stochastic linesearch (Vaswani et al. (2019); Jiang & Stich (2023); Tsukada & Iiduka (2025); Wang et al. (2025)), parameter-free stochastic algorithms (Lan et al. (2024)), and variance reduction (Fang et al. (2018); Wang et al. (2019)).

We use the normalized squared-ℓ_p norm as the proximal geometry

$$h(z):=\frac{1}{2(p-1)}\|z\|^2. \tag{5}$$

For $p\in(1,2]$, the Ball–Carlen–Lieb inequality (Ball et al., 1994) implies that h is 1-strongly convex with respect to $\|\cdot\|_p$.

For $x\in\mathrm{dom}(r)$, $g\in\mathbb{R}^d$, and $t>0$, define the proximal displacement

$$z(g,t;x):=\operatorname*{arg\,min}_{u:\,x+u\in\mathrm{dom}(r)}\left\{\langle g,u\rangle+r(x+u)-r(x)+\frac{1}{t}h(u)\right\}. \tag{6}$$

The objective in (6) is proper, closed, and strongly convex, so the minimizer exists and is unique. We use its measurable version throughout. The dual-norm averaging estimate and the KKT, decrease, and path-comparison properties of this displacement are stated and proved in Appendix A; see Lemmas 1 and 2.

3 Stochastic Proximal Linesearch with Minibatch Sampling

We present the stochastic proximal linesearch method as in Algorithm 1. In particular, the function value estimator $\widehat{f}_k$ and the gradient estimator $\widehat{g}_k$ are estimated via a minibatch of samples. s is a maximum trial stepsize, and K is an externally prescribed iteration budget. We also predetermine a sequence of positive integer batch sizes $\{N_k\}_{k=0,\dots,K-1}$, and a stepsize reset set $\mathcal{R}\subseteq\{0,1,\dots\}$ containing zero. Note that Algorithm 1 uses one fresh batch of samples $\mathcal{S}_k$ at each iteration and reuses it to estimate function value at all trial points of a backtracking linesearch in that iteration. The reset set controls the frequency of resetting the initial stepsize to the maximal level and balances the stepsize length and trial cost. The acceptance test is a proximal form of the Armijo sufficient-decrease condition (Armijo, 1966).

When $p=2$, the proximal term becomes $h(z)=\|z\|_2^2/2$, and (6) gives $x_k+z_k(t)=\mathrm{prox}_{tr}(x_k-t\widehat{g}_k)$. The Armijo decrease term is then $c\|z_k(t)\|_2^2/t$. Thus, with matching batch sizes and reset schedule, the proximal updates and acceptance test coincide with those of SLAM (Wang et al., 2025). For $1<p<2$, the same construction uses a strongly convex squared-ℓ_p proximal term. In particular, the dimension-dependent choice in (27) yields the near-ℓ_1/ℓ_∞ specialization of Section 5.1, while retaining $p>1$ and hence the strong convexity needed by the analysis. For an externally prescribed budget K, the method is *problem-parameter-free*: s,β,c, the reset schedule, and the batch choices below can all be specified without knowing L, σ_*, or the initial objective gap. In particular, s need not satisfy a smoothness-dependent upper bound, and no problem-dependent minimum batch size is imposed.

Algorithm 1 Stochastic Proximal Linesearch with Minibatch Sampling (SPL-MB)

Require: $x_0 \in \mathrm{dom}(r)$, $s > 0$, $\beta, c \in (0,1)$, horizon/iteration budget $K \geq 1$, positive integer batch sizes $N_0, \ldots, N_{K-1}$, and $\mathcal{R} \ni 0$; set $t_{-1} = s$

for $k = 0, \ldots, K-1$ **do**
  Draw a fresh i.i.d. batch $\mathcal{S}_k$ of size N_k
  Form $\widehat{f}_k(x) = N_k^{-1} \sum_{\xi \in \mathcal{S}_k} F(x, \xi)$ and $\widehat{g}_k = N_k^{-1} \sum_{\xi \in \mathcal{S}_k} G(x_k, \xi)$
  Set $\tau_k = s$ if $k \in \mathcal{R}$, and $\tau_k = t_{k-1}$ otherwise
  for $t = \tau_k, \beta\tau_k, \beta^2\tau_k, \ldots$ **do**
    Compute $z_k(t) = z(\widehat{g}_k, t; x_k)$ from (6)
    if $\widehat{f}_k(x_k + z_k(t)) + r(x_k + z_k(t)) \leq \widehat{f}_k(x_k) + r(x_k) - 2ch(z_k(t))/t$ **then**
      Set $t_k = t$, $\Delta_k = z_k(t)$, and $x_{k+1} = x_k + \Delta_k$; **break**
    end if
  end for
end for
Draw $Y \sim \mathrm{Unif}\{0, \ldots, K-1\}$ independently of the algorithmic history and return $x_{\mathrm{out}} = x_{Y+1}$

3.1 Well-definedness and convergence rate

The following proposition guarantees that the backtracking linesearch in each iteration of Algorithm 1 terminates in finitely many steps and is therefore well-defined, a consequence of the algorithm mechanism and the samplewise Lipschitz smoothness condition (3).

Proposition 1 (Finite termination)**.** *Under Assumption 1, every backtracking loop terminates almost surely. With*

$$\underline{t} := \min\left\{s, \frac{2\beta(1-c)}{L(p-1)}\right\}, \qquad \rho := \frac{s}{\underline{t}}, \tag{7}$$

the accepted stepsizes satisfy $\underline{t} \leq t_k \leq \tau_k \leq s$. At most

$$M_{\mathrm{trial}} := 1 + \left\lceil \log_{1/\beta}\left(\max\left\{1, \frac{sL(p-1)}{2(1-c)}\right\}\right)\right\rceil \tag{8}$$

trials are considered in one iteration.

For Algorithm 1, we can derive the following convergence rate. The coefficients $\widehat{\Gamma}$ and C_{res} are independent of K. Their explicit formulas are given in (50) and (53) in Appendix A.

Theorem 1 (Convergence rate in terms of stationarity residual)**.** *Suppose Assumptions 1 and 2 hold, and x_0 is deterministic. Define the constants,*

$$C_0 := \frac{8(p-1)C_{\mathrm{res}}^2}{c}(\Phi(x_0) - \Phi_{\mathrm{inf}}),$$
$$C_1 := \sigma_*^2\left(\frac{8(p-1)C_{\mathrm{res}}^2\widehat{\Gamma}}{c} + \frac{2}{p-1}\right). \tag{9}$$

For every $K \geq 1$ and every deterministic sequence $N_k \geq 1$, Algorithm 1 satisfies

$$\mathbb{E}\,\mathrm{dist}_{p^*}^2(0, \partial\Phi(x_{\mathrm{out}})) \leq \frac{C_0 + C_1\sum_{k=0}^{K-1} N_k^{-1}}{K}. \tag{10}$$

Remark 1. *Theorem 1 requires neither a strong growth condition nor interpolation. No boundedness assumption on the true or stochastic gradients or on the iterates is needed, nor is Lipschitz continuity of r on its domain. For a batch size N fixed independently of K, it gives $C_0/K + C_1/N$; for growing batches, stationarity follows whenever the average reciprocal batch size tends to zero.*

The main difficulty in the proof is that the current batch determines both the trial displacements and the accepted stepsize. Consequently, unbiasedness at a fixed point does not imply that the sampled error at the accepted point is centered. Proposition 1 reduces the backtracking search to at most M_{trial} potential stepsizes. Conditional on the past, each has a population-gradient candidate $z(\nabla f(x_k), t; x_k)$ independent of the fresh batch. The proofs involve controlling the stability of the proximal map and the function increment errors at the population candidates, and then transferring this control to the accepted trial point. The details are in Appendix A.

3.2 TWO PRESCRIBED BATCH CHOICES

Next we discuss the complexity of Algorithm 1. We count independent samples by $\mathcal{N}_{\mathrm{samp}}$, stochastic-gradient evaluations by $\mathcal{N}_G$, sampled-function evaluations by $\mathcal{N}_F$, and proximal solves by $\mathcal{N}_{\mathrm{prox}}$. Note that reusing a sample at another point adds an evaluation, not an independent sample. In Corollary 1, when the horizon K is known, a batch proportional to that horizon keeps the reciprocal sum bounded.

Corollary 1 (Fixed batchsize)**.** *Under the assumptions of Theorem 1, fix a numerical $c_b > 0$ and, for a given $K \geq 1$, set*

$$N_k = N := \lceil c_b K \rceil, \qquad 0 \leq k < K. \tag{11}$$

Then

$$\mathbb{E}\,\mathrm{dist}^2_{p^*}(0, \partial\Phi(x_{\mathrm{out}})) \leq \frac{C_0 + C_1/c_b}{K}, \qquad \mathcal{N}_{\mathrm{samp}} = \mathcal{N}_G = K\lceil c_b K \rceil. \tag{12}$$

For $0 < \varepsilon < 1$, the sufficient theoretical choice

$$K = \max\{1, \lceil (C_0 + C_1/c_b)/\varepsilon^2 \rceil\}$$

gives expected squared stationarity at most ε^2 and stochastic-oracle complexity

$$\mathcal{N}_{\mathrm{samp}} = \mathcal{N}_G = \mathcal{O}(\varepsilon^{-4}). \tag{13}$$

Alternatively, the following approach specifies the batches without using the terminal horizon. The harmonic reciprocal sum costs a logarithmic factor in the stationarity bound.

Corollary 2 (Increasing batches)**.** *Under the assumptions of Theorem 1, set*

$$N_k = k + 1. \tag{14}$$

Then

$$\mathbb{E}\,\mathrm{dist}^2_{p^*}(0, \partial\Phi(x_{\mathrm{out}})) \leq \frac{C_0 + C_1(1 + \log K)}{K}, \qquad \mathcal{N}_{\mathrm{samp}} = \mathcal{N}_G = \frac{K(K+1)}{2}. \tag{15}$$

For a fixed problem, a sufficient horizon has order $\varepsilon^{-2}\log(\mathrm{e} + 1/\varepsilon)$, and the stochastic-oracle complexity is $\mathcal{O}(\varepsilon^{-4}\log^2(\mathrm{e} + 1/\varepsilon))$.

In both cases,

$$\mathcal{N}_F \leq (M_{\mathrm{trial}} + 1)\sum_{k=0}^{K-1} N_k, \qquad \mathcal{N}_{\mathrm{prox}} \leq M_{\mathrm{trial}} K. \tag{16}$$

Note that repeated Armijo trials reuse the batch and do not multiply the independent sample or gradient counts. These counts exclude the deterministic arithmetic cost of a proximal solve.

4 STOCHASTIC PROXIMAL LINESEARCH WITH VARIANCE REDUCTION

Variance reduction propagates paired gradient differences between periodic SPIDER refreshes (Fang et al., 2018). The Algorithm 2 leveraging such a technique is presented below. It periodically (at multiples of q) rebuilds the gradient estimator v_k using B samples, and uses b sampled gradient differences to track gradient changes between refreshes. The affine correction in H_k aligns the Armijo model with this recursive estimator ($\nabla H_k(x_k) = v_k$). We first leave B, b, q as prescribed inputs; a specific choice (using neither L nor σ_*) is discussed later for developing theoretical guarantees. At every iteration, the batch consists of i.i.d. samples independent of the history of the algorithm. Inner gradient differences use the same samples at both points x_k and x_{k-1}, and every Armijo trial reuses the current batch.

Note that the inner estimator need not be conditionally unbiased. For the following statements involving arbitrary batch parameters, we use externally prescribed deterministic integers $B, b \geq 1$ and $2 \leq q \leq K$. Estimator refreshes at $k = 0, q, 2q, \ldots$ are separate from the prescribed stepsize reset set $\mathcal{R}$.

Use the natural filtration $\mathcal{F}_0 := \sigma(x_0), \mathcal{F}_k := \sigma(x_0, \mathcal{S}_0, \ldots, \mathcal{S}_{k-1}), k \geq 1$, with the fresh batches from Algorithm 2, and write $e_k = v_k - \nabla f(x_k)$. The previous estimator and all past iterates are

Algorithm 2 SPL with Variance Reduction (SPL-VR)

Require: $x_0 \in \mathrm{dom}(r)$, $K \geq 2$, $s > 0$, $c, \beta \in (0,1)$, and $\mathcal{R} \ni 0$; set $t_{-1} = s$; set B, b, q.

for $k = 0, \ldots, K-1$ **do**
  if $k \bmod q = 0$ **then**
    Draw a fresh i.i.d. batch $\mathcal{S}_k$ of size B; set $m_k = B$
    $v_k = B^{-1} \sum_{\xi \in \mathcal{S}_k} G(x_k, \xi)$ and $a_k = 0$
  else
    Draw a fresh i.i.d. batch $\mathcal{S}_k$ of size b; set $m_k = b$
    $a_k = v_{k-1} - b^{-1} \sum_{\xi \in \mathcal{S}_k} G(x_{k-1}, \xi)$
    $v_k = b^{-1} \sum_{\xi \in \mathcal{S}_k} G(x_k, \xi) + a_k$
  end if
  $H_k(x) = m_k^{-1} \sum_{\xi \in \mathcal{S}_k} F(x, \xi) + \langle a_k, x - x_k \rangle$
  Set $\tau_k = s$ if $k \in \mathcal{R}$, and $\tau_k = t_{k-1}$ otherwise
  for $t = \tau_k, \beta\tau_k, \beta^2\tau_k, \ldots$ **do**
    Compute $z_k(t) = z(v_k, t; x_k)$
    if $H_k(x_k + z_k(t)) + r(x_k + z_k(t)) \leq H_k(x_k) + r(x_k) - 2ch(z_k(t))/t$ **then**
      $t_k = t$, $\Delta_k = z_k(t)$, $x_{k+1} = x_k + \Delta_k$; **break**
    end if
  end for
end for
Draw $Y \sim \mathrm{Unif}\{0, \ldots, K-1\}$ independently and return $x_{\mathrm{out}} = x_{Y+1}$

$\mathcal{F}_k$-measurable; the current i.i.d. batch is independent of $\mathcal{F}_k$. The affine correction a_k may depend on that batch. Since the affine correction preserves smoothness and gives $\nabla H_k(x_k) = v_k$, proximal descent ensures finite backtracking without unbiasedness of v_k. The following proposition also relates estimator error and displacement to stationarity.

Proposition 2 (Finite termination for every positive batch size)**.** *Under Assumption 1, every line search in Algorithm 2, including its general-parameter version, terminates almost surely. With* $\underline{t}$, M_{trial} *from (7)–(8),*

$$\underline{t} \leq t_k \leq \tau_k \leq s,$$

and at most M_{trial} *trials are tested per iteration. Write* $e_k = v_k - \nabla f(x_k)$*. For every* $t > 0$*, the sampled and population candidates at the same base point satisfy*

$$\|z(v_k, t; x_k) - z(\nabla f(x_k), t; x_k)\| \leq t\|e_k\|_*.$$

At the accepted point,

$$\mathrm{dist}_{p^*}(0, \partial\Phi(x_{k+1})) \leq \left(L + \frac{1}{(p-1)t_k}\right) \|\Delta_k\| + \|e_k\|_*.$$

Next we present a theorem that combines weighted descent with the SPIDER error recursion, using lower boundedness (Assumption 2). Its quantity A_K collects the initial objective gap, refresh noise, and compensation for recursive error and selection using the same batch. Controlling A_K therefore controls stationarity and guides the batch choices; the proof is in Appendix B. The explicit nonnegative A_K, positive $\overline{\Theta}$, and C_{res} below are defined in (73), (67), and (53), respectively.

Theorem 2 (VR stationarity)**.** *Under Assumptions 1 and 2, with deterministic* x_0*, fix an integer* $K \geq 2$ *and deterministic integers* $B, b \geq 1$ *and* $2 \leq q \leq K$*. Consider the general-parameter version of Algorithm 2 described above. Then*

$$\frac{c}{4} \sum_{k=0}^{K-1} \mathbb{E} \frac{h(\Delta_k)}{t_k} \leq A_K, \tag{17}$$

$$\sum_{k=0}^{K-1} \mathbb{E}\|e_k\|_*^2 \leq \frac{2A_K}{\overline{\Theta}}, \tag{18}$$

$$\mathbb{E}\,\mathrm{dist}_{p^*}^2(0, \partial\Phi(x_{\mathrm{out}})) \leq \frac{1}{K}\left(\frac{16(p-1)C_{\mathrm{res}}^2}{c} + \frac{4}{\overline{\Theta}}\right) A_K. \tag{19}$$

Also $\mathbb{E}[\Phi(x_K) - \Phi_{\mathrm{inf}}] \leq A_K$.

The above bound holds for all the displayed parameters; it does not assert that every batch choice makes A_K/K vanish. For a prescribed horizon $K \geq 2$, we use batches and a refresh period as follows:

$$B = K, \qquad q = \left\lceil \sqrt{\frac{K}{\log(\mathrm{e}+K)}} \right\rceil, \qquad b = \lceil q \log(\mathrm{e}+K) \rceil. \tag{20}$$

Here $q/b \leq 1/\log(\mathrm{e}+K) \to 0$ controls the accumulated estimator error, while B/q and b balance refresh and inner-update work. The prescribed schedule gives the following uniform guarantee.

