

A Double-Inertial Variance-Reduced Proximal Alternating Method for Block-Structured Nonconvex Minimax Optimization

Yekini Shehu

yekini.shehu@zjnu.edu.cn

Zhejiang Normal University

Yonghong Yao

Tiangong University

Research Article

Keywords: nonconvex–nonconcave minimax optimization, inertial acceleration, variance reduction, Minty variational inequality, proximal alternating methods

Posted Date: October 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11246774/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

A Double-Inertial Variance-Reduced Proximal Alternating Method for Block-Structured Nonconvex Minimax Optimization

Yekini Shehu*

Yonghong Yao†

October 3, 2026

Abstract

We study nonconvex–nonconcave finite-sum minimax problems under the Minty variational inequality condition. We propose liSminimaxPALM, a double-inertial variance-reduced proximal alternating method whose query and post-update inertia are unrestricted in $[0, 1]$, the zero-inertia case included. With the SAGA estimator the method achieves evaluation complexity $N + \mathcal{O}(N^2 L^2 D^2 / \varepsilon^2)$ and an epoch SARAH variant achieves $\mathcal{O}(\sqrt{N} L^4 D^2 / \varepsilon^2)$, where D is the diameter of the domain around the Minty point. The N^2 single-loop scaling is proved optimal among unbiased table-based estimators *within anchor-type analyses* by a staleness lower bound: no unbiased estimator refreshing b entries per iteration contracts faster than $1 - b/N$ (Theorem A.2), and the ε^{-2} scaling is proved class-optimal at $N = 1$ by an explicit star-monotone nonconvex hard instance (Theorem E.3). Whether the Minty class admits a $\sqrt{N} L \Delta / \varepsilon^2$ lower bound matching Proposition 4.10 is stated as a conjecture with an explicit construction program (Appendix D); the strongest available bound of that form [3] is proved for a nonconvex zero-chain outside the Minty class. Under strong monotonicity of the saddle operator (covering strongly convex–strongly concave composite problems) the deterministic non-inertial variant converges R -linearly with certified rate $(1 + \eta\mu)^{-1}$, and the prox-center inertial variant is shown to preserve linear convergence for sufficiently small inertia, and on quadratic games the full range of inertia is certified in the decoupled case, and the tested coupled instances are certified by exact spectral computation (Theorem 5.6, Remark 5.7). For the main stochastic complexity theorem the prox-center parameter is taken $\alpha = 0$; the inertial prox-center case is left as an explicit open problem (Remark 4.2). Numerical experiments on a benchmark on which the Minty condition is *proved* confirm the oracle advantage at large N , the geometric decay, and convergence for every tested inertial configuration, including $\gamma = 0$ and the beyond-theory case $\alpha > 0$.

Keywords: nonconvex–nonconcave minimax optimization; inertial acceleration; variance reduction; Minty variational inequality; proximal alternating methods.

Mathematics Subject Classification: 90C26, 90C15, 90C47, 68Q25, 68T05.

1 Introduction

1.1 Problem and motivation

We consider

$$\min_{x \in \mathbb{R}^{m_1}} \max_{y \in \mathbb{R}^{m_2}} \Psi(x, y) = H(x, y) + f(x) - g(y), \quad H(x, y) = \frac{1}{N} \sum_{i=1}^N H_i(x, y), \quad (1)$$

* (Corresponding Author) School of Mathematical Sciences, Zhejiang Normal University, Jinhua 321004, China; e-mail: yekini.shehu@zjnu.edu.cn

† School of Mathematical Sciences, Tiangong University, Tianjin 300387, China; and Department of Medical Research, China Medical University Hospital, China Medical University, Taichung, Taiwan; e-mail: yy-htgu@hotmail.com

with each H_i smooth, and $f : \mathbb{R}^{m_1} \rightarrow (-\infty, +\infty]$, $g : \mathbb{R}^{m_2} \rightarrow (-\infty, +\infty]$ proper lower semicontinuous convex (possibly indicators of convex sets). Problem (1) arises in adversarial training [27], generative adversarial networks [20], robust tensor decomposition [23], and adversarial dictionary learning [26].

Positioning. Proximal alternating methods are classical for block nonconvex *minimization* [5, 32, 19, 35, 36, 4], and deterministic proximal alternating schemes exist for nonconvex–concave *minimax* [24, 6]. On the stochastic side, extragradient-type methods work under generalized monotonicity and (weak) Minty assumptions [15, 17, 8, 13], and inertial proximal methods are studied for monotone inclusions arising from GANs [10], double-inertial extrapolation being classical in variational-inequality methods [40], and variance reduction for saddle-point problems and monotone inclusions is classical [30, 25, 37, 1, 2, 7]. A variance-reduced inertial block-coordinate proximal alternating method under the Minty MVI condition, with matching lower-bound certificates, has been missing. This paper develops such a method. Three qualifications are stated at the outset and used throughout: our single-loop SAGA bound carries an N^2 dependence that we prove is information-theoretic for unbiased estimators (Appendix A); our linear-convergence theorem covers the deterministic non-inertial variant, with the inertial extension given for small inertial parameters (Appendix C) and the stochastic rate proved in [37] for the smooth non-inertial case; and the main stochastic complexity analysis assumes a vanishing prox-center inertial parameter ($\alpha = 0$), the inertial prox-center case being open in the anchored framework (Remark 4.2).

1.2 The Minty variational inequality

Definition 1.1 (Minty MVI [28, 17]). Problem (1) satisfies the Minty MVI condition if there exists $z^* = (x^*, y^*) \in \text{dom } \partial\Psi$ such that for all $z = (x, y) \in \text{dom } \partial\Psi$, all $u \in \partial f(x)$ and $v \in \partial g(y)$,

$$\langle \nabla_x H(x, y) + u, x - x^* \rangle + \langle -\nabla_y H(x, y) + v, y - y^* \rangle \geq 0. \quad (2)$$

If H is convex–concave and f, g convex, (2) holds at any saddle point. The condition excludes cycling and legitimizes the proximal gradient residual as a stationarity measure [17]; Wasserstein GANs with gradient penalty are a practical example satisfying a weak Minty condition [22].

Hierarchy of well-posedness conditions. (i) convex–concave $\Rightarrow$ (ii) Minty MVI (2); (iii) *weak Minty MVI* [17]: (2) holds for *some* z^* (rather than at every solution); (iv) *negative comonotonicity* [8, 13, 9]: $\langle F(z) - F(z'), z - z' \rangle \geq -\alpha \|F(z) - F(z')\|^2$, strictly weaker than Minty; (v) *interaction dominance* [21], a curvature condition guaranteeing convergence of proximal-point and descent–ascent schemes outside the Minty class.