Theorem 3 (VR complexity)**.** *Under Assumptions 1 and 2, with deterministic x_0 and (20), Algorithm 2 satisfies, for every $K \geq 2$,*

$$\mathbb{E}\,\mathrm{dist}^2_{p^*}(0, \partial\Phi(x_{\mathrm{out}})) \leq \frac{C}{K}, \qquad C := \left(\frac{16(p-1)C_{\mathrm{res}}^2}{c} + \frac{4}{\overline{\Theta}} \right) A_\infty, \tag{21}$$

where the finite horizon-independent A_∞ in (77) satisfies $A_K \leq A_\infty$. The oracle counts obey

$$\mathcal{N}_{\mathrm{samp}} \leq \mathcal{N}_G \leq 5K^{3/2}\sqrt{\log(\mathrm{e}+K)} + 3K. \tag{22}$$

For $0 < \varepsilon < 1$, the sufficient theoretical horizon

$$K = \max\{2, \lceil C/\varepsilon^2 \rceil\} \tag{23}$$

gives expected squared stationarity at most ε^2 and

$$\mathcal{N}_G, \mathcal{N}_{\mathrm{samp}} = \mathcal{O}\left((1 + C/\varepsilon^2)^{3/2} \sqrt{\log(\mathrm{e} + C/\varepsilon^2)} \right). \tag{24}$$

This is $\mathcal{O}(\varepsilon^{-3}\sqrt{\log(\mathrm{e}+1/\varepsilon)})$.

5 Geometry Specializations and Inexact Computation

5.1 Euclidean and near-ℓ_1/ℓ_∞ geometries

The choice $p = 2$ recovers Euclidean geometry. The normalized proximal function is singular at $p = 1$, which is an important case applicable to high dimensional problems Beck & Teboulle (2003); Xie et al. (2025). Even without the coefficient, the proximal function $\|\cdot\|_1^2$ is not strongly convex and the same analysis does not go through. On the other hand, we can recover this case via a near-ℓ_1/ℓ_∞ specialization.

For the near-ℓ_1/ℓ_∞ specialization, we can replace the smoothness and variance clauses of Assumption 1 by

$$\|G(x,\xi) - G(y,\xi)\|_\infty \leq L_{1,\infty}\|x - y\|_1, \tag{25}$$

$$\mathbb{E}\|G(x,\xi) - \nabla f(x)\|_\infty^2 \leq \sigma_\infty^2. \tag{26}$$

Here $L_{1,\infty} > 0$ and $\sigma_\infty \geq 0$ are finite; smoothness holds almost surely simultaneously for all feasible x, y, and the variance bound holds for every feasible x. These replace, rather than supplement, the original norm bounds. For $d \geq 2$, we can set

$$p = 1 + \frac{1}{2\log d}, \qquad p^* = 1 + 2\log d. \tag{27}$$

For this choice, for any $u, v \in \mathbb{R}^d$, we have $\|u\|_1 \leq \sqrt{\mathrm{e}}\|u\|_p$, $\|v\|_\infty \leq \|v\|_{p^*} \leq \sqrt{\mathrm{e}}\|v\|_\infty$. Applying both comparisons to the smoothness bound, and the dual comparison to the noise bound, gives the original oracle conditions with

$$L \leq \mathrm{e}L_{1,\infty}, \qquad \sigma_*^2 \leq \mathrm{e}\sigma_\infty^2. \tag{28}$$

Corollary 3 (ℓ_1/ℓ_∞ stationarity for the near-ℓ_1 choice)**.** *Let x_0 be deterministic and Assumption 2 hold. Suppose $d \geq 2$, choose p, p^* as in (27), and suppose the clauses of Assumption 1 hold with (3)–(4) replaced by (25)–(26). For the output of either method, Algorithm 1 or 2,*

$$\mathbb{E}\,\mathrm{dist}^2_\infty(0, \partial\Phi(x_{\mathrm{out}})) \leq \mathbb{E}\,\mathrm{dist}^2_{p^*}(0, \partial\Phi(x_{\mathrm{out}})). \tag{29}$$

Thus the stationarity rates and stochastic-oracle orders in Corollaries 1–2 and Theorem 3 hold also in ℓ_∞, with the converted L, σ_* in (28).

Appendix C proves the corollary and tracks all dimension factors. With s, β, c, c_b fixed and uniformly bounded $L_{1,\infty}$, σ_∞^2, and initial objective gaps across dimensions, the sample orders are

$$\mathcal{O}((\log d)^2\varepsilon^{-4}) \quad \text{(horizon-proportional minibatches)}, \qquad \widetilde{\mathcal{O}}((\log d)^{3/2}\varepsilon^{-3}) \quad \text{(VR)}.$$

5.2 Exact and inexact proximal computation

For $r \equiv 0$, the proximal displacement has the closed form

$$z_i(g,t) = -(p-1)t\|g\|_{p^*}^{2-p^*}|g_i|^{p^*-1}\operatorname{sgn}(g_i), \qquad i = 1,\dots,d, \tag{30}$$

with $z(g,t) = 0$ when $g = 0$. In particular,

$$z(g,t) = tz(g,1), \tag{31}$$

so a single $\mathcal{O}(d)$ computation supplies every trial displacement in one backtracking loop.

When r is an indicator function of a box region, the proximal displacement can also be computed efficiently (see Appendix D). A general convex r, however, need not admit such a closed-form formula.

Inexact solves must preserve both backtracking and stationarity, in addition to controlling proximal error as in deterministic methods (Schmidt et al., 2011; Bonettini et al., 2016). Fix $0 < \chi < 1 - c$ and $\eta_k = \eta_0/(k+1)$ with $\eta_0 > 0$. For $g_k = \widehat{g}_k$ or v_k, use a measurable trial candidate satisfying

$$\begin{aligned}&\operatorname{dist}_{p^*}\left(0, g_k + \frac{1}{t}\nabla h(\widetilde{z}_k(t)) + \partial r(x_k + \widetilde{z}_k(t))\right)\\ &\qquad \le \min\left\{\frac{\chi}{(p-1)t}\|\widetilde{z}_k(t)\|, \frac{1}{t}\eta_k\right\}.\end{aligned} \tag{32}$$

The relative tolerance ensures finite backtracking; by strong convexity, the absolute tolerance gives $\|\widetilde{z}_k(t) - z(g_k,t;x_k)\| \le \eta_k$. Since $\sum_k \eta_k^2 < \infty$, its cumulative contribution to the stationarity bound is finite. Both tolerances use no L or σ_*, and the Armijo tests are unchanged. Theorem 4 in Appendix E proves that the minibatch and VR stationarity rates and stochastic-oracle orders are preserved. Solver tolerances affect deterministic inner work, not the batch schedules.

6 Numerical Experiments

We evaluate SPL-MB and SPL-VR on stochastic Rosenbrock and nonconvex logistic-regression problems. Appendix F gives the full protocols, tuning rules, and numerical reference construction.

6.1 Fixed-minibatch stochastic Rosenbrock experiment

We compare SPL-MB, Euclidean SLAM (Wang et al., 2025), SLS0-1 (Vaswani et al., 2019), and tuned constant-stepsize SGD. SPL-MB uses $p = 1 + 1/(2\log d)$. All methods use batches of size 128 and $K = 24000$ iterations at eight dimensions between 100 and 2000; we report medians and interquartile ranges over five repetitions. These fixed-batch, final-iterate comparisons evaluate practical performance, not the prescribed-batch random-output complexity guarantees.

In Figure 1, SLAM has the smallest final objective and gradient-norm medians for $d \le 300$, while SPL-MB has the smallest medians at the tested dimensions $d \ge 500$. For the timing target $f(x) \le 10^{-3}$, SPL-MB is faster than SLAM at the tested dimensions $d \ge 500$. Both reach the target in all repetitions; SLS0-1 and tuned SGD do not reach it within the common cap of 96000 iterations.

6.2 Fixed-batch variance-reduced logistic-regression experiment

We compare SPL-VR, SpiderBoost, and SpiderBoost-M (Wang et al., 2019) on the nonconvex logistic-regression objectives for the LIBSVM data sets (Chang & Lin, 2011) `a9a` and `w8a`. All methods use full-gradient refreshes, inner batches of size 256, and the same refresh period; SPL-VR uses $p = 2$. We use five common-stream repetitions and numerical references f^*, not certified global minima. This fixed-batch protocol differs from the horizon-dependent schedule in Theorem 3.

Figure 2 shows that SPL-VR reaches a median objective gap below 10^{-8} in fewer refresh epochs and less algorithm time than both baselines on both data sets. All three methods eventually attain essentially the same final objective value. The additional ℓ_1-regularized experiment in Appendix F.3 shows rapid initial progress by SPL-VR and later performance comparable to Prox-SpiderBoost-M.

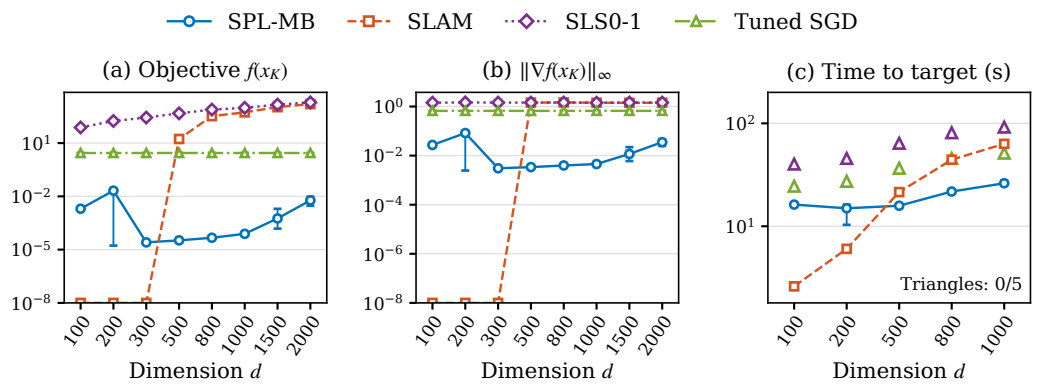


Figure 1: Rosenbrock: final population objective (left), final ℓ_∞ gradient norm (middle), and algorithm time to $f(x) \le 10^{-3}$ (right). Medians and interquartile ranges use five repetitions. Triangles denote median runtime at $K = 96000$ with $0/5$ target hits, not time to accuracy. The first two panels have a display floor of 10^{-8}.

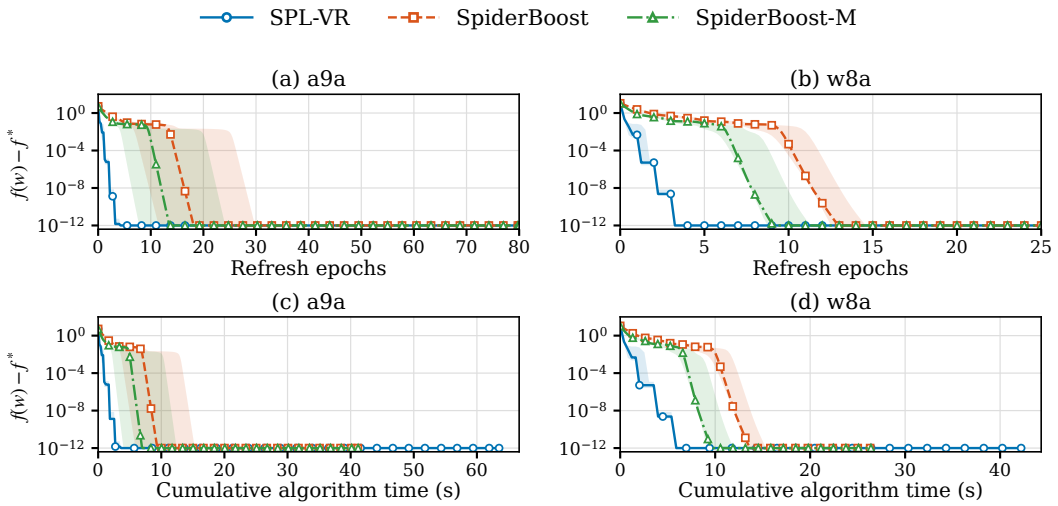


Figure 2: Nonconvex logistic-regression comparison. Top: median objective gap against refresh epochs on `a9a` (left) and `w8a` (right). Bottom: the same gaps against cumulative algorithm time. Shaded regions are interquartile ranges over five repetitions. Gaps below 10^{-12} are displayed at 10^{-12}.

7 Conclusion

We developed same-batch stochastic Armijo methods with normalized squared-ℓ_p proximal geometry, $1 < p \le 2$, in minibatch and variance-reduced variants. Their prescribed-budget execution requires neither smoothness nor variance constants, and the analysis avoids growth conditions and bounded gradients. For expected squared dual-norm stationarity at most ε^2, fixed-problem stochastic-oracle complexities are $\mathcal{O}(\varepsilon^{-4})$ for horizon-proportional minibatches and $\widetilde{\mathcal{O}}(\varepsilon^{-3})$ for variance reduction. Euclidean and near-ℓ_1/ℓ_∞ specializations, explicit proximal computations, and controlled inexact solves are included. Experiments illustrate their performance on smooth and nonsmooth composite problems.

AI USE STATEMENT

Generative AI tools assisted with checking and organizing proofs, organizing the initial draft, implementing experiments and checking experimental settings, preliminary analysis of results, literature searches, and translation. The authors reviewed and selectively adopted AI-assisted material and take full responsibility for the final content.

REPRODUCIBILITY STATEMENT

The assumptions, algorithms, and complexity guarantees are stated in Sections 2–4, with complete proofs in Appendices A and B. The proximal computations and inexact stopping criteria are documented in Section 5 and Appendices C–E. Complete experimental protocols, including data sets, parameter choices, tuning, sampling, and construction of numerical references, are provided in Appendix F. The anonymous supplementary code package contains the experimental implementations, saved numerical outputs, and commands for rerunning the experiments and regenerating the figures. Its README documents the dependencies and data sources.

REFERENCES

Larry Armijo. Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics*, 16(1):1–3, 1966. doi: 10.2140/pjm.1966.16.1.

Keith Ball, Eric A. Carlen, and Elliott H. Lieb. Sharp uniform convexity and smoothness inequalities for trace norms. *Inventiones Mathematicae*, 115(1):463–482, 1994. doi: 10.1007/BF01231769.

Amir Beck. *First-Order Methods in Optimization*, volume 25 of *MOS-SIAM Series on Optimization*. Society for Industrial and Applied Mathematics, 2017. doi: 10.1137/1.9781611974997.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003. doi: 10.1016/S0167-6377(02)00231-6.

A. S. Berahas, L. Cao, and K. Scheinberg. Global Convergence Rate Analysis of a Generic Line Search Algorithm with Noise. *SIAM Journal on Optimization*, 31(2):1489–1518, 2021. ISSN 1052-6234, 1095-7189. doi: 10.1137/19M1291832.

S. Bonettini, I. Loris, F. Porta, and M. Prato. Variable metric inexact line-search-based methods for nonsmooth optimization. *SIAM Journal on Optimization*, 26(2):891–921, 2016. doi: 10.1137/15M1019325.

Ekaterina Borodich and Dmitry Kovalev. Nesterov finds GRAAL: Optimal and adaptive gradient method for convex optimization. In *International Conference on Learning Representations*, 2026.

C. Cartis and K. Scheinberg. Global convergence rate analysis of unconstrained optimization methods based on probabilistic models. *Mathematical Programming*, 169(2):337–375, 2018. ISSN 0025-5610, 1436-4646. doi: 10.1007/s10107-017-1137-4.

Joao V. Cavalcanti, Laurent Lessard, and Ashia Wilson. Adaptive backtracking for faster optimization. In *International Conference on Learning Representations*, 2025.

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27, 2011. doi: 10.1145/1961189.1961199.

Aaron Defazio and Konstantin Mishchenko. Learning-rate-free learning by D-adaptation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 7449–7479, 2023.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. SPIDER: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Ilyas Fatkhullin and Niao He. Taming nonconvex stochastic mirror descent with general Bregman divergence. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3493–3501, 2024.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013. doi: 10.1137/120880811.

Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1–2):267–305, 2016. doi: 10.1007/s10107-014-0846-1.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Xiaowen Jiang and Sebastian U. Stich. Adaptive SGD with polyak stepsize and line-search: Robust convergence and variance reduction. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Billy Jin, Katya Scheinberg, and Miaolan Xie. High Probability Complexity Bounds for Line Search Based on Stochastic Oracles. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Sham M. Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Regularization techniques for learning with matrices. *Journal of Machine Learning Research*, 13(59):1865–1890, 2012. URL https://www.jmlr.org/papers/v13/kakade12a.html.

Ali Kavis, Stratis Skoulakis, Kimon Antonakopoulos, Leello Tadesse Dadi, and Volkan Cevher. Adaptive stochastic variance reduction for non-convex finite-sum minimization. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Guanghui Lan, Tianjiao Li, and Yangyang Xu. Projected gradient methods for nonconvex and stochastic smooth optimization: New complexities and auto-conditioned stepsizes, 2024. URL https://arxiv.org/abs/2412.14291. arXiv:2412.14291.

Lam M. Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2613–2621, 2017.

Courtney Paquette and Katya Scheinberg. A stochastic line search method with expected complexity analysis. *SIAM Journal on Optimization*, 30(1):349–376, 2020. doi: 10.1137/18M1216250.

Nhan H. Pham, Lam M. Nguyen, Dzung T. Phan, and Quoc Tran-Dinh. ProxSARAH: An efficient algorithmic framework for stochastic composite nonconvex optimization. *Journal of Machine Learning Research*, 21(110):1–48, 2020.

R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*, volume 317 of *Grundlehren der mathematischen Wissenschaften*. Springer, 1998. doi: 10.1007/978-3-642-02431-3.

H. H. Rosenbrock. An automatic method for finding the greatest or least value of a function. *The Computer Journal*, 3(3):175–184, 1960. doi: 10.1093/comjnl/3.3.175.

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

Yuki Tsukada and Hideaki Iiduka. Relationship between batch size and number of steps needed for nonconvex optimization of stochastic gradient descent using armijo-line-search learning rate. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=pqZ6nOm3WF.

Sharan Vaswani and Reza Babanezhad Harikandeh. Armijo line-search can make (Stochastic) gradient descent provably faster. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 61035–61070, 2025.

Sharan Vaswani, Aaron Mishkin, Issam Laradji, Mark Schmidt, Gauthier Gidel, and Simon Lacoste-Julien. Painless stochastic gradient: Interpolation, line-search, and convergence rates. In *Advances in Neural Information Processing Systems*, volume 32, pp. 3732–3745, 2019.

Qi Wang, Uday V. Shanbhag, and Yue Xie. A parameter-free stochastic linesearch method (SLAM) for minimizing expectation residuals, 2025. URL https://arxiv.org/abs/2512.14979. arXiv:2512.14979.

Zhe Wang, Kaiyi Ji, Yi Zhou, Yingbin Liang, and Vahid Tarokh. SpiderBoost and momentum: Faster variance reduction algorithms. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yue Xie, Jiawen Bi, and Hongcheng Liu. On tackling high-dimensional nonconvex stochastic optimization via stochastic first-order methods with non-smooth proximal terms and variance reduction, 2025. URL https://arxiv.org/abs/2509.13992. arXiv:2509.13992.

A PROOFS FOR SPL WITH MINIBATCH

The unbiasedness in (2) is an explicit oracle assumption. Jensen's inequality and (3) imply

$$\|\nabla f(x) - \nabla f(y)\|_* \le L\|x - y\|. \tag{33}$$

Consequently, integration along the segment joining x and y gives the sample and population Taylor bounds

$$|F(y,\xi) - F(x,\xi) - \langle G(x,\xi), y - x\rangle| \le \frac{L}{2}\|y - x\|^2, \tag{34}$$

$$|f(y) - f(x) - \langle \nabla f(x), y - x\rangle| \le \frac{L}{2}\|y - x\|^2. \tag{35}$$

For $p \in (1, 2]$, the Ball–Carlen–Lieb inequality (Ball et al., 1994) implies that h defined in (5) is 1-strongly convex with respect to $\|\cdot\|_p$:

$$h(u) - h(v) - \langle \nabla h(v), u - v\rangle \ge \frac{1}{2}\|u - v\|^2. \tag{36}$$

Moreover,

$$\langle \nabla h(z), z\rangle = \frac{1}{p-1}\|z\|^2 = 2h(z), \qquad \|\nabla h(z)\|_* = \frac{1}{p-1}\|z\|. \tag{37}$$

A.1 GEOMETRY AND PROXIMAL DISPLACEMENT

By Jensen's inequality, unbiasedness, and (3),

$$\|\nabla f(x) - \nabla f(y)\|_* \le \mathbb{E}\|G(x,\xi) - G(y,\xi)\|_* \le L\|x - y\|,$$

which proves (33). Since $\mathrm{dom}(r)$ is convex, integrating the sample gradient along $x + \theta(y - x)$ gives

$$\begin{aligned} &F(y,\xi) - F(x,\xi) - \langle G(x,\xi), y - x\rangle \\ &\quad = \int_0^1 \langle G(x + \theta(y - x),\xi) - G(x,\xi), y - x\rangle \, d\theta. \end{aligned}$$

The integrand is bounded in absolute value by $L\theta\|y - x\|^2$. Integration proves (34); the same argument using (33) proves (35).

For $p = 2$, (36) is the Euclidean quadratic identity. For $1 < p < 2$, fix $v, a \in \mathbb{R}^d$ and put $\phi(\theta) = h(v + \theta a)$. The case $a = 0$ is immediate. Omitting identically zero coordinates, away from the finitely many coordinate-zero crossings,

$$\begin{aligned} \phi''(\theta) &= \frac{2-p}{p-1}\|v + \theta a\|^{2-2p}\left(\sum_{j=1}^d |v_j + \theta a_j|^{p-1}\operatorname{sgn}(v_j + \theta a_j)a_j\right)^2 \\ &\quad + \|v + \theta a\|^{2-p}\sum_{j=1}^d |v_j + \theta a_j|^{p-2}a_j^2 \\ &\ge \|a\|^2. \end{aligned}$$

The last inequality follows from Hölder's inequality in the form

$$\|a\|_p^2 \le \left(\sum_{j=1}^d |z_j|^p\right)^{(2-p)/p}\sum_{j=1}^d |z_j|^{p-2}a_j^2.$$

At a crossing where $v + \theta a \neq 0$, the possible second-derivative singularities have order $|\theta - \theta_0|^{p-2}$ and are locally integrable because $p > 1$. If $v + \theta_0 a = 0$, then $\phi(\theta) = h(a)(\theta - \theta_0)^2$. Thus ϕ' is absolutely continuous across these crossings. Integrating the lower bound proves (36). For $z \neq 0$,

$$[\nabla h(z)]_j = \frac{1}{p-1}\|z\|^{2-p}|z_j|^{p-1}\operatorname{sgn}(z_j).$$

This formula, $(p-1)p^* = p$, and direct summation prove (37); both identities are immediate at $z = 0$.

Lemma 1 (Averaging in the dual norm)**.** *If $Z_1, \dots, Z_m$ are independent centered random vectors with $\mathbb{E}\|Z_i\|_*^2 \le \nu^2$, then*

$$\mathbb{E}\left\|\frac{1}{m}\sum_{i=1}^{m} Z_i\right\|_*^2 \le \frac{\nu^2}{(p-1)m}. \tag{38}$$

The same bound holds conditionally when the vectors are conditionally independent and centered and satisfy the corresponding conditional second-moment bound.

Proof. By strong convexity, the conjugate $h^*(v) = (p-1)\|v\|_*^2/2$ is 1-smooth in the dual norm (Kakade et al., 2012, Theorem 3), so $v \mapsto \|v\|_*^2/2$ is $1/(p-1)$-smooth. Set $S_i = m^{-1}\sum_{j=1}^{i} Z_j$. Conditional centering and smoothness give

$$\frac{1}{2}\mathbb{E}[\|S_i\|_*^2 \mid Z_1, \dots, Z_{i-1}] \le \frac{1}{2}\|S_{i-1}\|_*^2 + \frac{\mathbb{E}\|Z_i\|_*^2}{2(p-1)m^2}.$$

Iterating from $S_0 = 0$ proves (38). For the conditional statement, include the conditioning sigma-field in every filtration and repeat the same recursion. □

Lemma 2 (Proximal displacement properties)**.** *Let $z = z(g, t; x)$. There is $v \in \partial r(x+z)$ such that*

$$0 = g + v + \frac{1}{t}\nabla h(z), \tag{39}$$

and

$$\langle g, z\rangle + r(x+z) - r(x) + \frac{2}{t}h(z) \le 0. \tag{40}$$

If $0 < t_1 \le t_2$ and $z_i = z(g, t_i; x)$, then

$$\frac{t_1}{t_2}\|z_2\| \le \|z_1\| \le \|z_2\|. \tag{41}$$

Proof. The convex subdifferential sum rule gives (39). Pairing it with $-z$, using the subgradient inequality for r and (37), gives (40).

For the path comparison, uniqueness gives $z_1 = z_2$ if $t_1 = t_2$. For $t_1 < t_2$, put $\varphi(u) = \langle g, u\rangle + r(x+u) - r(x)$. Comparing the minimizers at t_1 and t_2 in their respective problems and adding yields

$$\left(\frac{1}{t_1} - \frac{1}{t_2}\right)(h(z_1) - h(z_2)) \le 0,$$

so $\|z_1\| \le \|z_2\|$. The optimality conditions give $-\nabla h(z_i)/t_i \in \partial\varphi(z_i)$. Monotonicity of $\partial\varphi$ and (37) imply

$$\frac{\|z_1\|^2}{t_1} + \frac{\|z_2\|^2}{t_2} \le \left(\frac{1}{t_1} + \frac{1}{t_2}\right)\|z_1\|\|z_2\|.$$

If $z_2 = 0$, then $z_1 = 0$. Otherwise, division by $\|z_2\|^2/t_1$ gives

$$\left(\frac{\|z_1\|}{\|z_2\|} - 1\right)\left(\frac{\|z_1\|}{\|z_2\|} - \frac{t_1}{t_2}\right) \le 0,$$

which proves (41). □

A.2 Finite candidates and function-increment error

For the analysis, let

$$\mathcal{F}_0 := \sigma(x_0), \qquad \mathcal{F}_k := \sigma(x_0, \mathcal{S}_0, \dots, \mathcal{S}_{k-1}), \quad k \ge 1, \tag{42}$$

completed by null sets. Then x_k, t_{k-1} and the population candidates are $\mathcal{F}_k$-measurable, while $\mathcal{S}_k$ is independent of $\mathcal{F}_k$.

Proof of Proposition 1. For a trial t, put $z = z_k(t)$. The sample function $\widehat{f}_k$ is L-smooth, so (34) and (40) give

$$\widehat{f}_k(x_k + z) + r(x_k + z) - \widehat{f}_k(x_k) - r(x_k) \le -\frac{2}{t}h(z) + L(p-1)h(z).$$

Thus the Armijo test holds whenever $t \le 2(1-c)/(L(p-1))$, proving finite termination and the trial-count bound. If a contraction occurs, the preceding trial exceeded $2(1-c)/(L(p-1))$, so the accepted trial exceeds $2\beta(1-c)/(L(p-1))$. If no contraction occurs, the accepted trial equals τ_k. Induction from $\tau_0 = s$ proves (7), independently of the reset set. □

For $i = 0, \dots, M_{\mathrm{trial}} - 1$, define

$$t_{k,i} := \beta^i \tau_k, \qquad z_{k,i} := z(\widehat{g}_k, t_{k,i}; x_k), \qquad \bar{z}_{k,i} := z(\nabla f(x_k), t_{k,i}; x_k). \tag{43}$$

If I_k is the accepted index, then $t_k = t_{k,I_k}$ and $\Delta_k = z_{k,I_k}$. Write $e_k := \widehat{g}_k - \nabla f(x_k)$.

Lemma 3 (Candidate stability)**.** *For every $x \in \mathrm{dom}(r)$, $t > 0$, and $g, g' \in \mathbb{R}^d$,*

$$\|z(g, t; x) - z(g', t; x)\| \le t\|g - g'\|_*.$$

In particular, every sampled and population candidate pair satisfies

$$\|z_{k,i} - \bar{z}_{k,i}\| \le t_{k,i}\|e_k\|_*. \tag{44}$$

Proof. Let $z = z(g, t; x)$ and $z' = z(g', t; x)$. Subtract the two proximal optimality conditions and pair with $z - z'$. Monotonicity of ∂r and (36) give

$$\frac{1}{t}\|z - z'\|^2 \le -\langle g - g', z - z'\rangle \le \|g - g'\|_*\|z - z'\|.$$

If $z = z'$ the claim is immediate; otherwise cancel $\|z - z'\|$. Taking $g = \widehat{g}_k$, $g' = \nabla f(x_k)$, and $t = t_{k,i}$ proves (44).

For every possible accepted index, (41) gives $\|\bar{z}_{k,i}\| \le \rho\|\bar{z}_{k,I_k}\|$: if $i \ge I_k$, the factor one suffices, while if $i < I_k$, the factor is at most $t_{k,i}/t_{k,I_k} \le \rho$. Hence

$$\|\bar{z}_{k,i}\| \le \rho\left(\|\Delta_k\| + s\|e_k\|_*\right).$$

Squaring, using $(a+b)^2 \le 2a^2 + 2b^2$ and $\|\Delta_k\|^2 = 2(p-1)h(\Delta_k)$, proves

$$\max_{0 \le i < M_{\mathrm{trial}}} \|\bar{z}_{k,i}\|^2 \le 4(p-1)\rho^2 h(\Delta_k) + 2\rho^2 s^2 \|e_k\|_*^2. \tag{45}$$

This path bound is also used for the inexact VR analysis in Appendix E. □

For later use, samplewise smoothness along $x + w + \theta(z - w)$ gives

$$\begin{aligned} &|F(x + z, \xi) - F(x + w, \xi) - \langle G(x, \xi), z - w\rangle| \\ &\quad \le L\|w\|\|z - w\| + \frac{L}{2}\|z - w\|^2. \end{aligned} \tag{46}$$

Indeed, $\|w + \theta(z - w)\| \le \|w\| + \theta\|z - w\|$ and integration over $\theta \in [0, 1]$ gives the displayed constants. The same bound holds for f and ∇f.

Define the same-batch function-increment error

$$\mathcal{D}_k(z) := \widehat{f}_k(x_k + z) - \widehat{f}_k(x_k) - f(x_k + z) + f(x_k). \tag{47}$$

Lemma 4 (Function-increment error)**.** *For $\mathcal{F}_k$-measurable feasible displacements z, w,*

$$\mathbb{E}[|\mathcal{D}_k(z) - \mathcal{D}_k(w)|^2 \mid \mathcal{F}_k] \le \frac{\sigma_*^2}{N_k}\|z - w\|^2. \tag{48}$$

Almost surely, the following bound holds simultaneously for all feasible z, w:

$$|\mathcal{D}_k(z) - \mathcal{D}_k(w)| \le \|\nabla\mathcal{D}_k(w)\|_*\|z - w\| + L\|z - w\|^2. \tag{49}$$

Proof. Conditional on $\mathcal{F}_k$, the segment $x_k + w + \theta(z-w)$ is fixed and lies in $\mathrm{dom}(r)$. The centered single-sample increment equals

$$\int_0^1 \langle G(x_k + w + \theta(z-w), \xi) - \nabla f(x_k + w + \theta(z-w)), z - w\rangle \, d\theta.$$

Its mean is zero and its second moment is at most $\sigma_*^2 \|z-w\|^2$, by integral Cauchy–Schwarz and (4). These bounds also justify interchanging expectation and integration. Independence of the scalar increments across samples divides this second moment by N_k, proving (48). Almost surely, $\nabla \mathcal{D}_k$ is $2L$-Lipschitz on the feasible displacement domain. Integrating its difference from $\nabla \mathcal{D}_k(w)$ along the same segment proves (49). □

A.3 Weighted descent and residual transfer

The analysis retains the accepted-step weight in the descent term. Define

$$\widehat{\Gamma} := \begin{cases} s, & L(p-1)s \le 1, \\ \dfrac{(p-1)sM_{\mathrm{trial}}}{2c} + \dfrac{s\sqrt{M_{\mathrm{trial}}}(1+\sqrt{p-1}) + Ls^2}{p-1}, & L(p-1)s > 1. \end{cases} \tag{50}$$

This case distinction concerns proof constants only. Neither the algorithm nor its batches need to know which case holds.

Lemma 5 (Weighted minibatch descent)**.** *For every deterministic integer $N_k \ge 1$,*

$$\mathbb{E}[\Phi(x_{k+1}) \mid \mathcal{F}_k] \le \Phi(x_k) - \frac{c}{2}\mathbb{E}\left[\frac{h(\Delta_k)}{t_k} \,\middle|\, \mathcal{F}_k\right] + \widehat{\Gamma}\frac{\sigma_*^2}{N_k}, \tag{51}$$

$$\mathbb{E}[\|e_k\|_*^2 \mid \mathcal{F}_k] \le \frac{\sigma_*^2}{(p-1)N_k}. \tag{52}$$

Define the weighted residual coefficient by

$$C_{\mathrm{res}}^2 := \max\left\{ s\left(L + \frac{1}{(p-1)s}\right)^2, \, \underline{t}\left(L + \frac{1}{(p-1)\underline{t}}\right)^2 \right\}. \tag{53}$$

At every accepted point,

$$\mathrm{dist}_{p^*}(0, \partial\Phi(x_{k+1})) \le \left(L + \frac{1}{(p-1)t_k}\right)\|\Delta_k\| + \|e_k\|_*, \tag{54}$$

$$\mathrm{dist}_{p^*}^2(0, \partial\Phi(x_{k+1})) \le 4(p-1)C_{\mathrm{res}}^2\frac{h(\Delta_k)}{t_k} + 2\|e_k\|_*^2. \tag{55}$$

Proof. The dual-averaging lemma at x_k gives (52). Within this proof put

$$A_k^{\mathrm{err}} := \max_i \frac{|\mathcal{D}_k(\bar{z}_{k,i})|}{\|\bar{z}_{k,i}\|}, \qquad B_k^{\mathrm{err}} := \max_i \|\nabla \mathcal{D}_k(\bar{z}_{k,i})\|_*,$$

with zero-displacement ratios set to zero. The candidates are $\mathcal{F}_k$-measurable. Lemma 4, dual averaging at $x_k + \bar{z}_{k,i}$, and $\max_i a_i^2 \le \sum_i a_i^2$ give

$$\mathbb{E}[(A_k^{\mathrm{err}})^2 \mid \mathcal{F}_k] \le \frac{M_{\mathrm{trial}}\sigma_*^2}{N_k}, \qquad \mathbb{E}[(B_k^{\mathrm{err}})^2 \mid \mathcal{F}_k] \le \frac{M_{\mathrm{trial}}\sigma_*^2}{(p-1)N_k}. \tag{56}$$

Candidate stability and the pathwise error-transfer inequality imply

$$|\mathcal{D}_k(\Delta_k)| \le A_k^{\mathrm{err}}\|\Delta_k\| + t_k(A_k^{\mathrm{err}} + B_k^{\mathrm{err}})\|e_k\|_* + Lt_k^2\|e_k\|_*^2. \tag{57}$$

Pointwise Young's inequality gives

$$A_k^{\mathrm{err}}\|\Delta_k\| \le \frac{c}{t_k}h(\Delta_k) + \frac{(p-1)t_k}{2c}(A_k^{\mathrm{err}})^2.$$

Use $t_k \le s$ only in the error terms. Conditional Cauchy–Schwarz, (52), and (56) yield

$$\mathbb{E}[-\mathcal{D}_k(\Delta_k) \mid \mathcal{F}_k] \le c\,\mathbb{E}\left[\frac{h(\Delta_k)}{t_k}\,\middle|\,\mathcal{F}_k\right] + \frac{\sigma_*^2}{N_k}\left[\frac{(p-1)sM_{\text{trial}}}{2c} + \frac{s\sqrt{M_{\text{trial}}}(1+\sqrt{p-1})+Ls^2}{p-1}\right].$$

Armijo therefore implies

$$\mathbb{E}[\Phi(x_{k+1}) \mid \mathcal{F}_k] \le \Phi(x_k) - c\,\mathbb{E}\left[\frac{h(\Delta_k)}{t_k}\,\middle|\,\mathcal{F}_k\right] + \frac{\sigma_*^2}{N_k}\left[\frac{(p-1)sM_{\text{trial}}}{2c} + \frac{s\sqrt{M_{\text{trial}}}(1+\sqrt{p-1})+Ls^2}{p-1}\right]. \tag{58}$$

No centering of $\langle e_k, \Delta_k\rangle$ has been used.