1.3 Contributions

- (i) **Algorithm.** liSminimaxPALM (Section 3): the double-inertial mechanism of [36] transferred to the minimax setting, with post-update inertia based on $x_{k+1} - x_k$; our analysis imposes no lower bound on the post-update inertia beyond $[0, 1)$, covering in particular $\gamma_i^k = 0$ (the inertial hypotheses of [35, 36], the latter as corrected, are not invoked here). Query inertia $\beta_i^k \in [0, 1)$ is covered; the complexity theorem takes the prox-center parameter $\alpha = 0$ (Remark 4.2).
- (ii) **Complexity, with both directions.** SAGA: $N + \mathcal{O}(N^2 L^2 D^2 / \varepsilon^2)$ (Theorem 4.8); epoch SARAH: $\mathcal{O}(\sqrt{N} L^4 D^2 / \varepsilon^2)$ (Proposition 4.10). The N^2 single-loop scaling is proved optimal among unbiased table-based estimators *within anchor-type analyses* (Theorem A.2, Corollary B.2); lower bounds for the Minty class transfer through the convex embedding (Proposition 4.11), and the ε^{-2} scaling is class-optimal at $N = 1$ via an explicit star-monotone

nonconvex instance (Theorem E.3); the optimal $\sqrt{N} L \Delta / \varepsilon^2$ bound is conjectured with a construction program (Appendix D).

- (iii) **Linear convergence.** Under strong monotonicity, the deterministic non-inertial variant converges R -linearly with certified rate $(1 + \eta\mu)^{-1}$ (Theorem 5.3), and the inertial variant preserves linear convergence for $\alpha \leq \eta\mu/50$ (Theorem C.1). On quadratic games, the decoupled case certifies the full range of inertia with acceleration $1 - \Theta(\kappa^{-1/2})$; coupled instances are certified for small coupling by the theorem and for the tested instance by exact spectral computation (Theorem 5.6, Remark 5.7).
- (iv) **Adaptive steps.** A backtracking variant retaining the complexity for initial step below a problem-dependent threshold (Theorem 6.1).
- (v) **Provably-Minty benchmark.** A nonconvex–nonconcave finite-sum game on which the Minty condition is *proved* (Proposition 7.1); all experiments use it (Section 7).

2 Preliminaries

We use the standard subdifferential calculus for proper l.s.c. convex functions [34].

Assumption 2.1 (Smoothness and block structure). (i) $H = \frac{1}{N} \sum_{i=1}^N H_i$ with each H_i finite-valued and C^1 ; (ii) $\nabla_x H(\cdot, y)$ is globally $L_1(y)$ -Lipschitz and $\nabla_y H(x, \cdot)$ is globally $L_2(x)$ -Lipschitz; ∇H is Lipschitz on bounded sets with constant M ; (iii) $\inf_{x,y} \Psi(x, y) > -\infty$; (iv) f and g are proper l.s.c. convex.

Assumption 2.2 (Boundedness). (i) $\text{dom } f$ and $\text{dom } g$ are bounded; or (ii) the sequence $\{z_k\}$ generated by Algorithm 1 is contained in a bounded set.

Remark 2.3. Under 2.2(i) the iterates and the extrapolated points $(\tilde{x}_k, \hat{x}_k, \bar{x}_{k+1})$ (resp. $(\tilde{y}_k, \hat{y}_k, \bar{y}_{k+1})$) remain in a bounded set: each is of the form $(1 + \theta)u - \theta v$ with $\theta \in [0, 1]$ and $u, v \in \text{dom } f$ (resp. $\text{dom } g$). We define the uniform constants $\bar{L}_1 = \sup_{y \in \mathcal{Y}} L_1(y)$, $\bar{L}_2 = \sup_{x \in \mathcal{X}} L_2(x)$, $\bar{M} = \sup \frac{\|\nabla H(z) - \nabla H(z')\|}{\|z - z'\|}$ over the bounded set, and

$$\bar{L} := \max\{\bar{L}_1, \bar{L}_2\} \leq \bar{M}. \quad (3)$$

If L is a per-component constant, $\|\nabla H_i(z) - \nabla H_i(z')\| \leq L\|z - z'\|$ for all i , then $\bar{L}_1, \bar{L}_2, \bar{M} \leq L$. All constants are $\mathcal{F}_k$ -measurable.

Definition 2.4 (Proximal gradient mapping). For $\eta_1, \eta_2 > 0$, $\mathcal{G}_\eta(z) := (\eta_1^{-1}(x - \text{Prox}_{\eta_1}^f(x - \eta_1 \nabla_x H(x, y))), \eta_2^{-1}(y - \text{Prox}_{\eta_2}^g(y + \eta_2 \nabla_y H(x, y))))$, and z is stationary iff $\mathcal{G}_\eta(z) = 0$.

Definition 2.5 (Block-partial first-order oracle). A query with (x, y, i, b) , $b \in \{x, y\}$, returns $\nabla_x H_i(x, y)$ or $\nabla_y H_i(x, y)$; querying both blocks of component i counts as two evaluations. Algorithm 1 draws two independent indices per iteration (one per block) and therefore uses two evaluations per iteration; a full gradient costs N component evaluations ($2N$ block queries).

Our estimator-error template follows [14], adapted to the minimax setting.

Definition 2.6 (Variance-reduced minimax estimator). A gradient estimator $\tilde{\nabla} H = (\tilde{\nabla}_x H, \tilde{\nabla}_y H)$ is variance-reduced with constants $V_1, V_\Gamma \geq 0$ and $\tau \in (0, 1]$ if there exist random vectors $B_k = (B_k^x, B_k^y)$ (deterministic error) and $S_k = (S_k^x, S_k^y)$ (stochastic error) such that

- (i) **Decomposition and unbiasedness.** $\tilde{\nabla}_x H(\hat{x}_k, \bar{y}_k) = \nabla_x H(z_{k+1}) - B_k^x - S_k^x$ and $\tilde{\nabla}_y H(\bar{x}_{k+1}, \hat{y}_k) = \nabla_y H(z_{k+1}) + B_k^y + S_k^y$, with $\mathbb{E}_k[S_k^x] = 0$, $\mathbb{E}_k[S_k^y] = 0$ (the two block queries use independent draws; conditionally on the x -draw the y -query point is fixed and the estimator is unbiased there, so $\mathbb{E}_k[S_k^y] = 0$ holds by the tower property);

(ii) **Second-moment bound.** there exist $\mathcal{F}_k$ -measurable v_k^i and $m_k \geq 1$ such that, with $\Gamma_k = \frac{1}{m_k} \sum_{i=1}^{m_k} \|v_k^i\|^2$,

$$\mathbb{E}_k [\|S_k^x\|^2 + \|S_k^y\|^2] \leq \Gamma_k + V_1 (\mathbb{E}_k \|z_{k+1} - z_k\|^2 + \|z_k - z_{k-1}\|^2 + \|z_{k-1} - z_{k-2}\|^2);$$

(iii) **Contraction.** $\mathbb{E}_k \Gamma_{k+1} - (1 - \tau)\Gamma_k \leq V_\Gamma (\mathbb{E}_k \|z_{k+1} - z_k\|^2 + \|z_k - z_{k-1}\|^2 + \|z_{k-1} - z_{k-2}\|^2)$.