A second bound follows from population smoothness and proximal decrease:

$$\Phi(x_{k+1}) - \Phi(x_k) \le -\left(\frac{2}{t_k} - L(p-1)\right) h(\Delta_k) - \langle e_k, \Delta_k\rangle. \tag{59}$$

If $L(p-1)s \le 1$, then $L(p-1) \le 1/t_k$, and

$$-\langle e_k, \Delta_k\rangle \le \frac{h(\Delta_k)}{2t_k} + (p-1)t_k\|e_k\|_*^2$$

shows that

$$\Phi(x_{k+1}) - \Phi(x_k) \le -\frac{h(\Delta_k)}{2t_k} + (p-1)s\|e_k\|_*^2. \tag{60}$$

Taking expectations proves (51) in this case since $c < 1$. Otherwise (58) proves it with the second branch of (50).

The KKT condition gives a vector in $\partial\Phi(x_{k+1})$ equal to $\nabla f(x_{k+1}) - \widehat{g}_k - t_k^{-1}\nabla h(\Delta_k)$. Population smoothness and (37) prove (54). Squaring and using $\|\Delta_k\|^2 = 2(p-1)h(\Delta_k)$ gives

$$\operatorname{dist}_{p^*}^2(0, \partial\Phi(x_{k+1})) \le 4(p-1)t_k\left(L + \frac{1}{(p-1)t_k}\right)^2 \frac{h(\Delta_k)}{t_k} + 2\|e_k\|_*^2.$$

The function $t \mapsto L^2 t + 2L/(p-1) + 1/((p-1)^2 t)$ is convex on $(0,\infty)$. Its maximum on $[\underline{t}, s]$ is attained at an endpoint. This proves (55) without separately maximizing t_k and t_k^{-1}. □

Conditional on a fixed history, the finite population candidates are finite vectors. Stability and (52) give a finite conditional displacement moment; (57) gives conditional integrability of its model error. Subtract Φ_{inf} in (51) and place the nonnegative displacement term on the left. Starting from deterministic x_0, induction gives finite expected objective gaps and displacement moments at every finite iteration. Thus all following sums are justified without a bounded-iterate assumption.

A.4 Finite-budget bounds and oracle counts

Proof of Theorem 1. Sum (51) and use $\Phi \ge \Phi_{\text{inf}}$:

$$\frac{c}{2}\sum_{k=0}^{K-1}\mathbb{E}\frac{h(\Delta_k)}{t_k} \le \Phi(x_0) - \Phi_{\text{inf}} + \widehat{\Gamma}\sigma_*^2\sum_{k=0}^{K-1}\frac{1}{N_k}. \tag{61}$$

Substituting (61) and (52) into the sum of (55) gives $C_0 + C_1\sum_k N_k^{-1}$. Independence and uniformity of Y divide this sum by K. □

Proof of Corollary 1. For $N = \lceil c_b K\rceil$, one has $\sum_{k=0}^{K-1} N_k^{-1} = K/N \le 1/c_b$. Apply Theorem 1. The exact independent-sample and gradient count is KN, and $KN \le c_b K^2 + K$. With the displayed accuracy choice, $K \le 1 + (C_0 + C_1/c_b)/\varepsilon^2$, proving (13), including the case $C_0 = C_1 = 0$. □

Proof of Corollary 2. The reciprocal sum satisfies

$$\sum_{k=0}^{K-1} N_k^{-1} = \sum_{k=1}^{K} \frac{1}{k} \le 1 + \log K,$$

whereas $\sum_{k=0}^{K-1} N_k = K(K+1)/2$. Theorem 1 proves (15). For $K \ge 1$, $(C_0 + C_1(1+\log K))/K$ is nonincreasing. For fixed C_0, C_1, a sufficiently large constant multiple of $\varepsilon^{-2}\log(\mathrm{e}+1/\varepsilon)$ makes this bound at most ε^2, since the logarithm of that horizon is $\mathcal{O}(\log(\mathrm{e}+1/\varepsilon))$. Squaring the horizon order gives the claimed sample count. If $C_0 = C_1 = 0$, take $K = 1$. □

For either schedule, the gradients at x_k and the sampled value there can be cached. Each iteration uses at most M_{trial} proximal solves and $N_k(M_{\mathrm{trial}}+1)$ sampled-function evaluations. Summation proves (16).

B Proofs for SPL with Variance Reduction

Proof of Proposition 2. The affine term does not change the L-smoothness of H_k, and $\nabla H_k(x_k) = v_k$. Proximal optimality and the sampled Taylor bound give, for $z = z(v_k, t; x_k)$,

$$H_k(x_k+z) + r(x_k+z) - H_k(x_k) - r(x_k) \le -\left(\frac{2}{t} - L(p-1)\right) h(z).$$

Thus every trial with $t \le 2(1-c)/(L(p-1))$ is accepted, for every positive m_k. The reset and backtracking argument in Proposition 1 gives the stated stepsize and trial-count bounds. Lemma 3 applies with gradients v_k and $\nabla f(x_k)$. Finally, the KKT argument for (54) uses population smoothness and $e_k = v_k - \nabla f(x_k)$, not unbiasedness of v_k, so it gives the displayed stationarity bound as well. □

B.1 Model error and weighted descent

At a refresh let $U_k = e_k$, and at an inner iteration let $U_k = e_k - e_{k-1}$. Then

$$\mathbb{E}[U_k \mid \mathcal{F}_k] = 0, \qquad v_k - U_k = \begin{cases} \nabla f(x_k), & k \bmod q = 0, \\ \nabla f(x_k) + e_{k-1}, & k \bmod q \ne 0. \end{cases} \tag{62}$$

For $t_{k,i} = \beta^i \tau_k$, $0 \le i < M_{\mathrm{trial}}$, define the history-measurable comparison candidates

$$\widetilde{z}_{k,i} := z(v_k - U_k, t_{k,i}; x_k). \tag{63}$$

At an inner iteration they retain the historical error e_{k-1}. The algorithm computes neither U_k nor these comparison candidates.

Use the centered Taylor-remainder error

$$\mathcal{E}_k(z) := H_k(x_k+z) - H_k(x_k) - f(x_k+z) + f(x_k) - \langle e_k, z\rangle. \tag{64}$$

Since $v_k = m_k^{-1}\sum_{\xi\in\mathcal{S}_k} G(x_k,\xi) + a_k$, the affine correction cancels identically in this expression. For $Q(z,\xi) := F(x_k+z,\xi) - F(x_k,\xi) - \langle G(x_k,\xi), z\rangle$,

$$\mathcal{E}_k(z) = \frac{1}{m_k}\sum_{\xi\in\mathcal{S}_k} \big(Q(z,\xi) - \mathbb{E}[Q(z,\xi)\mid\mathcal{F}_k]\big)$$

at every history-measurable feasible z, where the expectation uses a fresh oracle sample. This centering requires neither conditional unbiasedness of v_k nor history measurability of a_k.

Lemma 6 (Compensated finite-candidate remainder)**.** *For every $\lambda > 0$ and any family of at most M_{trial} history-measurable feasible displacements,*

$$\mathbb{E}\left[\max_i \big(-\mathcal{E}_k(\widetilde{z}_{k,i}) - \lambda\|\widetilde{z}_{k,i}\|^2\big)_+ \,\middle|\, \mathcal{F}_k\right] \le \frac{\mathrm{e}M_{\mathrm{trial}}\sigma_*^2}{\lambda m_k}\exp\left(-\frac{m_k\lambda^2}{8L^2}\right). \tag{65}$$

Proof. Condition on $\mathcal{F}_k$ and fix a nonzero displacement z. Let $X_\ell = -Q(z, \xi_{k,\ell}) + \mathbb{E}Q(z, \xi)$. The Taylor bound puts Q in an interval of length $L\|z\|^2$, so $\mathbb{E}X_\ell = 0$, $|X_\ell| \le L\|z\|^2$, and the range length of X_ℓ is at most $L\|z\|^2$. Also

$$Q(z, \xi) - \mathbb{E}Q(z, \xi) = \int_0^1 \langle G(x_k + \theta z, \xi) - \nabla f(x_k + \theta z) - G(x_k, \xi) + \nabla f(x_k), z\rangle \, d\theta.$$

The uniform noise second moment and integral Cauchy–Schwarz imply $\mathbb{E}X_\ell^2 \le 4\sigma_*^2\|z\|^2$.

The sample mean satisfies $m_k^{-1}\sum_{\ell=1}^{m_k} X_\ell = -\mathcal{E}_k(z)$. If $\lambda \ge L$, then $-\mathcal{E}_k(z) \le L\|z\|^2 \le \lambda\|z\|^2$ and $(-\mathcal{E}_k(z) - \lambda\|z\|^2)_+ = 0$. Otherwise choose $u = m_k\lambda/(L^2\|z\|^2)$. Centering, Taylor's formula, and Hoeffding's lemma (Hoeffding, 1963) give, for $v > 0$,

$$1 \le \mathbb{E}e^{vX_\ell}, \qquad \mathbb{E}e^{vX_\ell} - 1 \le \frac{v^2 e^{vL\|z\|^2}}{2}\mathbb{E}X_\ell^2, \qquad \mathbb{E}e^{vX_\ell} \le e^{v^2L^2\|z\|^4/8}.$$

Expand $\prod_{\ell=1}^{m_k} \mathbb{E}e^{uX_\ell/m_k} - 1$ as a telescoping sum. Conditional independence and $\mathbb{E}\mathcal{E}_k(z) = 0$ yield

$$\mathbb{E}(e^{-u\mathcal{E}_k(z)} - 1 + u\mathcal{E}_k(z)) \le \frac{2u^2\sigma_*^2\|z\|^2}{m_k}\exp\left(\frac{uL\|z\|^2}{m_k} + \frac{u^2L^2\|z\|^4}{8m_k}\right). \tag{66}$$

For $y \ge 0$, Jensen's inequality with density $2(1-v)$ on $[0,1]$ shows that

$$e^y - 1 - y = y^2\int_0^1 (1-v)e^{vy}\,dv \ge \frac{y^2}{2}e^{y/3}.$$

For $-\mathcal{E}_k(z) \ge \lambda\|z\|^2$, one also has $-\mathcal{E}_k(z) - \lambda\|z\|^2 \le \mathcal{E}_k(z)^2/(4\lambda\|z\|^2)$. For $-\mathcal{E}_k(z) < \lambda\|z\|^2$, $e^{-u\mathcal{E}_k(z)} - 1 + u\mathcal{E}_k(z) \ge 0$. Hence, in both cases,

$$(-\mathcal{E}_k(z) - \lambda\|z\|^2)_+ \le \frac{e^{-u\lambda\|z\|^2/3}}{2\lambda\|z\|^2u^2}(e^{-u\mathcal{E}_k(z)} - 1 + u\mathcal{E}_k(z)).$$

Combine this with (66) and use $uL\|z\|^2/m_k = \lambda/L < 1$:

$$\mathbb{E}(-\mathcal{E}_k(z) - \lambda\|z\|^2)_+ \le \frac{\mathrm{e}\sigma_*^2}{\lambda m_k}\exp\left(-\frac{5m_k\lambda^2}{24L^2}\right) \le \frac{\mathrm{e}\sigma_*^2}{\lambda m_k}\exp\left(-\frac{m_k\lambda^2}{8L^2}\right).$$

For $z = 0$ the left side is zero. Sum this bound over the candidates and use $\max_i a_i \le \sum_i a_i$ for nonnegative a_i. No independence between different candidates is required. □

Define analysis-only coefficients

$$\begin{aligned}
\lambda &:= \frac{c}{8s}, \qquad w(m) := \frac{\mathrm{e}M_{\mathrm{trial}}\sigma_*^2}{\lambda m}\exp\left(-\frac{m\lambda^2}{8L^2}\right),\\
\overline{\Theta} &:= \begin{cases}(p-1)s, & L(p-1)s \le 1,\\ (p-1)s/c, & L(p-1)s > 1,\end{cases}\\
\Theta_0 &:= \begin{cases}0, & L(p-1)s \le 1,\\ 3Ls^2 + 4(p-1)L^2s^3/c + cs/4, & L(p-1)s > 1.\end{cases}
\end{aligned} \tag{67}$$

The quadratic compensation is applied separately to each candidate; there is no need to compare its size with that of the largest trial.

Lemma 7 (Weighted variance-reduced descent)**.** *For every positive m_k,*

$$\begin{aligned}
\mathbb{E}[\Phi(x_{k+1}) \mid \mathcal{F}_k] \le \Phi(x_k) &- \frac{c}{2}\mathbb{E}\left[\frac{h(\Delta_k)}{t_k}\,\middle|\,\mathcal{F}_k\right]\\
&+ \overline{\Theta}\,\mathbb{E}[\|e_k\|_*^2 \mid \mathcal{F}_k] + \Theta_0\,\mathbb{E}[\|U_k\|_*^2 \mid \mathcal{F}_k] + w(m_k).
\end{aligned} \tag{68}$$

Proof. Apply Lemma 3 at x_k with gradients v_k and $v_k - U_k$ to obtain

$$\|z_k(t_{k,i}) - \widetilde{z}_{k,i}\| \le t_{k,i}\|U_k\|_*.$$

At the accepted index, $\|\Delta_k - \widetilde{z}_{k,I_k}\| \le s\|U_k\|_*$ and

$$\|\widetilde{z}_{k,I_k}\|^2 \le 2\|\Delta_k\|^2 + 2s^2\|U_k\|_*^2 = 4(p-1)h(\Delta_k) + 2s^2\|U_k\|_*^2.$$

Pointwise, the negative model error at this candidate is at most $\lambda\|\widetilde{z}_{k,I_k}\|^2$ plus the maximum in (65). Consequently

$$\begin{aligned}\mathbb{E}[-\mathcal{E}_k(\widetilde{z}_{k,I_k}) \mid \mathcal{F}_k] \le \frac{c(p-1)}{2s}\mathbb{E}[h(\Delta_k) \mid \mathcal{F}_k] \\ + \frac{cs}{4}\mathbb{E}[\|U_k\|_*^2 \mid \mathcal{F}_k] + w(m_k).\end{aligned}$$

The random index has not been moved outside the conditional expectation.