Lemma 2.7 (SAGA satisfies Definition 2.6). *Let L be a per-component Lipschitz constant on the bounded set. The SAGA estimator [16]—two independent uniform draws per iteration (one per block), table initialized at $z_0 = z_{-1}$, both block queries reading the table in its state at the start of the iteration, and exactly one entry (the sampled x -index) refreshed after the queries at the x -query point—satisfies Definition 2.6 with*

$$V_1 = 24L^2, \quad \tau = \frac{1}{2N}, \quad V_\Gamma = 16NL^2, \quad \frac{V_\Gamma}{\tau} = 32N^2L^2. \quad (4)$$

Proof. Let $g_j^k = \nabla H_j(\phi_j^k)$ be the table entry and $\bar{g}^k = \frac{1}{N} \sum_j g_j^k$. For a query point w and uniform index j , $\tilde{\nabla}H(w) = \nabla H_j(w) - g_j^k + \bar{g}^k$; with $X_j = \nabla H_j(w) - g_j^k$,

$$\mathbb{E}_k \|\tilde{\nabla}H(w) - \nabla H(w)\|^2 = \mathbb{E}_k \|X_j - \mathbb{E}_k X_j\|^2 \leq \mathbb{E}_k \|X_j\|^2 = \frac{1}{N} \sum_j \|\nabla H_j(w) - g_j^k\|^2.$$

With $a = \nabla H_j(w) - \nabla H_j(z_{k-1})$ and $b = \nabla H_j(z_{k-1}) - g_j^k$, $(a + b)^2 \leq 2a^2 + 2b^2$ gives, for $\Gamma_k := \frac{1}{N} \sum_j \|\nabla H_j(\phi_j^k) - \nabla H_j(z_{k-1})\|^2$,

$$\mathbb{E}_k \|\tilde{\nabla}H(w) - \nabla H(w)\|^2 \leq 2L^2 \|w - z_{k-1}\|^2 + 2\Gamma_k. \quad (5)$$

For the x -query $w_k^x = (\hat{x}_k, \bar{y}_k)$: $\hat{x}_k - x_{k-1} = (1 + \beta_1^k)\Delta_k^x$, $\bar{y}_k - y_{k-1} = (1 + \gamma_2^{k-1})\Delta_k^y$, hence $\|w_k^x - z_{k-1}\|^2 \leq 4\|\Delta_k\|^2$. For the y -query $w_k^y = (\bar{x}_{k+1}, \hat{y}_k)$: $\bar{x}_{k+1} - x_{k-1} = \Delta_{k+1}^x + (1 + \gamma_1^k)\Delta_k^x$, $\hat{y}_k - y_{k-1} = (1 + \beta_2^k)\Delta_k^y$, hence $\|w_k^y - z_{k-1}\|^2 \leq 2\|\Delta_{k+1}\|^2 + 8\|\Delta_k\|^2$. Since the x -query is $\mathcal{F}_k$ -measurable and the y -query is fixed conditionally on the x -index,

$$\mathbb{E}_k \|S_k^x\|^2 \leq 8L^2 \|\Delta_k\|^2 + 2\Gamma_k, \quad \mathbb{E}_k \|S_k^y\|^2 \leq 4L^2 \mathbb{E}_k \|\Delta_{k+1}\|^2 + 16L^2 \|\Delta_k\|^2 + 2\Gamma_k.$$

Summing and rescaling $\Gamma \leftarrow 4\Gamma$ yields property (ii) with variance term Γ_k and $V_1 = 24L^2$ (which dominates the $4L^2$ coefficient of $\mathbb{E}_k \|\Delta_{k+1}\|^2$). Property (i) holds: $\mathbb{E}_k[S_k^x] = 0$ since the x -query is $\mathcal{F}_k$ -measurable, and $\mathbb{E}_k[S_k^y] = 0$ conditionally on the x -draw since the y -index is an independent uniform draw reading the start-of-iteration table.

For property (iii), let w_k be the stored point of the refreshed index; then $\|w_k - z_k\|^2 = \bar{\beta}^2 \|\Delta_k^x\|^2 + \bar{\gamma}^2 \|\Delta_k^y\|^2 \leq \|\Delta_k\|^2$. With probability $1 - \frac{1}{N}$ the table is unchanged and only the reference moves; with probability $\frac{1}{N}$ one entry is replaced by w_k . Hence

$$\mathbb{E}_k \Gamma_{k+1} = \frac{N-1}{N} \cdot \frac{1}{N} \sum_j \|\nabla H_j(\phi_j^k) - \nabla H_j(z_k)\|^2 + \frac{1}{N} \cdot \frac{1}{N} \sum_j \|\nabla H_j(w_k) - \nabla H_j(z_k)\|^2.$$

Young's inequality $\|u+v\|^2 \leq (1+a)\|u\|^2 + (1+\frac{1}{a})\|v\|^2$ with $a = \frac{1}{2N}$, $u = \nabla H_j(\phi_j^k) - \nabla H_j(z_{k-1})$, $v = \nabla H_j(z_{k-1}) - \nabla H_j(z_k)$ bounds the first term by $\frac{N-1}{N} [(1 + \frac{1}{2N})\Gamma_k + (1 + 2N)L^2 \|\Delta_k\|^2]$. Since $(1 - \frac{1}{N})(1 + \frac{1}{2N}) = 1 - \frac{1}{2N} - \frac{1}{2N^2} \leq 1 - \frac{1}{2N}$ and $(1 - \frac{1}{N})(1 + 2N)L^2 \leq 3NL^2$, while the refresh term is at most $\frac{L^2}{N} \|\Delta_k\|^2 \leq NL^2 \|\Delta_k\|^2$,

$$\mathbb{E}_k \Gamma_{k+1} \leq \left(1 - \frac{1}{2N}\right) \Gamma_k + 4NL^2 \|\Delta_k\|^2.$$

Rescaling by the same factor 4 gives (iii) with $\tau = \frac{1}{2N}$, $V_\Gamma = 16NL^2$, hence $V_\Gamma/\tau = 32N^2L^2$. $\square$

Algorithm 1 IISminimaxPALM

Require: Initial points $(x_{-1}, y_{-1}) = (x_0, y_0)$; inertial parameters $\alpha_i^k, \beta_i^k, \gamma_i^k \in [0, 1]$; step sizes $\eta_i^k > 0$.

```

1: for  $k = 0, 1, 2, \dots$  do
2:    $\tilde{x}_k = x_k + \alpha_1^k(x_k - x_{k-1}); \quad \hat{x}_k = x_k + \beta_1^k(x_k - x_{k-1})$ 
3:    $x_{k+1} \in \text{Prox}_{\eta_1^k}^f(\tilde{x}_k - \eta_1^k \tilde{\nabla}_x H(\hat{x}_k, \bar{y}_k))$ 
4:    $\bar{x}_{k+1} = x_{k+1} + \gamma_1^k(x_{k+1} - x_k)$ 
5:    $\tilde{y}_k = y_k + \alpha_2^k(y_k - y_{k-1}); \quad \hat{y}_k = y_k + \beta_2^k(y_k - y_{k-1})$ 
6:    $y_{k+1} \in \text{Prox}_{\eta_2^k}^g(\tilde{y}_k + \eta_2^k \tilde{\nabla}_y H(\bar{x}_{k+1}, \hat{y}_k))$  ▷ proximal ascent
7:    $\bar{y}_{k+1} = y_{k+1} + \gamma_2^k(y_{k+1} - y_k)$ 
8: end for

```

3 The IISminimaxPALM algorithm

Remark 3.1 (Design). Compared with [36] the maximizer update (line 6) uses the proximal operator for g , reflecting the ascent nature of the inner maximization. The post-update inertia (lines 4 and 8) uses $x_{k+1} - x_k$ rather than $x_{k+1} - \bar{x}_k$; consequently the analysis of Section 4 requires no nonzero lower bound on γ_i^k , and in particular $\gamma_i^k = 0$ is covered (for the inertial hypotheses of [35, 36], including the correction to the latter, we refer to those works). With $\alpha_i^k = \beta_i^k = \gamma_i^k = 0$ and exact gradients the method is proximal gradient descent–ascent; with stochastic gradients it is SPRING [18] adapted to minimax. Throughout the convergence analysis of Section 4 we set the prox-center parameter to $\alpha_i^k = 0$ (Assumption 4.1); the query and post-update parameters $\beta_i^k, \gamma_i^k \in [0, 1)$ remain unrestricted. The inertial prox-center case $\alpha_i^k > 0$ is open in the anchored framework and is discussed in Remark 4.2.

4 Convergence analysis

Table 1: Notation for Section 4.

Symbol	Meaning
$z_k = (x_k, y_k), z^*, \bar{z}_k$	iterate, Minty point, post-update extrapolation
$\Delta_k = z_k - z_{k-1}$	displacement
$\Lambda_k = \frac{1}{2}\ z_k - z^*\ ^2; C_0, \rho'$	distance potential; summation constants (Lemma 4.5)
B_k, S_k, Γ_k	estimator bias, stochastic error, table potential (Def. 2.6)
D	diameter $\sup_k \ z_k - z^*\ $ (7)
$\tau, V_1, V_\Gamma, \Theta = V_\Gamma/\tau$	estimator constants (Def. 2.6)
η, β_i, γ_i	step size and inertial parameters ($\alpha_i = 0$)
$c = 3\eta\bar{M}, a_0 = 1 - c; c_0, \rho, c_k, p_1, p_2, q$	Lyapunov coefficients (Lemmas 4.4–4.5)
$L, \bar{L}_1, \bar{L}_2, \bar{L}, \bar{M}$	smoothness constants (Remark 2.3)
μ_0	residual-bound constant (Lemma 4.7)

Write $\Delta_{k+1} = z_{k+1} - z_k$, $\Delta_k = z_k - z_{k-1}$, $\Delta_{k-1} = z_{k-1} - z_{k-2}$, $\tilde{z}_k = z_k + (\alpha_1 \Delta_k^x, \alpha_2 \Delta_k^y)$ and

$$w_k := z_{k+1} - \tilde{z}_k = \Delta_{k+1} - (\alpha_1 \Delta_k^x, \alpha_2 \Delta_k^y). \quad (6)$$

Throughout, $z^* = (x^*, y^*)$ is the Minty point of Definition 1.1 and

$$D := \sup_{k \geq 0} \|z_k - z^*\| < \infty \quad (7)$$

(finite by Assumption 2.2 and Remark 2.3; under 2.2(i), D is at most the diameter of the set of extrapolations, a pure problem constant). In this section the prox-center parameter vanishes, $\alpha_i = 0$ (Assumption 4.1), so that $\tilde{z}_k = z_k$, $w_k = \Delta_{k+1}$, and the distance potential is the plain

$\Lambda_k = \frac{1}{2}\|z_k - z^*\|^2$. Let $u \in \partial f(x_{k+1})$ and $v \in \partial g(y_{k+1})$ be the subgradients furnished by the first-order conditions of the two proximal subproblems, and let $B_k = (B_k^x, B_k^y)$ and $S_k = (S_k^x, S_k^y)$ be the decomposition of Definition 2.6(i), so that

$$\tilde{\nabla}_x H = \nabla_x H(z_{k+1}) - B_k^x - S_k^x, \quad \tilde{\nabla}_y H = \nabla_y H(z_{k+1}) + B_k^y + S_k^y \quad (8)$$

at the respective query points (the asymmetric signs reflect the descent/ascent structure).

Assumption 4.1 (Parameter conditions). The prox-center parameters vanish, $\alpha_i^k \equiv 0$; the query and post-update inertial parameters satisfy $0 \leq \beta_i^k, \gamma_i^k \leq 1$. The step sizes are fixed, $\eta_1^k = \eta_2^k = \eta$, where

$$\eta \leq \min\left\{1, \frac{1}{300\bar{M}}, \frac{1}{256\sqrt{\Theta}}\right\}. \quad (9)$$

Remark 4.2 (Why $\alpha = 0$, and what fails for $\alpha > 0$). With an inertial prox-center ($\alpha_i > 0$) the anchor naturally produces the subtracted potential $\frac{1}{2}\|z_k - z^*\|^2 - \frac{\alpha_1}{2}\|x_{k-1} - x^*\|^2 - \frac{\alpha_2}{2}\|y_{k-1} - y^*\|^2$, and bounding the bias inner product $2\eta\langle B_k, z_{k+1} - z^* \rangle$ by Young's inequality produces a term proportional to $\|z_{k+1} - z^*\|^2$. Converting this term back into the subtracted potential regenerates, at every iteration, a bounded term proportional to $\eta\bar{M}\bar{\alpha}D^2$ ($\bar{\alpha} = \max \alpha_i$): indeed $\|z_k - z^*\|^2 = 2\Lambda_k + (\alpha_1\|x_{k-1} - x^*\|^2 + \alpha_2\|y_{k-1} - y^*\|^2)$, and the second summand is bounded only by D^2 . The standard repairs (perturbing the potential, absorbing into the slack of $\|w_k\|^2$, adding $\|z_k - z^*\|^2$ to the Lyapunov with a small weight) all regenerate the same per-iteration constant, which would cap the stationarity guarantee at a nonzero floor. For $\alpha = 0$ the argument closes exactly (Lemma 4.4), because the unmatched term enters with a *negative* sign in the summed inequality and is dropped legitimately. Closing the inertial prox-center case therefore requires either a different estimator split or a different potential, we state it as an open problem.

Lemma 4.3 (Anchor inequality). *For every $k \geq 0$,*

$$2(\Lambda_{k+1} - \Lambda_k) + \|\Delta_{k+1}\|^2 \leq 2\eta\langle B_k + S_k, z_{k+1} - z^* \rangle. \quad (10)$$

Proof. The optimality conditions of the x - and y -subproblems give

$$\frac{x_{k+1} - \tilde{x}_k}{\eta} = -(\tilde{\nabla}_x H(\hat{x}_k, \bar{y}_k) + u), \quad \frac{y_{k+1} - \tilde{y}_k}{\eta} = \tilde{\nabla}_y H(\bar{x}_{k+1}, \hat{y}_k) - v. \quad (11)$$

Since $\alpha_i = 0$ we have $\tilde{x}_k = x_k$, $\tilde{y}_k = y_k$; testing the two identities in (11) against $x_{k+1} - x^*$ and $y_{k+1} - y^*$ and using the three-point identity $2\langle a - c, a - b \rangle = \|a - c\|^2 - \|b - c\|^2 + \|a - b\|^2$ with $a = x_{k+1}$, $b = x_k$, $c = x^*$ (and likewise for y) gives, after adding the two blocks and using (8),

$$2(\Lambda_{k+1} - \Lambda_k) + \|\Delta_{k+1}\|^2 = -2\eta\langle F(z_{k+1}), z_{k+1} - z^* \rangle + 2\eta\langle B_k + S_k, z_{k+1} - z^* \rangle,$$

with $F(z_{k+1}) = (\nabla_x H(z_{k+1}) + u, -\nabla_y H(z_{k+1}) + v)$. The first term on the right is nonpositive by the Minty condition (2) applied at $z = z_{k+1} \in \text{dom } \partial\Psi$; dropping it yields (10). $\square$