The sample and population remainder-difference bounds (46) give

$$\begin{aligned}|\mathcal{E}_k(\Delta_k) - \mathcal{E}_k(\widetilde{z}_{k,I_k})| &\le 2L\|\widetilde{z}_{k,I_k}\|\|\Delta_k - \widetilde{z}_{k,I_k}\| + L\|\Delta_k - \widetilde{z}_{k,I_k}\|^2 \\ &\le 2L\|\Delta_k\|\|\Delta_k - \widetilde{z}_{k,I_k}\| + 3L\|\Delta_k - \widetilde{z}_{k,I_k}\|^2 \\ &\le \frac{c}{2s}h(\Delta_k) + \left(3Ls^2 + \frac{4L^2(p-1)s^3}{c}\right)\|U_k\|_*^2.\end{aligned}$$

Also

$$-\langle e_k, \Delta_k\rangle \le \frac{c}{2s}h(\Delta_k) + \frac{(p-1)s}{c}\|e_k\|_*^2.$$

Armijo implies

$$\Phi(x_{k+1}) - \Phi(x_k) \le -\frac{2c}{t_k}h(\Delta_k) - \langle e_k, \Delta_k\rangle - \mathcal{E}_k(\Delta_k).$$

The three positive displacement coefficients sum to $c(p+1)/(2s) \le 3c/(2t_k)$. This proves (68) when $L(p-1)s > 1$. No error paired with Δ_k has been centered. If $L(p-1)s \le 1$, apply (60) to v_k and add the nonnegative $w(m_k)$ to obtain the other branch. □

B.2 SPIDER RECURSION AND STATIONARITY

Lemma 8 (Finite-horizon SPIDER error)**.** *For the general-parameter version of Algorithm 2,*

$$\begin{aligned}\sum_{k=0}^{K-1}\mathbb{E}\|e_k\|_*^2 \le \frac{\sigma_*^2 K}{(p-1)B} \\ + \min\left\{\frac{8L^2(q-1)}{b}\sum_{k=0}^{K-1}\mathbb{E}h(\Delta_k), \frac{4\sigma_*^2 K(q-1)}{(p-1)b}\right\}.\end{aligned} \tag{69}$$

The fresh errors also satisfy

$$\begin{aligned}\sum_{k=0}^{K-1}\mathbb{E}\|U_k\|_*^2 \le \frac{\sigma_*^2\lceil K/q\rceil}{(p-1)B} \\ + \min\left\{\frac{8L^2}{b}\sum_{k=0}^{K-1}\mathbb{E}h(\Delta_k), \frac{4\sigma_*^2(K - \lceil K/q\rceil)}{(p-1)b}\right\}.\end{aligned} \tag{70}$$

Proof. At a refresh, dual averaging gives $\mathbb{E}[\|e_k\|_*^2 \mid \mathcal{F}_k] \le \sigma_*^2/((p-1)B)$. At an inner iteration, write

$$Z_{k,\ell} := G(x_k, \xi_{k,\ell}) - G(x_{k-1}, \xi_{k,\ell}) - \nabla f(x_k) + \nabla f(x_{k-1}), \quad 1 \le \ell \le b.$$

These vectors are conditionally independent and centered. Samplewise smoothness and the uniform noise second moment give

$$\mathbb{E}[\|Z_{k,\ell}\|_*^2 \mid \mathcal{F}_k] \le \min\{8(p-1)L^2h(\Delta_{k-1}),\, 4\sigma_*^2\}. \tag{71}$$

For the first bound, use $\|Z_{k,\ell}\|_* \le 2L\|\Delta_{k-1}\|$; for the second, write it as the difference of the centered gradient noises at the two points.

Reveal the b samples one at a time in the identity $e_k = e_{k-1} + b^{-1}\sum_\ell Z_{k,\ell}$. The conjugate $h^*(v) = (p-1)\|v\|_*^2/2$ is 1-smooth. At each revelation its linear term has conditional mean zero. Summing the b smoothness inequalities yields

$$\mathbb{E}[\|e_k\|_*^2 \mid \mathcal{F}_k] \le \|e_{k-1}\|_*^2 + \min\left\{\frac{8L^2}{b}h(\Delta_{k-1}), \frac{4\sigma_*^2}{(p-1)b}\right\}.$$

This uses centering of the increments, not centering of e_k. The fresh error at an inner iteration is $U_k = b^{-1}\sum_{\ell=1}^b Z_{k,\ell}$. Conditional dual averaging gives

$$\mathbb{E}[\|U_k\|_*^2 \mid \mathcal{F}_k] \le \min\left\{\frac{8L^2}{b}h(\Delta_{k-1}), \frac{4\sigma_*^2}{(p-1)b}\right\}.$$

At a refresh $U_k = e_k$. Summing these bounds over $\lceil K/q\rceil$ refreshes and $K - \lceil K/q\rceil$ inner iterations proves (70): each preceding displacement enters at most once, without a factor $q-1$.

Partition the run into refresh epochs, each of length at most q. Apply the displacement branch throughout each epoch and sum. Each displacement occurs at most $q-1$ times, and the refresh contribution is at most $\sigma_*^2 K/((p-1)B)$. Alternatively use the variance branch: each iteration has accumulated at most $q-1$ increments, giving $4\sigma_*^2 K(q-1)/((p-1)b)$. Taking the smaller of these two bounds proves (69), including a partial final epoch. □

The initial refresh has a finite error second moment. At later steps the recursion just proved propagates this property from previously integrable displacements. Conditional finite termination and stability give finite conditional displacement moments. Apply (68) after subtracting $\Phi_{\inf}$, with the weighted displacement on the nonnegative left-hand side. Induction gives finite unconditional objective, displacement, and error moments for every finite horizon. This justifies the expectations without an additional integrability assumption.

The number of refreshes and the compensation terms are

$$R_K := \left\lceil \frac{K}{q} \right\rceil,$$

$$E_K := \frac{\sigma_*^2 K}{p-1}\left(\frac{4[\overline{\Theta}(q-1)+\Theta_0]}{b} - \frac{c}{8sL^2}\right)_+, \tag{72}$$

$$A_K := \Phi(x_0) - \Phi_{\inf} + \frac{\sigma_*^2(\overline{\Theta}K + \Theta_0 R_K)}{(p-1)B} + R_K w(B) + (K-R_K)w(b) + E_K. \tag{73}$$

Here $(u)_+ = \max\{u, 0\}$. The finite-budget bound below retains this positive part. It is an analysis term; the algorithm does not evaluate it.

Proof of Theorem 2. For nonnegative u, v and positive θ,

$$\min\{u, v\} \le \theta u + (1-\theta)_+ v. \tag{74}$$

If $\theta \ge 1$, use $\min\{u,v\} \le u \le \theta u$. For $0 < \theta < 1$, the right-hand side minus the left-hand side is $(1-\theta)(v-u) \ge 0$ when $u \le v$, and $\theta(u-v) \ge 0$ when $u \ge v$.

Multiply (69) by $\overline{\Theta}$ and (70) by Θ_0, and add. The sum of two minima is at most the minimum of the sums of their respective branches. Use $K - R_K \le K$ in the variance branch and $h(\Delta_k) \le sh(\Delta_k)/t_k$ in the displacement branch. Apply (74) with

$$u = \frac{8sL^2[\overline{\Theta}(q-1)+\Theta_0]}{b}\sum_{k=0}^{K-1}\mathbb{E}\frac{h(\Delta_k)}{t_k},$$

$$v = \frac{4\sigma_*^2 K[\overline{\Theta}(q-1)+\Theta_0]}{(p-1)b}, \qquad \theta = \frac{cb}{32sL^2[\overline{\Theta}(q-1)+\Theta_0]}.$$

Since $q \geq 2$ and $\overline{\Theta} > 0$, $\theta > 0$. The resulting bound is

$$\overline{\Theta} \sum_{k=0}^{K-1} \mathbb{E}\|e_k\|_*^2 + \Theta_0 \sum_{k=0}^{K-1} \mathbb{E}\|U_k\|_*^2 \leq \frac{\sigma_*^2(\overline{\Theta}K + \Theta_0 R_K)}{(p-1)B} + \frac{c}{4} \sum_{k=0}^{K-1} \mathbb{E}\frac{h(\Delta_k)}{t_k} + E_K. \tag{75}$$

Sum (68). Since $\sum_k w(m_k) = R_K w(B) + (K - R_K)w(b)$, substituting (75) yields

$$\mathbb{E}[\Phi(x_K) - \Phi_{\inf}] + \frac{c}{4} \sum_{k=0}^{K-1} \mathbb{E}\frac{h(\Delta_k)}{t_k} \leq A_K.$$

This proves the objective and displacement bounds. In (75), the refresh term and E_K sum to at most A_K, and the displacement term is at most another A_K. Dropping the nonnegative fresh-error term proves (18). Sum the weighted bound (55) and use independence and uniformity of Y to obtain (19). □

For the general-parameter version, the exact independent-sample and gradient counts are

$$\begin{aligned} \mathcal{N}_{\text{samp}} &= R_K B + (K - R_K)b, \\ \mathcal{N}_G &= R_K B + 2(K - R_K)b. \end{aligned} \tag{76}$$

A refresh uses B samples and B gradients; an inner iteration uses b samples and paired gradients at two points. These counts include a partial final epoch and do not depend on the stepsize reset set.

B.3 The prescribed schedule and oracle complexity

Define the finite horizon-independent bound

$$A_\infty := \Phi(x_0) - \Phi_{\inf} + \frac{\sigma_*^2}{p-1}\left[\overline{\Theta} + \Theta_0 + 4\overline{\Theta}\exp\left(\frac{64s\overline{\Theta}L^2}{c}\right) + \frac{64sL^2\Theta_0^2}{c}\right] + \frac{8M_{\text{trial}}\sigma_*^2 L^2}{\lambda^3}. \tag{77}$$

Proof of Theorem 3. For $K \geq 2$, $1 < \log(\mathrm{e} + K) < K$. Thus $2 \leq q \leq K$ and

$$q \geq \sqrt{K/\log(\mathrm{e}+K)}, \qquad b \geq \sqrt{K\log(\mathrm{e}+K)} \geq \sqrt{K}, \qquad R_K \leq \sqrt{K\log(\mathrm{e}+K)} + 1.$$

Moreover $b \leq \sqrt{K\log(\mathrm{e}+K)} + \log(\mathrm{e}+K) + 1 \leq 2\sqrt{K\log(\mathrm{e}+K)} + 1$. Substitution in (76) proves (22).

Since $q/b \leq 1/\log(\mathrm{e}+K)$,

$$\frac{4[\overline{\Theta}(q-1) + \Theta_0]}{b} \leq \frac{4\overline{\Theta}}{\log(\mathrm{e}+K)} + \frac{4\Theta_0}{\sqrt{K}}.$$

Using $(u+v)_+ \leq u_+ + v_+$ and splitting the negative term equally gives

$$E_K \leq \frac{\sigma_*^2}{p-1}\left[K\left(\frac{4\overline{\Theta}}{\log(\mathrm{e}+K)} - \frac{c}{16sL^2}\right)_+ + \left(4\Theta_0\sqrt{K} - \frac{cK}{16sL^2}\right)_+\right]. \tag{78}$$

If the first positive part is nonzero, then $\log(\mathrm{e}+K) < 64s\overline{\Theta}L^2/c$ and hence $K < \exp(64s\overline{\Theta}L^2/c)$. Since $\log(\mathrm{e}+K) > 1$, its contribution inside the brackets is at most $4\overline{\Theta}\exp(64s\overline{\Theta}L^2/c)$; the same bound holds when the positive part is zero. For the second term,

$$\sup_{u \geq 0}\left(4\Theta_0 u - \frac{cu^2}{16sL^2}\right)_+ = \frac{64sL^2\Theta_0^2}{c},$$

as follows by completing the square. Substitution into (78) gives, for every $K \geq 2$,

$$E_K \leq \frac{\sigma_*^2}{p-1}\left[4\overline{\Theta}\exp\left(\frac{64s\overline{\Theta}L^2}{c}\right)+\frac{64sL^2\Theta_0^2}{c}\right]. \tag{79}$$

The function w is decreasing. Since $B, b \geq \sqrt{K}$,

$$R_K w(B)+(K-R_K)w(b) \leq \frac{\mathrm{e}M_{\mathrm{trial}}\sigma_*^2}{\lambda}\sqrt{K}\exp\left(-\frac{\lambda^2\sqrt{K}}{8L^2}\right). \tag{80}$$

Using $\sup_{u\geq 0} ue^{-\lambda^2 u/(8L^2)} = 8L^2/(\mathrm{e}\lambda^2)$ in (80) yields

$$R_K w(B)+(K-R_K)w(b) \leq \frac{8M_{\mathrm{trial}}\sigma_*^2 L^2}{\lambda^3} \qquad (K \geq 2).$$

Finally, $B = K$ and $R_K \leq K$ give

$$\frac{\sigma_*^2(\overline{\Theta}K+\Theta_0 R_K)}{(p-1)B} \leq \frac{\sigma_*^2(\overline{\Theta}+\Theta_0)}{p-1}.$$

Substitute these bounds and (79) into (73) to obtain $A_K \leq A_\infty$ for every $K \geq 2$. Equation 19 now gives the uniform C/K bound. For $0 < \varepsilon < 1$, the sufficient theoretical horizon

$$K = \max\{2, \lceil C/\varepsilon^2 \rceil\} \tag{81}$$

has $C/K \leq \varepsilon^2$ and $K \leq 2 + C/\varepsilon^2$. Insertion into (22) gives

$$\mathcal{N}_G, \mathcal{N}_{\mathrm{samp}} = \mathcal{O}\left((1+C/\varepsilon^2)^{3/2}\sqrt{\log(\mathrm{e}+C/\varepsilon^2)}\right), \tag{82}$$

including $C = 0$. For a fixed problem this is the claimed $\mathcal{O}(\varepsilon^{-3}\sqrt{\log(\mathrm{e}+1/\varepsilon)})$ order. □

Finally,

$$\mathcal{N}_F \leq (M_{\mathrm{trial}}+1)\mathcal{N}_{\mathrm{samp}}, \qquad \mathcal{N}_{\mathrm{prox}} \leq M_{\mathrm{trial}}K.$$

Repeated Armijo trials reuse the same batch and do not multiply the sample or gradient counts. All these counts omit the arithmetic cost of an individual proximal solve.

C Details for the Near-ℓ_1/ℓ_∞ Specialization

For every $p \in (1, 2]$,

$$\|u\|_1 \leq d^{1/p^*}\|u\|_p, \qquad \|v\|_{p^*} \leq d^{1/p^*}\|v\|_\infty. \tag{83}$$

Thus (3)–(4) hold with

$$L := d^{2/p^*}L_{1,\infty}, \qquad \sigma_* := d^{1/p^*}\sigma_\infty. \tag{84}$$

For every $p \in (1, 2]$, the norm comparisons in (83) and the endpoint bounds (25)–(26) give

$$\begin{aligned}\|G(x,\xi)-G(y,\xi)\|_{p^*} &\leq d^{1/p^*}\|G(x,\xi)-G(y,\xi)\|_\infty \\ &\leq d^{2/p^*}L_{1,\infty}\|x-y\|_p = L\|x-y\|_p, &(85)\\ \mathbb{E}\|G(x,\xi)-\nabla f(x)\|_{p^*}^2 &\leq d^{2/p^*}\sigma_\infty^2 = \sigma_*^2, &(86)\\ \|\nabla f(x)-\nabla f(y)\|_{p^*} &\leq d^{2/p^*}L_{1,\infty}\|x-y\|_p = L\|x-y\|_p. &(87)\end{aligned}$$

The last inequality follows by Jensen's inequality and (85). These are precisely the ℓ_p/ℓ_{p^*} smoothness and second-moment estimates used in the minibatch, fixed-budget variance-reduced, and inexact analyses; they are consequences of the endpoint oracle assumption rather than additional assumptions.

For (27), $p \in (1, 2]$, and for every $z, v \in \mathbb{R}^d$,

$$\|z\|_p \leq \|z\|_1 \leq d^{1/p^*}\|z\|_p \leq \sqrt{\mathrm{e}}\,\|z\|_p, \tag{88}$$

$$\|v\|_\infty \leq \|v\|_{p^*} \leq d^{1/p^*}\|v\|_\infty \leq \sqrt{\mathrm{e}}\,\|v\|_\infty, \tag{89}$$

because $d^{1/p^*} = \exp(\log d/(1+2\log d)) < \sqrt{\mathrm{e}}$. Consequently, for every nonempty $S \subseteq \mathbb{R}^d$,

$$\mathrm{dist}_\infty(0,S) \le \mathrm{dist}_{p^*}(0,S) \le \sqrt{\mathrm{e}}\ \mathrm{dist}_\infty(0,S). \tag{90}$$

The effective constants satisfy

$$L \le \mathrm{e}L_{1,\infty}, \qquad \sigma_*^2 \le \mathrm{e}\sigma_\infty^2. \tag{91}$$

Proof of Corollary 3. The preceding conversion establishes the smoothness and variance clauses of Assumption 1 with (84); its remaining clauses hold by hypothesis. For every output, (90) with $S = \partial\Phi(x_{\mathrm{out}})$ gives $\mathrm{dist}_\infty(0,S) \le \mathrm{dist}_{p^*}(0,S)$. Squaring and taking expectations, then applying Corollaries 1 and 2 and Theorems 2 and 3, proves the exact rates and oracle orders. The same comparison applies to Theorem 4. It introduces no additional oracle calls. □

For the choice in (27),

$$\frac{1}{p-1} = 2\log d, \qquad \frac{1}{(p-1)^2} = 4(\log d)^2. \tag{92}$$

In particular, (52) gives

$$\mathbb{E}[\|e_k\|_{p^*}^2 \mid \mathcal{F}_k] \le \frac{2\mathrm{e}\sigma_\infty^2 \log d}{N_k}. \tag{93}$$

The averaging factor has not disappeared under normalization. The sharper descent analysis instead avoids multiplying it by another inverse power of $p-1$ in the final noise coefficient.

To quantify dimension dependence, fix s, β, c, c_b, and consider a family with uniformly bounded $L_{1,\infty}$, σ_∞^2, and $\Phi(x_0) - \Phi_{\inf}$. These are conditions for comparing dimensions, not additional assumptions for the fixed-problem theorems. By (91), L is uniformly bounded. Since $0 < p-1 \le 1$, the lower stepsize bound is uniformly positive, and ρ and M_{trial} are uniformly bounded. In (50), either $\widehat{\Gamma} = s$, or $L(p-1)s > 1$ implies $(p-1)^{-1} < Ls$, so the other branch is also uniformly bounded. Therefore

$$C_{\mathrm{res}} = \mathcal{O}((p-1)^{-1}), \qquad \widehat{\Gamma} = \mathcal{O}(1), \qquad \frac{\overline{\Theta}}{p-1} = \mathcal{O}(1), \qquad \frac{\Theta_0}{p-1} = \mathcal{O}(1). \tag{94}$$

Indeed, $(p-1)s \le \overline{\Theta} \le (p-1)s/c$. If $L(p-1)s \le 1$, then $\Theta_0 = 0$. Otherwise, $(p-1)^{-1} < Ls$ is uniformly bounded, so the expression for Θ_0 in (67) proves the last bound.