Lemma 4.4 (Master inequality). $\|B_k^x\| \leq \bar{M}(\|\Delta_{k+1}\| + \|\Delta_k\|)$ and $\|B_k^y\| \leq \bar{M}(2\|\Delta_{k+1}\| + \|\Delta_k\|)$, hence

$$\|B_k\| \leq 3\bar{M}(\|\Delta_{k+1}\| + \|\Delta_k\|). \quad (12)$$

Set

$$c := 3\eta\bar{M} \leq \frac{1}{100}, \quad a_0 := 1 - c, \quad c_0 := \frac{1}{64\Theta}, \quad (13)$$

$$\rho := \frac{1}{2} - c - \frac{\eta^2}{2c_0} - \frac{c_0 V_1}{2}, \quad c_k := c + \frac{c_0 V_1}{2}. \quad (14)$$

Then

$$a_0 \mathbb{E}_k \|z_{k+1} - z^*\|^2 - \|z_k - z^*\|^2 + \rho \mathbb{E}_k \|\Delta_{k+1}\|^2 \leq c_k \|\Delta_k\|^2 + \frac{c_0 V_1}{2} \|\Delta_{k-1}\|^2 + \frac{c_0}{2} \Gamma_k, \quad (15)$$

and, under Assumption 4.1, $\rho \geq \frac{2}{5}$ and $c_k \leq \frac{1}{20}$.

Proof. $B_k^x = \nabla_x H(z_{k+1}) - \nabla_x H(\hat{x}_k, \bar{y}_k)$ is controlled by the displacement $(x_{k+1} - \hat{x}_k, y_{k+1} - \bar{y}_k) = (\Delta_{k+1}^x - \beta_1^k \Delta_k^x, \Delta_{k+1}^y - \gamma_2^{k-1} \Delta_k^y)$, of norm at most $\|\Delta_{k+1}\| + \|\Delta_k\|$; similarly $B_k^y = \nabla_y H(z_{k+1}) - \nabla_y H(\bar{x}_{k+1}, \hat{y}_k)$ is controlled by $(\gamma_1^k \Delta_{k+1}^x, \Delta_{k+1}^y - \beta_2^k \Delta_k^y)$, of norm at most $2\|\Delta_{k+1}\| + \|\Delta_k\|$ (using $\beta, \gamma \leq 1$). Hence (12).

Take $\mathbb{E}_k$ in (10) and split

$$\langle B_k + S_k, z_{k+1} - z^* \rangle = \langle B_k, z_{k+1} - z^* \rangle + \langle S_k, \Delta_{k+1} \rangle + \langle S_k, z_k - z^* \rangle;$$

the last term has mean zero since $z_k - z^*$ is $\mathcal{F}_k$ -measurable and $\mathbb{E}_k[S_k^x] = \mathbb{E}_k[S_k^y] = 0$ (the tower property applied to the independent y -draw). By (12) and Young's inequality with parameter $c > 0$,

$$2\eta \langle B_k, z_{k+1} - z^* \rangle \leq c\|z_{k+1} - z^*\|^2 + \frac{\eta^2}{c} \|B_k\|^2 \leq c\|z_{k+1} - z^*\|^2 + 2c(\|\Delta_{k+1}\|^2 + \|\Delta_k\|^2),$$

because $\frac{\eta^2}{c} \|B_k\|^2 \leq \frac{\eta^2}{3\eta\bar{M}} \cdot 9\bar{M}^2(\|\Delta_{k+1}\| + \|\Delta_k\|)^2 = 3\eta\bar{M}(\|\Delta_{k+1}\| + \|\Delta_k\|)^2 \leq 2c(\|\Delta_{k+1}\|^2 + \|\Delta_k\|^2)$. Young's inequality with c_0 gives $2\eta \langle S_k, \Delta_{k+1} \rangle \leq c_0 \|S_k\|^2 + \frac{\eta^2}{c_0} \|\Delta_{k+1}\|^2$, and Definition 2.6(ii) bounds $\mathbb{E}_k \|S_k\|^2$. Since $\alpha_i = 0$, $w_k = \Delta_{k+1}$. Moving the term $c \mathbb{E}_k \|z_{k+1} - z^*\|^2$ to the left produces the factor $a_0 = 1 - c$ in (15); dividing by 2 and collecting gives the claim. The bounds on ρ, c_k follow from $\eta\bar{M} \leq \frac{1}{300}$ (so $c \leq \frac{1}{100}$), $\eta\sqrt{\Theta} \leq \frac{1}{256}$ (so $\frac{\eta^2}{2c_0} \leq \frac{1}{2048}$), $c_0\Theta = \frac{1}{64}$ and $c_0V_1 = \frac{3}{256N^2} \leq \frac{3}{256}$ (so $\frac{c_0V_1}{2} \leq \frac{3}{512}$):

$$\rho \geq \frac{1}{2} - \frac{1}{100} - \frac{1}{2048} - \frac{3}{512} \geq \frac{2}{5}, \quad c_k \leq \frac{1}{100} + \frac{3}{512} \leq \frac{1}{20}.$$

□

Lemma 4.5 (Lyapunov summation). *With*

$$q := \frac{c_0}{2\tau}, \quad p_2 := c_0(V_1 + \Theta), \quad p_1 := 2c_k + c_0V_1 + 2c_0\Theta, \quad (16)$$

$$\rho' := \rho - c_k - \frac{c_0V_1}{2} - \frac{3c_0\Theta}{2} \geq \frac{1}{4} > 0, \quad (17)$$

there exists $C_0 = \mathcal{O}(1 + c_0\Theta)$ such that for all $K \geq 1$ (with $\Delta_0 = 0$)

$$\rho' \sum_{k=1}^K \mathbb{E} \|\Delta_{k+1}\|^2 \leq C_0 D^2. \quad (18)$$

Proof. Sum (15) over $k = 1, \dots, K$. To the summed inequality add the telescoping identities

$$\sum_{k=1}^K (\mathbb{E} \|\Delta_{k+1}\|^2 - \|\Delta_k\|^2) = \mathbb{E} \|\Delta_{K+1}\|^2 - \mathbb{E} \|\Delta_1\|^2, \quad \sum_{k=1}^K (\|\Delta_k\|^2 - \|\Delta_{k-1}\|^2) = \|\Delta_K\|^2,$$

weighted by $\frac{p_1}{2}$ and $\frac{p_2}{2}$ respectively, and use Definition 2.6(iii) in the form $q(\mathbb{E} \Gamma_{k+1} - \Gamma_k) \leq -\frac{c_0}{2} \Gamma_k + \frac{c_0\Theta}{2} (\mathbb{E} \|\Delta_{k+1}\|^2 + \|\Delta_k\|^2 + \|\Delta_{k-1}\|^2)$. The choices (16) cancel the Γ_k -, $\|\Delta_{k-1}\|^2$ - and $\|\Delta_k\|^2$ -brackets exactly: the Γ_k -bracket is $\frac{c_0}{2} - q\tau = 0$; the $\|\Delta_{k-1}\|^2$ -bracket is $\frac{c_0V_1}{2} - \frac{p_2}{2} + qV_\Gamma = \frac{c_0V_1}{2} - \frac{c_0V_1}{2} - \frac{c_0\Theta}{2} + \frac{c_0\Theta}{2} = 0$; the $\|\Delta_k\|^2$ -bracket is $c_k - \frac{p_1}{2} + \frac{p_2}{2} + qV_\Gamma = c_k - c_k - \frac{c_0V_1}{2} - c_0\Theta + \frac{c_0V_1}{2} + \frac{c_0\Theta}{2} + \frac{c_0\Theta}{2} = 0$, where $\frac{p_2}{2} = \frac{c_0V_1}{2} + \frac{c_0\Theta}{2}$ and $qV_\Gamma = \frac{c_0\Theta}{2}$. The remaining coefficient of $\sum_k \mathbb{E} \|\Delta_{k+1}\|^2$ is $\rho - (c_k + \frac{c_0V_1}{2} + \frac{3c_0\Theta}{2}) = \rho'$, the three Θ -terms coming from qV_Γ acting on $\mathbb{E} \|\Delta_{k+1}\|^2, \|\Delta_k\|^2, \|\Delta_{k-1}\|^2$.

It remains to control the distance terms. The summed distance contribution is

$$\sum_{k=1}^K (a_0 \mathbb{E} \|z_{k+1} - z^*\|^2 - \|z_k - z^*\|^2) \geq -c \sum_{k=1}^K \mathbb{E} \|z_{k+1} - z^*\|^2 - \mathbb{E} \|z_1 - z^*\|^2 \geq -\mathbb{E} \|z_1 - z^*\|^2,$$

where the nonpositive sum $-c \sum_k \mathbb{E} \|z_{k+1} - z^*\|^2$ has been dropped from the left-hand side (if $A + P \leq B$ with $P \geq 0$ then $A \leq B$). The terms $\mathbb{E} \|\Delta_{K+1}\|^2$, $\|\Delta_K\|^2$ and $\mathbb{E} \Gamma_{K+1}$ are nonnegative and stay on the left, while the one-time terms $\mathbb{E} \|z_1 - z^*\|^2 \leq D^2$, $\frac{p_1}{2} \mathbb{E} \|\Delta_1\|^2 \leq 2p_1 D^2$ and $q \mathbb{E} \Gamma_1 \leq 2c_0 \Theta D^2$ (using $\Gamma_1 \leq V_\Gamma \mathbb{E} \|\Delta_1\|^2$ when $\Gamma_0 = 0$, and $\mathbb{E} \|\Delta_1\|^2 \leq 4D^2$) are absorbed into $C_0 D^2$ with $C_0 = 1 + 2p_1 + 2c_0 \Theta = \mathcal{O}(1 + c_0 \Theta)$. Numerically, $\rho' \geq \frac{2}{5} - \frac{1}{20} - \frac{3}{512} - \frac{3}{128} \geq \frac{1}{4}$. $\square$

Corollary 4.6 (Summability). *For all $K \geq 1$ (with $\Delta_0 = 0$),*

$$\frac{1}{K} \sum_{k=1}^K \mathbb{E} \|\Delta_{k+1}\|^2 \leq \frac{C_0 D^2}{\rho' K}, \quad \frac{1}{K} \sum_{k=1}^K \mathbb{E} \Gamma_{k+1} \leq \frac{\Gamma_1}{\tau K} + \frac{3\Theta C_0 D^2}{\rho' K}. \quad (19)$$

Proof. The first bound is (18) divided by K . For the second, unroll Definition 2.6(iii):

$$\mathbb{E} \Gamma_{k+1} \leq (1 - \tau)^{k-1} \Gamma_1 + V_\Gamma \sum_{j=1}^{k-1} (1 - \tau)^{k-1-j} (\mathbb{E} \|\Delta_{j+1}\|^2 + \mathbb{E} \|\Delta_j\|^2 + \mathbb{E} \|\Delta_{j-1}\|^2),$$

sum over $k = 1, \dots, K$, use $\sum_{k>j} (1 - \tau)^{k-1-j} \leq \frac{1}{\tau}$ and the fact that each $\mathbb{E} \|\Delta_j\|^2$ appears for at most three values of j :

$$\tau \sum_{k=1}^K \mathbb{E} \Gamma_{k+1} \leq \tau \Gamma_1 + 3V_\Gamma \sum_{j=1}^K \mathbb{E} \|\Delta_{j+1}\|^2.$$

Divide by τK and apply the first bound of (19). $\square$

Lemma 4.7 (Subgradient residual bound). *There exists $\mu_0 = \mathcal{O}(\eta^{-2} + \bar{M}^2 + V_1)$ such that for all $k \geq 1$,*

$$\mathbb{E} [\text{dist}(0, \partial \Psi(z_{k+1}))^2] \leq \mu_0 \mathbb{E} [\|\Delta_{k+1}\|^2 + \|\Delta_k\|^2 + \|\Delta_{k-1}\|^2] + 3 \mathbb{E} \Gamma_k. \quad (20)$$

Proof. From the x -proximal optimality condition, $u_{k+1} := -\tilde{\nabla}_x H(\hat{x}_k, \bar{y}_k) - \frac{1}{\eta}(x_{k+1} - \tilde{x}_k) \in \partial f(x_{k+1})$, so

$$\nabla_x H(z_{k+1}) + u_{k+1} = [\nabla_x H(z_{k+1}) - \nabla_x H(\hat{x}_k, \bar{y}_k)] + [\nabla_x H(\hat{x}_k, \bar{y}_k) - \tilde{\nabla}_x H(\hat{x}_k, \bar{y}_k)] - \frac{1}{\eta}(x_{k+1} - \tilde{x}_k).$$

Since $\alpha_i = 0$ we have $x_{k+1} - \tilde{x}_k = \Delta_{k+1}^x$, and the query displacement has norm at most $\sqrt{2}(\|\Delta_{k+1}\| + \|\Delta_k\|)$, so

$$\left\| \frac{1}{\eta}(x_{k+1} - \tilde{x}_k) \right\| \leq \frac{1}{\eta} \|\Delta_{k+1}\|, \quad \|\nabla_x H(z_{k+1}) - \nabla_x H(\hat{x}_k, \bar{y}_k)\| \leq \sqrt{2} \bar{M} (\|\Delta_{k+1}\| + \|\Delta_k\|).$$

The y -block is handled identically (its query point differs from z_{k+1} by a vector of norm at most $\sqrt{2}(\|\Delta_{k+1}\| + \|\Delta_k\|)$). Applying $(a + b + c)^2 \leq 3(a^2 + b^2 + c^2)$ to the concatenated residual and then Definition 2.6(ii) gives (20). $\square$

Theorem 4.8 (Complexity with the SAGA estimator). *Suppose Assumptions 2.1, 2.2, and 4.1 hold, and $\tilde{\nabla} H$ is the SAGA estimator of Lemma 2.7 (table initialized at $z_0 = z_{-1}$, so that $\Gamma_0 = 0$). Then*

$$\min_{2 \leq k \leq K+2} \mathbb{E} [\text{dist}(0, \partial \Psi(z_k))^2] \leq \frac{C_{\text{complex}}}{K},$$

where $C_{\text{complex}} = \mathcal{O}((\eta^{-2} + \bar{M}^2 + \Theta)D^2)$. Consequently, reaching $\mathbb{E} [\text{dist}(0, \partial \Psi)^2] \leq \varepsilon^2$ costs at most

$$N + \mathcal{O}\left(\frac{N^2 L^2 D^2}{\varepsilon^2}\right) \quad (21)$$

single-component gradient evaluations.