The explicit minibatch constants now give

$$C_0 = \mathcal{O}\left(\frac{\Phi(x_0) - \Phi_{\inf}}{p-1}\right), \qquad C_1 = \mathcal{O}\left(\frac{\sigma_\infty^2}{p-1}\right). \tag{95}$$

For example, $(p-1)C_{\mathrm{res}}^2 = \mathcal{O}((p-1)^{-1})$ in both terms of (9). For $N_k = \lceil c_b K \rceil$, the finite sum of reciprocal batch sizes is at most $1/c_b$, so

$$\mathbb{E}\,\mathrm{dist}_\infty^2(0, \partial\Phi(x_{\mathrm{out}})) \le \frac{C_0 + C_1/c_b}{K} = \mathcal{O}\left(\frac{(\Phi(x_0) - \Phi_{\inf} + \sigma_\infty^2)\log d}{K}\right). \tag{96}$$

There is no finite-prefix objective or critical index in this calculation. For $N_k = k+1$, the corresponding bound is $\mathcal{O}((\Phi(x_0) - \Phi_{\inf} + \sigma_\infty^2(1+\log K))\log(d)/K)$.

For VR, (69) has displacement coefficient $8L^2(q-1)/b$, with no inverse power of $p-1$. In the full bound (77), $\lambda = c/(8s)$ is independent of d, and

$$\frac{\overline{\Theta}}{p-1}\exp(64s\overline{\Theta}L^2/c) = \mathcal{O}(1), \qquad \frac{L^2\Theta_0^2}{p-1} = \mathcal{O}(1).$$

The refresh terms and finite-candidate compensation are uniformly bounded as well. Thus

$$A_\infty = \mathcal{O}(\Phi(x_0) - \Phi_{\inf} + \sigma_\infty^2).$$

Both coefficients multiplying A_∞ in (21) have order $\mathcal{O}((p-1)^{-1})$. Consequently,

$$\mathbb{E}\,\mathrm{dist}_\infty^2(0, \partial\Phi(x_{\mathrm{out}})) \le \frac{C}{K} = \mathcal{O}\left(\frac{(\Phi(x_0) - \Phi_{\inf} + \sigma_\infty^2)\log d}{K}\right), \qquad K \ge 2. \tag{97}$$

The constants are uniform in d under the stated family bounds, but can grow rapidly with the common smoothness bound.

For the inexact methods, fix $\chi \in (0, 1-c)$ and η_0 independently of d. The same argument bounds $\underline{t}_\chi$ away from zero and $M_{\mathrm{trial},\chi}, \rho_\chi$ above. The constants $\widehat{\Gamma}_\chi, D_{\mathrm{mb},\chi}, \widehat{\Theta}_\chi, D_\chi$ are uniformly bounded, and $\overline{\Theta}_\chi = (p-1)\widehat{\Theta}_\chi$ with $\widehat{\Theta}_\chi \geq s/(1-\chi)$. The bound (147) is therefore $\mathcal{O}(\Phi(x_0) - \Phi_{\inf} + \sigma_\infty^2 + \eta_0^2)$ uniformly in d: λ_χ is bounded away from zero, and $\overline{\Theta}_\chi/(p-1)$ and its exponential factor are uniformly bounded. In (140) and (146), substitute these bounds and $\sum_k \eta_k^2 = \mathcal{O}(\eta_0^2)$ to obtain

$$C_{\mathrm{mb}}^{\mathrm{in}},\ C_{\mathrm{vr}}^{\mathrm{in}} = \mathcal{O}\big((\Phi(x_0) - \Phi_{\inf} + \sigma_\infty^2 + \eta_0^2)\log d\big). \tag{98}$$

No additional batch scaling or stricter tolerance schedule is used.

For $0 < \varepsilon < 1$, the displayed uniform bounds permit sufficient budgets $K = \mathcal{O}(\log(d)/\varepsilon^2)$ for both horizon-proportional minibatches and fixed-budget VR, with the constants depending on the stated family bounds. Substitution into (12) and (22) gives, respectively for horizon-proportional minibatches and fixed-budget VR,

$$\begin{aligned} \mathcal{N}_{\mathrm{samp}} = \mathcal{N}_G &= \mathcal{O}\left((\log d)^2 \varepsilon^{-4}\right), \\ \mathcal{N}_{\mathrm{samp}},\ \mathcal{N}_G &= \mathcal{O}\left((\log d)^{3/2}\varepsilon^{-3}\sqrt{\log\left(\mathrm{e} + \frac{\log d}{\varepsilon^2}\right)}\right). \end{aligned} \tag{99}$$

For linear minibatches, a sufficient budget is $K = \mathcal{O}(\log(d)\varepsilon^{-2}\log(\mathrm{e} + \log(d)/\varepsilon^2))$, and (15) gives

$$\mathcal{N}_{\mathrm{samp}} = \mathcal{N}_G = \mathcal{O}\left((\log d)^2 \varepsilon^{-4}\left[\log\left(\mathrm{e} + \frac{\log d}{\varepsilon^2}\right)\right]^2\right).$$

These are sufficient stochastic-oracle upper bounds, not optimality claims or observable stopping rules. Deterministic work on a vector still depends on d. For $p = 2$ under the endpoint assumption alone, the conversion $L = dL_{1,\infty}$ and $\sigma_*^2 = d\sigma_\infty^2$ can instead introduce polynomial dimension dependence. The near-ℓ_1 conclusion uses (27), not normalization alone.

D Proximal Subproblem Computation

D.1 Unconstrained formula

When $r \equiv 0$, the trial subproblem is

$$\min_{z \in \mathbb{R}^d}\left\{\langle g, z\rangle + \frac{1}{2(p-1)t}\|z\|_p^2\right\}. \tag{100}$$

The squared-norm conjugacy formula (Beck, 2017) gives $h^* = (p-1)\|\cdot\|_{p^*}^2/2$ for $h = \|\cdot\|_p^2/[2(p-1)]$. Its gradient satisfies

$$[\nabla h^*(v)]_i = (p-1)\|v\|_{p^*}^{2-p^*}|v_i|^{p^*-1}\operatorname{sgn}(v_i), \tag{101}$$

with $\nabla h^*(0) = 0$. The optimality condition for (100) is $-tg = \nabla h(z)$, or equivalently $z = \nabla h^*(-tg)$. The one-homogeneity of the gradient of a squared norm therefore gives (30) and (31).

For $g \neq 0$, an equivalent evaluation avoids multiplying separately large and small powers:

$$z_i = -(p-1)t\|g\|_{p^*}\operatorname{sgn}(g_i)\left(\frac{|g_i|}{\|g\|_{p^*}}\right)^{p^*-1}.$$

The ratios lie in $[0, 1]$. Computing the norm after scaling by $\max_i |g_i|$ also avoids forming large powers of the original coordinates. This changes the numerical evaluation, not the proximal solution.

D.2 Box-constrained scalar reduction

For a displacement box $\mathcal{B} = [\ell, u]$ with $\ell \leq 0 \leq u$ componentwise, consider

$$\min_{z \in \mathcal{B}}\left\{\langle g, z\rangle + \frac{1}{2(p-1)t}\|z\|_p^2\right\}. \tag{102}$$

The zero vector is optimal if and only if

$$-g \in N_{\mathcal{B}}(0), \qquad \text{equivalently} \qquad \Pi_{\mathcal{B}}(-g) = 0. \tag{103}$$

Otherwise, for $\theta > 0$, define

$$y_i(\theta) := -\operatorname{sgn}(g_i)\left(\frac{|g_i|}{\theta}\right)^{1/(p-1)}, \qquad z_i(\theta) := \Pi_{[\ell_i,u_i]}(y_i(\theta)), \tag{104}$$

$$G(\theta) := \theta - \frac{1}{(p-1)t}\|z(\theta)\|_p^{2-p}. \tag{105}$$

For $p = 2$, the second term in (105) is understood as $1/t$ on the nonzero branch.

Proposition 3 (Scalar reduction for a box)**.** *If (103) fails, then G is continuous and strictly increasing on $(0, \infty)$, with*

$$\lim_{\theta\downarrow 0} G(\theta) < 0, \qquad \lim_{\theta\uparrow\infty} G(\theta) = +\infty. \tag{106}$$

Hence G has a unique positive root $\theta^\star$, and the unique solution of (102) is $z^\star = z(\theta^\star)$.

The scalar error also controls the proximal KKT residual:

$$\begin{aligned}\operatorname{dist}_{p^*}\left(0, g + \frac{1}{t}\nabla h(z(\theta)) + N_{\mathcal{B}}(z(\theta))\right) \\ \le |G(\theta)|\,\|z(\theta)\|_p^{p-1}.\end{aligned} \tag{107}$$

The gradient of h is

$$[\nabla h(z)]_i = \frac{1}{p-1}\|z\|_p^{2-p}|z_i|^{p-1}\operatorname{sgn}(z_i), \tag{108}$$

with $\nabla h(0) = 0$. Hence $z^\star = 0$ solves (102) exactly when $-g \in N_{\mathcal{B}}(0)$, which is equivalent to the projection test (103).

Proof of Proposition 3. For fixed $\theta > 0$, the coordinate $z_i(\theta)$ in (104) is the unique minimizer over $[\ell_i, u_i]$ of

$$g_i v + \frac{\theta}{p}|v|^p.$$

It satisfies

$$0 \in g_i + \theta|z_i(\theta)|^{p-1}\operatorname{sgn}(z_i(\theta)) + N_{[\ell_i,u_i]}(z_i(\theta)). \tag{109}$$

Because each interval contains zero, $|z_i(\theta)|$ is continuous and nonincreasing in θ. Therefore $[(p-1)t]^{-1}\|z(\theta)\|_p^{2-p}$ is continuous and nonincreasing. For $0 < \theta_1 < \theta_2$,

$$\begin{aligned}G(\theta_2) - G(\theta_1) &= (\theta_2 - \theta_1) \\ &\quad - \frac{1}{(p-1)t}\left(\|z(\theta_2)\|_p^{2-p} - \|z(\theta_1)\|_p^{2-p}\right) \\ &\ge \theta_2 - \theta_1 > 0,\end{aligned}$$

so G is continuous and strictly increasing.

Failure of (103) means that some coordinate has a feasible nonzero descent direction: either $g_i > 0$ and $\ell_i < 0$, or $g_i < 0$ and $u_i > 0$. In the first case $y_i(\theta) \to -\infty$ as $\theta \downarrow 0$; in the second it tends to $+\infty$. Projection onto the corresponding interval therefore has a nonzero limiting magnitude, possibly infinite when that side is unbounded. For $1 < p < 2$,

$$\liminf_{\theta\downarrow 0} \frac{1}{(p-1)t}\|z(\theta)\|_p^{2-p} > 0,$$

while for $p = 2$ that term equals $1/t$. This proves the first limit in (106). As $\theta \uparrow \infty$, $y(\theta) \to 0$, and projection onto an interval containing zero cannot increase magnitude. Hence $z(\theta) \to 0$ and $[(p-1)t]^{-1}\|z(\theta)\|_p^{2-p} \to 0$ for $1 < p < 2$; for $p = 2$ it remains $1/t$. Thus $G(\theta) \to +\infty$ in either case.

Let $\theta^\star$ be the unique root. Then $\theta^\star = [(p-1)t]^{-1}\|z(\theta^\star)\|_p^{2-p}$. Substituting this identity into (109) yields the KKT system of (102). Strong convexity of the objective makes the solution unique, so $z^\star = z(\theta^\star)$. □

D.3 Bracketed bisection and safeguarded Newton steps

Start the scalar search from $\theta_0 = 1/[(p-1)t]$. If $G(\theta_0) = 0$, return $z(\theta_0)$. If $G(\theta_0) < 0$, use θ_0 as the lower endpoint and repeatedly double the upper endpoint until its function value is positive. If $G(\theta_0) > 0$, use it as the upper endpoint and repeatedly halve the lower endpoint until its function value is negative. The limits in (106) guarantee finite bracketing. Bisection then maintains

$$G(\theta_{\mathrm{lo}}) < 0 < G(\theta_{\mathrm{hi}}). \tag{110}$$

After an initial bracket $[\theta_{\mathrm{lo}}^{(0)}, \theta_{\mathrm{hi}}^{(0)}]$ has been obtained, the number of bisection iterations needed to reduce its width to at most $\delta_\theta > 0$ is at most

$$\max\left\{0, \left\lceil \log_2 \frac{\theta_{\mathrm{hi}}^{(0)} - \theta_{\mathrm{lo}}^{(0)}}{\delta_\theta} \right\rceil\right\}. \tag{111}$$

Each evaluation of G costs $\mathcal{O}(d)$ operations. If N_{bracket} doubling or halving steps initialize the bracket, the arithmetic cost is

$$\mathcal{O}\left(d\left[1 + N_{\mathrm{bracket}} + \log \max\left\{1, \frac{\theta_{\mathrm{hi}}^{(0)} - \theta_{\mathrm{lo}}^{(0)}}{\delta_\theta}\right\}\right]\right). \tag{112}$$

This deterministic arithmetic is separate from $\mathcal{N}_G$, $\mathcal{N}_F$, and the independent-sample count. The bracket width is not itself the proximal accuracy certificate: the inexact algorithm stops using the KKT criterion (119).

To verify (107), define $q_i(z) := |z_i|^{p-1}\operatorname{sgn}(z_i)$. Equation 109 gives

$$0 \in g + \theta q(z(\theta)) + N_{\mathcal{B}}(z(\theta)),$$

whereas (105) gives

$$\frac{1}{(p-1)t}\|z(\theta)\|_p^{2-p} = \theta - G(\theta).$$

Using the same normal-cone element in the exact proximal KKT residual leaves $-G(\theta)q(z(\theta))$. Its norm is unchanged by this sign. Finally,

$$\|q(z)\|_{p^*}^{p^*} = \sum_{i=1}^{d} |z_i|^{(p-1)p^*} = \sum_{i=1}^{d} |z_i|^p,$$

so $\|q(z)\|_{p^*} = \|z\|_p^{p-1}$ and (107) follows.

Away from clipping breakpoints, safeguarded Newton iterations can accelerate the bisection. Let

$$\mathcal{I}(\theta) := \{i : \ell_i < y_i(\theta) < u_i\}, \qquad S(\theta) := \sum_{i=1}^{d} |z_i(\theta)|^p. \tag{113}$$

On an interval containing no clipping breakpoint,

$$z_i'(\theta) = -\frac{z_i(\theta)}{(p-1)\theta}, \quad i \in \mathcal{I}(\theta), \qquad S'(\theta) = -\frac{p}{(p-1)\theta}\sum_{i\in\mathcal{I}(\theta)} |z_i(\theta)|^p.$$

For $1 < p < 2$ and $S(\theta) > 0$, differentiation of $G(\theta) = \theta - [(p-1)t]^{-1}S(\theta)^{(2-p)/p}$ yields

$$G'(\theta) = 1 + \frac{2-p}{t(p-1)^2\theta} S(\theta)^{(2-p)/p-1} \sum_{i\in\mathcal{I}(\theta)} |z_i(\theta)|^p. \tag{114}$$

Use $\theta - G(\theta)/G'(\theta)$ only when the derivative is defined and the proposal lies in the central half of the current bracket; otherwise use the midpoint. If G is zero, return that root; otherwise its sign updates the corresponding endpoint. Each nonterminal update then reduces the bracket width by a factor at most $3/4$, ensuring convergence even when Newton proposals are rejected or cross clipping breakpoints. The count in (111) refers to pure bisection; for this hybrid rule the same width bound holds with $\log_{4/3}$ in place of $\log_2$. For $p = 2$, $G'(\theta) = 1$, $\theta^\star = 1/t$, and

$$z^\star = \Pi_{\mathcal{B}}(-tg). \tag{115}$$

E Proofs for Inexact Proximal Subproblems

Throughout, the tolerances are

$$0 < \chi < 1 - c, \qquad \eta_k := \frac{\eta_0}{k+1}, \qquad \eta_0 > 0. \tag{116}$$

Theorem 4 (Inexact proximal variants)**.** *Under Assumptions 1 and 2, replace each exact proximal solve in Algorithm 1 or Algorithm 2 by a measurable candidate satisfying (32). Let x_0 be deterministic, take any deterministic positive minibatch sequence, and use (20) for fixed-budget VR. There are finite fixed-problem constants $C^{\mathrm{in}}_{\mathrm{mb}}$ and $C^{\mathrm{in}}_{\mathrm{vr}}$ such that*

$$\mathbb{E}\,\mathrm{dist}^2_{p^*}(0, \partial\Phi(x_{\mathrm{out}})) \le \frac{C^{\mathrm{in}}_{\mathrm{mb}}}{K}\left(1 + \sum_{k=0}^{K-1} N_k^{-1}\right), \tag{117}$$

$$\mathbb{E}\,\mathrm{dist}^2_{p^*}(0, \partial\Phi(x_{\mathrm{out}})) \le \frac{C^{\mathrm{in}}_{\mathrm{vr}}}{K}, \qquad K \ge 2, \tag{118}$$

with the minibatch bound holding for every $K \ge 1$. Explicit finite-budget bounds for general deterministic batch parameters are given in Appendix E. For minibatches, (11) and (14) give stochastic-oracle complexities $\mathcal{O}(\varepsilon^{-4})$ and $\mathcal{O}(\varepsilon^{-4}\log^2(\mathrm{e} + 1/\varepsilon))$, respectively. The fixed-budget VR order is $\mathcal{O}(\varepsilon^{-3}\sqrt{\log(\mathrm{e} + 1/\varepsilon)})$.