Proof. Averaging (20) over $k = 1, \dots, K$ and applying Corollary 4.6,

$$\min_{2 \leq k \leq K+2} \mathbb{E}[\text{dist}(0, \partial\Psi(z_k))^2] \leq \frac{9\mu_0 C_0 D^2}{\rho' K} + \frac{9\Theta C_0 D^2}{\rho' K} + \frac{3\Gamma_1}{\tau K}.$$

For the SAGA table initialized at $z_0 = z_{-1}$ one has $\Gamma_0 = 0$, hence $\Gamma_1 \leq V_\Gamma \mathbb{E} \|\Delta_1\|^2 \leq 4V_\Gamma D^2$ and $\frac{3\Gamma_1}{\tau K} \leq \frac{12\Theta D^2}{K}$. By Lemma 2.7, $\Theta = 32N^2 L^2$, so $\eta^{-1} = O(NL)$ by (9), $\mu_0 = \mathcal{O}(N^2 L^2)$, and both remaining terms are $\mathcal{O}(N^2 L^2 D^2 / K)$. Hence $K = \mathcal{O}(N^2 L^2 D^2 / \varepsilon^2)$ iterations suffice, each costing two evaluations (Definition 2.5), plus N initialization evaluations. $\square$

Remark 4.9 (The N^2 scaling is optimal for single-loop unbiased VR). The class contains nonconvex *minimization* (H independent of y , $g \equiv 0$), for which single-loop SAGA achieves $N + \mathcal{O}(N^{2/3} L^2 \Delta / \varepsilon^2)$ via function-value anchoring [33]; hence no lower bound above $N^{2/3}$ can hold for the algorithm class, and the gap of Theorem 4.8 is analytical, not informational. Appendix A shows that within anchor-type analyses, any unbiased table-based estimator refreshing b entries per iteration contracts at rate at most $\tau \leq b/(N-1)$ (Theorem A.2), with $V_\Gamma = \Omega(L^2/\tau)$ (Corollary B.2); hence the total cost of any single-loop unbiased method analyzed through (15) is $\Omega(N^2 L^2 D^2 / (b\varepsilon^2)) - \Omega(N^2 L^2 D^2 / \varepsilon^2)$ at unit cost. Theorem 4.8 matches this lower bound up to absolute constants: single-loop unbiased VR under Minty MVI is exactly N^2 , and epoch-based biased VR (Proposition 4.10) is exactly $\sqrt{N}$.

4.1 Epoch SARAH variant

SARAH does not satisfy Definition 2.6 at the per-iteration level: within an epoch its estimation error accumulates over all previous inner steps and its estimator is biased ($\mathbb{E}[S_t | \mathcal{F}_{t-1}] = S_{t-1} \neq 0$). We analyze it at the epoch level.

Proposition 4.10 (Complexity of the epoch SARAH variant). *Consider Algorithm 1 with $\alpha_i^k = \beta_i^k = \gamma_i^k = 0$, run in epochs of length $m = \lceil \sqrt{N} \rceil$ with minibatch size $b = \lceil \sqrt{N} \rceil$; epochs are chained by $z_{k+1,0} := z_{k,m}$. Each epoch begins with a full gradient $v_0 = \nabla H(z_{k,0})$ and inner step t uses*

$$v_t = v_{t-1} + \frac{1}{b} \sum_{i \in I_t} (\nabla H_i(z_{k,t-1}) - \nabla H_i(z_{k,t-2})),$$

producing $z_{k,t}$ from $z_{k,t-1}$. Let $\delta_{k,t} := z_{k,t} - z_{k,t-1}$ and suppose

$$\eta \leq \min \left\{ \frac{1}{64L^2}, \frac{1}{12M}, \frac{1}{32} \right\}. \quad (22)$$

Then for every $K \geq 1$,

$$\sum_{t=1}^m \mathbb{E} \|\delta_{k,t}\|^2 \leq 3 \left(\mathbb{E} \|z_{k,0} - z^*\|^2 - \mathbb{E} \|z_{k,m} - z^*\|^2 \right) \quad \text{for every } k, \quad (23)$$

$$\sum_{k=1}^K \sum_{t=1}^m \mathbb{E} \|\delta_{k,t}\|^2 \leq 3D^2, \quad (24)$$

and the total evaluation cost to reach $\mathbb{E}[\text{dist}(0, \partial\Psi(z_{k,\tau}))^2] \leq \varepsilon^2$ at a uniformly random inner step (k, τ) is

$$\mathcal{O}\left(\frac{\sqrt{N} L^4 D^2}{\varepsilon^2}\right), \quad (25)$$

the optimal $\sqrt{N}$ scaling up to problem-dependent constants.

Proof. Step 1 (per-inner-step contraction). With zero inertia the anchor gives, for $\Lambda(z) = \frac{1}{2}\|z - z^*\|^2$,

$$\mathbb{E}[\Lambda(z_{k,t}) - \Lambda(z_{k,t-1})] + \frac{1}{2} \mathbb{E} \|\delta_{k,t}\|^2 \leq \eta \mathbb{E} \langle B_t + S_t, \delta_{k,t} \rangle + \eta \mathbb{E} \langle B_t + S_t, z_{k,t-1} - z^* \rangle,$$

where $S_t = \nabla H(z_{k,t-1}) - v_t$ and, accounting for the y -query at the fresh point $(x_{k,t}, y_{k,t-1})$ and the estimator reference $z_{k,t-1}$, $B_t = (\nabla_x H(z_{k,t}) - \nabla_x H(z_{k,t-1}), -[\nabla_y H(z_{k,t}) - \nabla_y H(z_{k,t-1})] - [\nabla_y H(x_{k,t}, y_{k,t-1}) - \nabla_y H(z_{k,t})])$ satisfies $\|B_t\| \leq 3\bar{M}\|\delta_{k,t}\|$. No unbiasedness is used. By Young's inequality and $\|B_t\| \leq 3\bar{M}\|\delta_{k,t}\|$,

$$\eta \langle B_t, z_{k,t-1} - z^* \rangle \leq \frac{3\eta\bar{M}}{2} (\|\delta_{k,t}\|^2 + \|z_{k,t-1} - z^*\|^2),$$

which is absorbed by the $-\frac{1}{2}\|z_{k,t-1} - z^*\|^2$ on the left (as $3\eta\bar{M} \leq \frac{1}{4} \leq \frac{1}{2}$, since $\eta\bar{M} \leq \frac{1}{12}$); $\eta \langle B_t, \delta \rangle \leq 3\eta\bar{M}\|\delta\|^2$; Young with parameter $1/4$ gives $\eta \langle S_t, \delta \rangle \leq 2\eta \mathbb{E} \|S_t\|^2 + \frac{\eta}{8}\|\delta\|^2$ and $\eta \langle S_t, z_{k,t-1} - z^* \rangle \leq 2\eta \mathbb{E} \|S_t\|^2 + \frac{\eta}{8}\|z_{k,t-1} - z^*\|^2$ (the latter absorbed). Using $\mathbb{E} \|S_t\|^2 \leq \frac{L^2}{b} \sum_{s \leq t} \mathbb{E} \|\delta_{k,s}\|^2$ [29],

$$\frac{1}{2} (\mathbb{E} \|z_{k,m} - z^*\|^2 - \mathbb{E} \|z_{k,0} - z^*\|^2) + \left(\frac{1}{2} - 3\eta\bar{M} - \frac{\eta}{8} - \frac{4\eta mL^2}{b} \right) \sum_{t=1}^m \mathbb{E} \|\delta_{k,t}\|^2 \leq 0.$$