In the box case, (107) shows that the scalar solver may stop once

$$|G(\theta)|\|z(\theta)\|_p^{p-1} \le \min\left\{\frac{\chi}{(p-1)t}\|z(\theta)\|_p, \frac{1}{t}\eta_k\right\}. \tag{119}$$

The Armijo test is unchanged, with $\widetilde{z}_k(t)$ replacing the exact candidate. Thus an accepted minibatch displacement satisfies

$$\widehat{f}_k(x_k + \Delta_k) + r(x_k + \Delta_k) \le \widehat{f}_k(x_k) + r(x_k) - \frac{2c}{t_k}h(\Delta_k), \tag{120}$$

and the variance-reduced test replaces $\widehat{f}_k$ by H_k. After the exact zero test, the box root has $\|z(\theta^\star)\|_p > 0$. Continuity then makes the right-hand side of (119) positive near the root while its left-hand side tends to zero. A contracting bracket therefore reaches this criterion in finitely many steps in exact arithmetic. This statement does not provide a universal finite-cost solver for a general convex r.

Set

$$\underline{t}_\chi := \min\left\{s, \frac{2\beta(1-c-\chi)}{L(p-1)}\right\}, \qquad \rho_\chi := \frac{s}{\underline{t}_\chi}, \tag{121}$$

$$M_{\mathrm{trial},\chi} := 1 + \left\lceil \log_{1/\beta}\left(\max\left\{1, \frac{sL(p-1)}{2(1-c-\chi)}\right\}\right)\right\rceil. \tag{122}$$

Proposition 4 (Inexact termination and candidate accuracy)**.** *Provided that every trial candidate satisfies (32), each inexact backtracking loop terminates almost surely, and*

$$\underline{t}_\chi \le t_k \le \tau_k \le s, \qquad I_k < M_{\mathrm{trial},\chi}. \tag{123}$$

For a potential trial $t_{k,i} = \beta^i\tau_k$, let $z_{k,i} := z(g_k, t_{k,i}; x_k)$ and $\widetilde{z}_{k,i} := \widetilde{z}_k(t_{k,i})$. Then

$$\|\widetilde{z}_{k,i} - z_{k,i}\| \le \eta_k. \tag{124}$$

Define

$$C_{\mathrm{res},\chi} := L + \frac{1}{(p-1)\underline{t}_\chi}. \tag{125}$$

At every accepted inexact iteration,

$$\mathrm{dist}^2_{p^*}(0, \partial\Phi(x_{k+1})) \le 6(p-1)C^2_{\mathrm{res},\chi}h(\Delta_k) + 3\|e_k\|_*^2 + \frac{3}{\underline{t}_\chi^2}\eta_k^2, \tag{126}$$

with $e_k = v_k - \nabla f(x_k)$ for the variance-reduced method.

Proof of Proposition 4. Choose a proximal KKT residual whose dual norm is bounded by (32). Pairing the inexact KKT relation with $\widetilde{z}_k(t)$ and using the subgradient inequality for r gives

$$\langle g_k, \widetilde{z}_k(t)\rangle + r(x_k + \widetilde{z}_k(t)) - r(x_k) + \frac{2}{t}h(\widetilde{z}_k(t)) \le \frac{\chi}{(p-1)t}\|\widetilde{z}_k(t)\|^2 = \frac{2\chi}{t}h(\widetilde{z}_k(t)).$$

The current sampled model is L-smooth because it is an average of samplewise L-smooth functions plus, in an inner variance-reduced iteration, an affine term. Its objective change is therefore at most

$$-\frac{2(1-\chi)}{t}h(\widetilde{z}_k(t)) + L(p-1)h(\widetilde{z}_k(t)).$$

The unchanged Armijo condition holds whenever $t \le 2(1-c-\chi)/(L(p-1))$. The geometric trial sequence then gives (123) and the definitions in (121)–(122).

For the candidate accuracy, subtract the exact and inexact proximal KKT relations and pair with $\widetilde{z}_{k,i} - z_{k,i}$. Monotonicity of ∂r, strong monotonicity of ∇h, and the absolute part of (32) yield

$$\frac{1}{t_{k,i}}\|\widetilde{z}_{k,i} - z_{k,i}\|^2 \le \frac{1}{t_{k,i}}\eta_k\|\widetilde{z}_{k,i} - z_{k,i}\|.$$

Division proves (124); the claim is immediate when the candidate difference is zero. □

For every potential inexact trial, retain the exact population candidate

$$\bar{z}_{k,i} := z(\nabla f(x_k), t_{k,i}; x_k), \qquad i = 0, \dots, M_{\mathrm{trial},\chi} - 1. \tag{127}$$

For the minibatch method, e_k and $\mathcal{D}_k$ are defined in (43) and (47). For the variance-reduced method, use $e_k = v_k - \nabla f(x_k)$ and $\mathcal{E}_k$ from (64). In both cases the population candidates are $\mathcal{F}_k$-measurable.

E.1 Inexact minibatch descent

Set

$$\Gamma_{0,\chi} := \frac{sM_{\mathrm{trial},\chi}}{2c} + 2s\sqrt{M_{\mathrm{trial},\chi}} + 2Ls^2 + sM_{\mathrm{trial},\chi},$$
$$\widehat{\Gamma}_\chi := \frac{s + Ls\Gamma_{0,\chi}}{1-\chi}, \qquad D_{\mathrm{mb},\chi} := 2L + \frac{1}{s}. \tag{128}$$

For every positive N_k, we claim

$$\mathbb{E}[\Phi(x_{k+1}) \mid \mathcal{F}_k] \le \Phi(x_k) - \frac{c}{2s}\mathbb{E}[h(\Delta_k) \mid \mathcal{F}_k] + \widehat{\Gamma}_\chi\frac{\sigma_*^2}{N_k} + D_{\mathrm{mb},\chi}\eta_k^2. \tag{129}$$

Indeed, sampled-to-population stability and (124) give

$$\|\Delta_k - \bar{z}_{k,I_k}\| \le s\|e_k\|_* + \eta_k. \tag{130}$$

Use the two finite maxima from the proof of Lemma 5, now over $M_{\mathrm{trial},\chi}$ candidates. The same argument as (57) gives

$$|\mathcal{D}_k(\Delta_k)| \le A_k^{\mathrm{err}}\|\Delta_k\| + (A_k^{\mathrm{err}} + B_k^{\mathrm{err}})(s\|e_k\|_* + \eta_k) + L(s\|e_k\|_* + \eta_k)^2.$$

The additional linear term satisfies

$$\eta_k(A_k^{\mathrm{err}} + B_k^{\mathrm{err}}) \le \frac{s}{2}\big((A_k^{\mathrm{err}})^2 + (B_k^{\mathrm{err}})^2\big) + \frac{\eta_k^2}{s}.$$

Together with $L(s\|e_k\|_* + \eta_k)^2 \le 2Ls^2\|e_k\|_*^2 + 2L\eta_k^2$, the conditional moment bounds yield

$$\mathbb{E}[\Phi(x_{k+1}) \mid \mathcal{F}_k] \le \Phi(x_k) - \frac{c}{s}\mathbb{E}[h(\Delta_k) \mid \mathcal{F}_k] + \frac{\Gamma_{0,\chi}\sigma_*^2}{(p-1)N_k} + D_{\mathrm{mb},\chi}\eta_k^2. \tag{131}$$

Here the coefficient of the new maximum-moment term is at most $sM_{\mathrm{trial},\chi}$ after factoring out $\sigma_*^2/((p-1)N_k)$.

The relative KKT criterion and population smoothness also give, pathwise,

$$\Phi(x_{k+1}) - \Phi(x_k) \le -\left(\frac{2(1-\chi)}{t_k} - L(p-1)\right) h(\Delta_k) - \langle e_k, \Delta_k\rangle. \tag{132}$$

If $L(p-1)s \le 1-\chi$, apply

$$-\langle e_k, \Delta_k\rangle \le \frac{1-\chi}{2s} h(\Delta_k) + \frac{(p-1)s}{1-\chi}\|e_k\|_*^2.$$

Since $1-\chi > c$, the resulting descent coefficient is at least $c/(2s)$. In the complementary case, $(p-1)^{-1} < Ls/(1-\chi)$ in (131). These two cases prove (129), without an algorithmic test or batch restriction.

E.2 Inexact variance-reduced descent

For the variance-reduced method retain the exact population candidates (127). Path comparison and (130), with e_k, give

$$\max_i \|\bar{z}_{k,i}\|^2 \le 4(p-1)\rho_\chi^2 h(\Delta_k) + 4\rho_\chi^2 s^2 \|e_k\|_*^2 + 4\rho_\chi^2 \eta_k^2. \tag{133}$$

Define

$$\Theta_{0,\chi} := \frac{s}{c} + \frac{cs}{2} + 6Ls^2 + \frac{8L^2 s^3}{c}, \qquad D_\chi := \frac{c}{2s} + 6L + \frac{8L^2 s}{c}, \tag{134}$$

$$\widehat{\Theta}_\chi := \frac{s + Ls\Theta_{0,\chi}}{1-\chi}, \qquad \overline{\Theta}_\chi := (p-1)\widehat{\Theta}_\chi, \tag{135}$$

$$\lambda_\chi := \frac{c}{8s\rho_\chi^2}, \qquad w_\chi(m) := \frac{\mathrm{e}M_{\mathrm{trial},\chi}\sigma_*^2}{\lambda_\chi m}\exp\left(-\frac{m\lambda_\chi^2}{8L^2}\right). \tag{136}$$

For every positive batch size $m_k = |\mathcal{S}_k|$,

$$\begin{aligned}\mathbb{E}[\Phi(x_{k+1}) \mid \mathcal{F}_k] \le \Phi(x_k) - \frac{c}{2s}\mathbb{E}[h(\Delta_k) \mid \mathcal{F}_k] + \overline{\Theta}_\chi \mathbb{E}[\|e_k\|_*^2 \mid \mathcal{F}_k] \\ + D_\chi \eta_k^2 + w_\chi(m_k).\end{aligned} \tag{137}$$

To verify this, apply Lemma 6 with λ_χ and $M_{\mathrm{trial},\chi}$, then (133). The population-candidate error is at most

$$\frac{c(p-1)}{2s}\mathbb{E}[h(\Delta_k) \mid \mathcal{F}_k] + \frac{cs}{2}\mathbb{E}[\|e_k\|_*^2 \mid \mathcal{F}_k] + \frac{c}{2s}\eta_k^2 + w_\chi(m_k).$$

For $\delta_k = \Delta_k - \bar{z}_{k,I_k}$, the remainder difference bound gives

$$\begin{aligned}|\mathcal{E}_k(\Delta_k) - \mathcal{E}_k(\bar{z}_{k,I_k})| &\le 2L\|\Delta_k\|\|\delta_k\| + 3L\|\delta_k\|^2 \\ &\le \frac{c}{2s}h(\Delta_k) + \left(3L + \frac{4L^2(p-1)s}{c}\right)\|\delta_k\|^2.\end{aligned}$$

Use $\|\delta_k\|^2 \le 2s^2\|e_k\|_*^2 + 2\eta_k^2$. The Armijo test and Young's inequality for $-\langle e_k, \Delta_k\rangle$ then yield (137) with $\Theta_{0,\chi}$ in place of $\overline{\Theta}_\chi$. Finally use (132) when $L(p-1)s \le 1-\chi$, and $\Theta_{0,\chi} < (p-1)Ls\Theta_{0,\chi}/(1-\chi)$ otherwise. This proves the displayed coefficient $\overline{\Theta}_\chi$.

E.3 Residual transfer and summation

Lemma 9 (Inexact residual transfer)**.** *At every accepted inexact minibatch iteration,*

$$\mathrm{dist}_{p^*}(0, \partial\Phi(x_{k+1})) \le C_{\mathrm{res},\chi}\|\Delta_k\| + \|e_k\|_* + \frac{\eta_k}{\underline{t}_\chi}. \tag{138}$$

Consequently, (126) holds. The same bounds hold with $e_k = v_k - \nabla f(x_k)$ for the variance-reduced method.

Proof. The absolute KKT bound supplies $u_{k+1} \in \partial r(x_{k+1})$ such that

$$\|\nabla f(x_{k+1}) + u_{k+1}\|_* \le \left(L + \frac{1}{(p-1)t_k}\right)\|\Delta_k\| + \|e_k\|_* + \frac{\eta_k}{t_k}.$$

Use $t_k \ge \underline{t}_\chi$, then square with $(a+b+c)^2 \le 3(a^2+b^2+c^2)$ and $\|\Delta_k\|^2 = 2(p-1)h(\Delta_k)$. □

The conditional integrability arguments in Appendices A and B still apply. Conditional on the history, the exact population candidates are finite, and (130) gives a finite conditional displacement moment. The descent bounds above propagate integrability of $\Phi(x_k) - \Phi_{\inf}$ and $h(\Delta_k)$. The SPIDER recursion (71) depends only on the historical displacement, not on exact proximal optimality, so it propagates finite estimator moments for the inexact iterates as well. Lemma 9 completes the induction.

Proof of Theorem 4: minibatch case. Summing (129) from the first iteration gives

$$\frac{c}{2s}\sum_{k=0}^{K-1}\mathbb{E}h(\Delta_k) \le \Phi(x_0) - \Phi_{\inf} + \widehat{\Gamma}_\chi\sigma_*^2\sum_{k=0}^{K-1}N_k^{-1} + D_{\mathrm{mb},\chi}\sum_{k=0}^{K-1}\eta_k^2. \tag{139}$$

Sum (126) and substitute this bound and (52). For every $K \ge 1$,

$$\begin{aligned}
&K\mathbb{E}\,\mathrm{dist}_{p^*}^2(0, \partial\Phi(x_{\mathrm{out}})) \\
&\le \frac{12s(p-1)C_{\mathrm{res},\chi}^2}{c}\left[\Phi(x_0) - \Phi_{\inf} + \widehat{\Gamma}_\chi\sigma_*^2\sum_{k=0}^{K-1}N_k^{-1} + D_{\mathrm{mb},\chi}\sum_{k=0}^{K-1}\eta_k^2\right] \\
&\quad + \frac{3\sigma_*^2}{p-1}\sum_{k=0}^{K-1}N_k^{-1} + \frac{3}{\underline{t}_\chi^2}\sum_{k=0}^{K-1}\eta_k^2.
\end{aligned} \tag{140}$$

The tolerance series

$$\sum_{k=0}^{\infty}\eta_k^2 = \eta_0^2\sum_{k=1}^{\infty}k^{-2} < \infty \tag{141}$$

is finite. Define

$$\begin{aligned}
C_{\mathrm{mb}}^{\mathrm{in}} := &\frac{12s(p-1)C_{\mathrm{res},\chi}^2}{c}\left[\Phi(x_0) - \Phi_{\inf} + D_{\mathrm{mb},\chi}\sum_{k=0}^{\infty}\eta_k^2 + \widehat{\Gamma}_\chi\sigma_*^2\right] \\
&+ \frac{3\sigma_*^2}{p-1} + \frac{3}{\underline{t}_\chi^2}\sum_{k=0}^{\infty}\eta_k^2.
\end{aligned} \tag{142}$$

The nonnegative constant and noise coefficients in (140) are each at most $C_{\mathrm{mb}}^{\mathrm{in}}$, proving (117) for every positive deterministic batch sequence. For (11), the reciprocal sum is at most $1/c_b$; for (14), it is at most $1 + \log K$. The same counting and accuracy-budget arguments as Corollaries 1 and 2 give $\mathcal{O}(\varepsilon^{-4})$ and $\mathcal{O}(\varepsilon^{-4}\log^2(\mathrm{e} + 1/\varepsilon))$, respectively. □

For the variance-reduced case, use $R_K = \lceil K/q \rceil$ and define

$$E_{K,\chi} := \frac{\sigma_*^2 K}{p-1}\left(\frac{4\overline{\Theta}_\chi(q-1)}{b} - \frac{c}{8sL^2}\right)_+, \tag{143}$$

$$A_{K,\chi}^{\mathrm{in}} := \Phi(x_0) - \Phi_{\inf} + \frac{\overline{\Theta}_\chi\sigma_*^2 K}{(p-1)B} + R_K w_\chi(B) + (K - R_K)w_\chi(b) + E_{K,\chi} + D_\chi\sum_{k=0}^{K-1}\eta_k^2. \tag{144}$$

Proof of Theorem 4: VR case. The error recursion (69) holds on the inexact iterates. Multiply it by $\overline{\Theta}_\chi$ and apply (74) to its minimum, with

$$\theta = \frac{cb}{32s\overline{\Theta}_\chi L^2(q-1)}.$$

The displacement contribution is $c/(4s)$ times the displacement sum; the remaining variance contribution is exactly $E_{K,\chi}$. Consequently,

$$\overline{\Theta}_\chi \sum_{k=0}^{K-1} \mathbb{E}\|e_k\|_*^2 \le \frac{\overline{\Theta}_\chi \sigma_*^2 K}{(p-1)B} + \frac{c}{4s} \sum_{k=0}^{K-1} \mathbb{E}h(\Delta_k) + E_{K,\chi}.$$

Sum (137), use the lower bound $\Phi_{\inf}$, and substitute the last display. This gives

$$\sum_{k=0}^{K-1} \mathbb{E}h(\Delta_k) \le \frac{4s}{c} A_{K,\chi}^{\mathrm{in}}, \qquad \sum_{k=0}^{K-1} \mathbb{E}\|e_k\|_*^2 \le \frac{2A_{K,\chi}^{\mathrm{in}}}{\overline{\Theta}_\chi}. \tag{145}$$

The second inequality follows by substituting the first into the preceding error bound. Residual transfer and the independent uniform output imply, for every $K \ge 2$ and all $B, b \ge 1$, $2 \le q \le K$,

$$\begin{aligned} &K\mathbb{E}\,\mathrm{dist}_{p^*}^2(0, \partial\Phi(x_{\mathrm{out}})) \\ &\quad \le \left(\frac{24s(p-1)C_{\mathrm{res},\chi}^2}{c} + \frac{6}{\overline{\Theta}_\chi} \right) A_{K,\chi}^{\mathrm{in}} + \frac{3}{\underline{t}_\chi^2} \sum_{k=0}^{K-1} \eta_k^2. \end{aligned} \tag{146}$$

Under (20), $q/b \le 1/\log(\mathrm{e}+K)$, so

$$E_{K,\chi} \le \frac{\sigma_*^2 K}{p-1} \left(\frac{4\overline{\Theta}_\chi}{\log(\mathrm{e}+K)} - \frac{c}{8sL^2} \right)_+.$$

When the positive part is nonzero, $K < \exp(32s\overline{\Theta}_\chi L^2/c)$. Since $\log(\mathrm{e}+K) > 1$, this proves

$$E_{K,\chi} \le \frac{4\overline{\Theta}_\chi \sigma_*^2}{p-1} \exp(32s\overline{\Theta}_\chi L^2/c) \qquad (K \ge 2).$$

Also $B, b \ge \sqrt{K}$, and w_χ is decreasing. Hence

$$R_K w_\chi(B) + (K - R_K) w_\chi(b) \le \frac{\mathrm{e} M_{\mathrm{trial},\chi} \sigma_*^2}{\lambda_\chi} \sqrt{K} \exp\left(-\frac{\lambda_\chi^2 \sqrt{K}}{8L^2} \right) \le \frac{8 M_{\mathrm{trial},\chi} \sigma_*^2 L^2}{\lambda_\chi^3}.$$

Together with $B = K$ and (141), this gives $A_{K,\chi}^{\mathrm{in}} \le A_{\infty,\chi}^{\mathrm{in}} < \infty$, where

$$\begin{aligned} A_{\infty,\chi}^{\mathrm{in}} := \Phi(x_0) - \Phi_{\inf} + \frac{\overline{\Theta}_\chi \sigma_*^2}{p-1} \left[1 + 4\exp(32s\overline{\Theta}_\chi L^2/c) \right] \\ + \frac{8 M_{\mathrm{trial},\chi} \sigma_*^2 L^2}{\lambda_\chi^3} + D_\chi \sum_{k=0}^{\infty} \eta_k^2. \end{aligned} \tag{147}$$

Define

$$C_{\mathrm{vr}}^{\mathrm{in}} := \left(\frac{24s(p-1)C_{\mathrm{res},\chi}^2}{c} + \frac{6}{\overline{\Theta}_\chi} \right) A_{\infty,\chi}^{\mathrm{in}} + \frac{3}{\underline{t}_\chi^2} \sum_{k=0}^{\infty} \eta_k^2. \tag{148}$$

This proves (118) for every $K \ge 2$. The theoretical horizon $K = \max\{2, \lceil C_{\mathrm{vr}}^{\mathrm{in}}/\varepsilon^2 \rceil\}$ then certifies the same stationarity accuracy. The sample and gradient counts remain (76); the function and proximal-solve bounds use $M_{\mathrm{trial},\chi}$. Substitution into (22) gives the fixed-problem order $\mathcal{O}(\varepsilon^{-3}\sqrt{\log(\mathrm{e}+1/\varepsilon)})$. □

F Experimental Protocols and Detailed Results

F.1 Fixed-minibatch stochastic Rosenbrock experiment

We compare four methods:

1. SPL-MB, the direct ℓ_p proximal line search with $p = 1 + 1/(2\log d)$, normalized proximal function $h(z) = \|z\|_p^2/[2(p-1)]$, $s = 1$, $c = 0.7$, $\beta = 0.9$, and a stepsize reset every 50 iterations;

2. `slam-p50`, Euclidean SLAM (Wang et al., 2025) with $s = 1$, $c = 0.1$, $\beta = 0.9$, and the same reset period;

3. `sls0-1`, Algorithm 1 with $\mathrm{opt} = 0$ of the stochastic line-search method (Vaswani et al., 2019), with $\eta_{\max} = 1$, $c = 0.1$, and $\beta = 0.9$, whose accepted stepsize is carried to the next iteration without a periodic reset;

4. `sgd-const-tuned`, minibatch SGD with a tuned constant stepsize.

All four methods use a fixed batch size of 128, giving them the same stochastic-gradient sampling cost per iteration without batch growth. Since the Armijo parameters and initial-trial rules differ, the comparison evaluates the stated configurations rather than isolating the effect of geometry alone. A fresh batch is drawn at every iteration. Each line-search method reuses the current batch in all Armijo trials of that iteration.

For each dimension, SGD is run with every constant stepsize in $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ for one fifth of the total iteration budget. Each candidate is evaluated over five independent tuning runs, and the stepsize with the smallest mean final population objective is used in the comparison. The tuning streams are independent of the reported runs, and tuning time is excluded from the reported algorithm time.

We use a stochastic variant of the Rosenbrock objective (Rosenbrock, 1960),

$$F(x, \xi) := \sum_{i=1}^{d-1} \left[(100 + \xi)(x_{i+1} - x_i^2)^2 + (1 - x_i)^2 \right], \qquad \xi \sim \mathcal{N}(0, 10^2), \tag{149}$$

whose population minimizer is the all-ones vector and whose minimum value is zero. All methods start from $x_0 = (6, \ldots, 6)^\top$. The fixed iteration budget is $K = 24000$, and

$$d \in \{100, 200, 300, 500, 800, 1000, 1500, 2000\}.$$

Each configuration is repeated five times. We report the median population objective and median $\|\nabla f(x)\|_\infty$ at the final iterate, together with their interquartile ranges. Since the sampled Rosenbrock objective and gradient are affine in the Gaussian coefficient, their batch averages are generated from the exact Gaussian law of the sample mean. This is distributionally equivalent to explicitly averaging 128 independent Gaussian samples. Population quantities are evaluated only for reporting.

Under its stated assumptions, Theorem 1 gives the bound $C_0/K + C_1/N$ for a fixed batch size N; its vanishing-error complexity guarantees use horizon-proportional or linearly increasing batches. The unconstrained stochastic Rosenbrock model has unbounded curvature and state-dependent gradient noise, so it does not satisfy the global samplewise smoothness and uniform variance bounds in that assumption. The reported final-iterate values therefore serve as a finite-budget performance comparison, not a direct verification of the random-output stationarity theorem.

Figure 1 shows that SLAM attains the smallest final objective and stationarity values for $d \leq 300$. From $d = 500$ onward, SPL-MB has the smallest median value of both reported quantities among the four methods. Over $d \in \{500, 800, 1000, 1500, 2000\}$, its median final objective ranges from approximately 3×10^{-5} to 6×10^{-3}, and its median ℓ_∞ gradient norm ranges from approximately 3×10^{-3} to 4×10^{-2}. Since every method uses the same fixed batch and iteration budget, each uses $128K$ stochastic-gradient samples; their sampled-function and line-search costs can differ.

For the timing comparison, we use $d \in \{100, 200, 500, 800, 1000\}$ and stop a run when the population objective first satisfies $f(x) \leq 10^{-3}$, subject to the common cap $K = 96000$. The 128 samples are evaluated explicitly in this experiment. Algorithm time includes stochastic-gradient computation, proximal-direction computation, and same-batch Armijo evaluations, but excludes tuning and reporting. Figure 1 shows that SLAM is faster for $d = 100$ and $d = 200$. SPL-MB becomes faster from $d = 500$ onward; at $d = 800$ and $d = 1000$, its median time is approximately one half and two fifths, respectively, of the SLAM time. SPL-MB and SLAM reach the target in all five repetitions, whereas `sls0-1` and `sgd-const-tuned` do not reach it before the iteration cap.

F.2 Fixed-batch variance-reduced logistic-regression experiment

Following the smooth nonconvex experiment in the SpiderBoost study (Wang et al., 2019), we use the binary logistic-regression objective

$$f(w) := \frac{1}{n}\sum_{i=1}^{n} \log\bigl(1+\exp(-y_i x_i^\top w)\bigr) + 0.1\sum_{j=1}^{d}\frac{w_j^2}{1+w_j^2}, \tag{150}$$

where $x_i \in \mathbb{R}^d$ and $y_i \in \{-1,1\}$. The nonconvex regularizer is treated as part of the smooth finite-sum objective, and $r \equiv 0$. We use the LIBSVM training sets (Chang & Lin, 2011) `a9a`, with $n = 32561$ and $d = 123$, and `w8a`, with $n = 49749$ and $d = 300$, without an intercept or further feature scaling.

We compare three methods without additional parameter tuning:

1. SPL-VR with $p = 2$, $s = 1$, $c = 0.1$, $\beta_{\mathrm{ls}} = 0.9$, and a stepsize reset every 50 iterations;
2. `SpiderBoost`, Algorithm 1 of the SpiderBoost study with constant stepsize $\eta = 0.05$;
3. `SpiderBoost-M`, the smooth specialization of Prox-SpiderBoost-M, namely Algorithm 3 with $r \equiv 0$ and $\alpha_k = 2/(\lceil k/q\rceil + 1)$. We use

$$\beta_k = 0.05, \qquad \lambda_k = (1+\alpha_{k+1})\beta_k, \tag{151}$$

   Here λ_k uses the epochwise coefficient α_{k+1} from the interpolation step.

For every method, a refresh uses the full gradient, the inner batch has size $|\xi_k| = 256$, and

$$q := \left\lceil \frac{2n}{|\xi_k|} \right\rceil = \left\lceil \frac{2n}{256} \right\rceil. \tag{152}$$

Inner samples are drawn with replacement. In SPL-VR, the same inner batch is used for the SPIDER gradient difference, the affine sampled model, and every Armijo trial at that iteration. This experiment uses fixed full-gradient refreshes and fixed inner batches rather than the prescribed horizon-dependent schedule in (20).

We use five independent repetitions with shared standard-normal initial points and common sampled indices across methods. The budget is 80 refresh epochs for `a9a` and 25 refresh epochs for `w8a`. One refresh epoch consists of q iterations, including one full-gradient refresh, not one pass through the data. Estimator refreshes and stepsize resets are separate mechanisms.

Independent numerical references f^* are computed using 13 deterministic `fminunc` runs per data set: one zero and twelve pseudorandom initializations at three fixed scales, followed by trust-region refinement of the best point. The objective values agree within 10^{-10}, giving $f^* = 0.5057912583706641$ for `a9a` and $f^* = 0.4656997893435263$ for `w8a`, with ℓ_∞ gradient norms below 1.4×10^{-9}. These are numerical references, not certificates of global optimality.

Figure 2 reports median gaps $f(w) - f^*$ and interquartile ranges at common checkpoints, against refresh epochs and median cumulative algorithm time. Algorithm time includes sampling, gradient evaluations, proximal updates, and all Armijo trials; it excludes diagnostics used only for reporting, reference computation, data loading, and file output. Gaps below 10^{-12} are displayed at 10^{-12}.

SPL-VR reaches a median gap below 10^{-8} in fewer refresh epochs and less algorithm time than both baselines on both data sets, despite its additional Armijo evaluations. All three methods eventually attain essentially the same final objective value.

F.3 Fixed-batch variance reduction with an ℓ_1 regularizer

To evaluate the methods on a nonsmooth composite problem, we add an ℓ_1 penalty to (150), following the experiment in Figure 3(a)–(b) of Wang et al. (2019):

$$\Phi(w) := f(w) + 0.1\|w\|_1. \tag{153}$$

Here $r(w) = 0.1\|w\|_1$ is handled by the proximal update. We retain the preceding experiment's data, parameters, full-gradient refreshes, inner batches, budgets, and five paired random streams,

without retuning. All methods are rerun on the composite objective; no trajectories or timings are reused from the smooth experiment.

The three methods now use the following composite updates:

1. SPL-VR solves each trial subproblem by soft thresholding:
$$[w_k + z_k(t)]_j = \operatorname{sign}(w_{k,j} - tv_{k,j}) \max\{|w_{k,j} - tv_{k,j}| - 0.1t, 0\}.$$
It recomputes this step at each trial stepsize and tests
$$H_k(w_k + z_k(t)) + r(w_k + z_k(t)) \le H_k(w_k) + r(w_k) - \frac{c}{t}\|z_k(t)\|_2^2.$$
2. Prox-SpiderBoost, Algorithm 2 of Wang et al. (2019), uses $\eta = 0.05$ and the update $w_{k+1} = \operatorname{prox}_{\eta r}(w_k - \eta v_k)$.
3. Prox-SpiderBoost-M uses Algorithm 3 of Wang et al. (2019) with (151), retaining its coupled momentum update. Function values are evaluated at its interpolated sequence z_k, where the gradient estimator is constructed, and at w_k for the other methods.

Independent references Φ_{ref} are computed by deterministic proximal-gradient solves from zero and twelve Gaussian initializations, four at each scale 0.1, 1, and 3, stopping when $\operatorname{dist}_\infty(0, \partial\Phi(w)) \le 10^{-10}$. The attained values agree within 10^{-10} and are cross-checked with L-BFGS-B on an equivalent bound-constrained formulation, giving $\Phi_{\text{ref}} = 0.6494559882763125$ for `a9a` and $\Phi_{\text{ref}} = 0.6809400097426740$ for `w8a`. They are not certificates of global optimality.

Figure 3 reports the composite objective gap $\Phi(w) - \Phi_{\text{ref}}$ with the same timing convention, using medians and interquartile ranges. Epoch curves use recorded checkpoints; time curves linearly interpolate objective values on a common time grid for each method, without extrapolation. Neither uses a best-so-far envelope or estimates the theorem's random-output stationarity measure.

On both data sets, SPL-VR reaches a gap of 10^{-2} faster than either baseline. At 10^{-6}, it is faster than Prox-SpiderBoost and comparable to Prox-SpiderBoost-M. At 10^{-8}, all five runs of every method succeed; SPL-VR remains faster than Prox-SpiderBoost, while the momentum method is faster in algorithm time, especially on `w8a`. All methods attain essentially the same final objective, with SPL-VR taking slightly more time over the complete budgets. The results demonstrate its use with $r \neq 0$, not a uniform speed advantage at every accuracy.

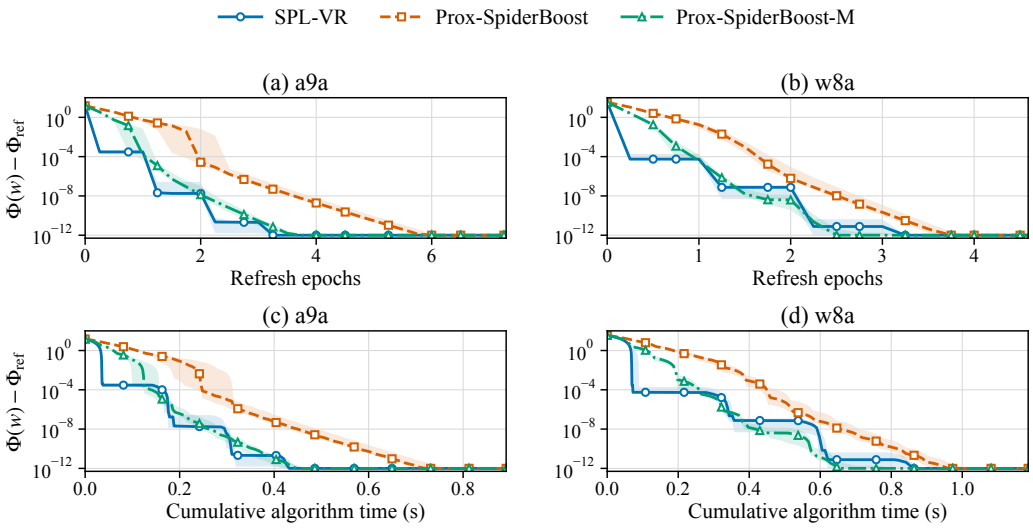


Figure 3: Nonconvex logistic regression with an ℓ_1 regularizer. Top: median composite objective gap against refresh epochs on `a9a` (left) and `w8a` (right). Bottom: the same gap measure against cumulative algorithm time. Shaded regions are interquartile ranges over five repetitions. Gaps below 10^{-12} are displayed at 10^{-12}. All runs complete the stated budgets; only the common flat terminal tail below the display floor is omitted.