By (22) the bracket is $\geq \frac{1}{2} - \frac{1}{4} - \frac{1}{256} - \frac{1}{16} = \frac{47}{256} \geq \frac{1}{6}$ (using $3\eta\bar{M} \leq \frac{1}{4}$, $\eta \leq \frac{1}{32}$, and $m = b$ gives $4\eta L^2 \leq \frac{1}{16}$); this proves (23), and summing over the chained epochs telescopes the distances and gives (24) with constant 3. In particular, for every epoch k and every $t \leq m$,

$$\mathbb{E} \|S_t\|^2 \leq \frac{L^2}{b} \sum_{s \leq t} \mathbb{E} \|\delta_{k,s}\|^2 \leq \frac{3L^2}{b} (\mathbb{E} \|z_{k,0} - z^*\|^2 - \mathbb{E} \|z_{k,m} - z^*\|^2) \leq \frac{3D^2L^2}{b}. \quad (26)$$

Step 2 (residual at a random inner step). Lemma 4.7 (zero inertia, no variance term beyond $3\mathbb{E} \|S_t\|^2$) gives, for a uniform (k, τ) ,

$$\begin{aligned} \mathbb{E} \text{dist}(0, \partial\Psi(z_{k,\tau}))^2 &\leq \left(\frac{c}{\eta^2} + \bar{M}^2 \right) \frac{1}{mK} \sum_{k,t} \mathbb{E} \|\delta_{k,t}\|^2 + \frac{3}{mK} \sum_{k,t} \mathbb{E} \|S_t\|^2 \\ &\leq \frac{3D^2}{mK} \left(\frac{c}{\eta^2} + \bar{M}^2 + \frac{3L^2}{b} \right), \end{aligned}$$

using (24) and (26). With $m = b = \lceil \sqrt{N} \rceil$ and $\eta^{-1} = \mathcal{O}(L^2)$ this is $\mathcal{O}(D^2L^4/(\sqrt{N}K))$, so $K = \mathcal{O}(D^2L^4/(\sqrt{N}\varepsilon^2))$ epochs suffice; each costs $N + mb = 2N$ component evaluations (Definition 2.5), giving (25). $\square$

4.2 Lower bounds for the Minty class, and the status of $\sqrt{N}$

Proposition 4.11 (Lower bounds for the Minty class via embedding). *Every lower bound proved on convex finite sums applies to the class of Definition 1.1 through Lemma D.1. Quantitatively: for any randomized algorithm, any $N \geq 2$, and any $\varepsilon \lesssim LB$, there exist a dimension $d = \mathcal{O}(\frac{L^2B^2}{\varepsilon^2} \log \frac{NLB}{\varepsilon})$ and a Minty instance with domain $\{\|x\| \leq B\}$, L -smooth components and L -smooth sum, on which finding an ε -stationary point requires*

$$\Omega\left(N + \sqrt{\frac{NLB}{\varepsilon}}\right) \quad (27)$$

component gradient evaluations; on tight instances $\Delta = \Theta(LB^2)$ this reads $\Omega(N + \sqrt{N}(L\Delta)^{1/4}\varepsilon^{-1/2})$. For $N = 1$ the convex-stationarity complexity is $\tilde{\Theta}(\sqrt{L\Delta}/\varepsilon)$, with accelerated gradient descent

optimal [11, 12]. Two facts should be kept in mind. (i) Convex instances cannot witness Conjecture D.6: (27) is the strongest bound convex instances allow, since ε -stationarity of a convex L -smooth function implies εB -optimality (Remark D.4). (ii) The $\sqrt{N} L \Delta / \varepsilon^2$ bound of [3] is not available this way: its hard instance is nonconvex and not known to be Minty. The ε^{-2} scaling is nevertheless class-optimal at $N = 1$: Theorem E.3 constructs an explicit star-monotone nonconvex instance requiring $\Omega(L \Delta / \varepsilon^2)$ evaluations (zero-respecting; randomized in dimension $c(L \Delta / \varepsilon^2)^2$).

For convex Minty instances the picture is complete: the lower bound (27) is matched by accelerated variance reduction, which achieves ε -optimality in $O(N + \sqrt{N L R^2 / \epsilon})$ evaluations [38], so the convex-stationarity complexity is $\Theta(N + \sqrt{N L R^2 / \epsilon})$ (up to the dimension requirement). The complexity landscape of the Minty class is therefore: convex instances are settled at $\varepsilon^{-1/2}$ -type scaling; the $N = 1$ nonconvex case is settled at ε^{-2} (Theorem E.3); and the jointly-optimal $\sqrt{N} L \Delta / \varepsilon^2$ corner is Conjecture D.6.

Proof. Lemma D.1 shows the class contains every convex finite-sum instance, so lower bounds for convex sums transfer verbatim. [38, Theorem 7] construct convex γ -smooth components on $\{\|x\| \leq B\}$ such that, in dimension $\mathcal{O}(\frac{\gamma^2 B^4}{\epsilon^2} \log \frac{N \gamma B^2}{\epsilon})$, any randomized algorithm needs $\Omega(N + \sqrt{N \gamma B^2 / \epsilon})$ queries for ϵ -optimality. An ε -stationary point of a convex L -smooth f on $\{\|x\| \leq B\}$ is εB -optimal, since $f(x) - f^* \leq \langle \nabla f(x), x - x^* \rangle \leq \varepsilon B$; applying the lower bound with $\epsilon = \varepsilon B$ and $\gamma = L$ gives (27) and the stated dimension. The tight-regime form uses $B = \Theta(\sqrt{\Delta / L})$; the $N = 1$ statement is the lower bound of [11, Section 3] together with the optimality of accelerated gradient descent [12]. $\square$

5 Linear convergence under strong monotonicity

Two tiers. The inertial theory under strong monotonicity has two tiers, which we state before the results. For *general* strongly monotone problems, only small inertia is certified ($\alpha \leq \eta \mu / 50$, Theorem C.1 in Appendix C); this is the content of Theorem 5.3 for the non-inertial variant. For *quadratic games with small coupling* — the class on which the numerical experiments run — the full practical range of inertia is certified, with acceleration (Theorem 5.6). The intermediate regime, general strongly monotone problems at practical inertia, remains open (Remark 5.8).

Definition 5.1 (Strong monotonicity). The operator $F = (\nabla_x H + \partial f, -\nabla_y H + \partial g)$ is μ -strongly monotone, $\mu > 0$, and $\bar{M}$ -Lipschitz on $\text{dom } \partial \Psi$: for all z, z' and all $\zeta \in F(z), \zeta' \in F(z')$, $\langle \zeta - \zeta', z - z' \rangle \geq \mu \|z - z'\|^2$ and $\|\zeta - \zeta'\| \leq \bar{M} \|z - z'\|$.

Remark 5.2 (SC–SC and the two-sided PL condition). If $H(\cdot, y)$ is μ_x -strongly convex, $H(x, \cdot)$ is μ_y -strongly concave, and ∇H is $\bar{M}$ -Lipschitz, then Definition 5.1 holds with $\mu = \min\{\mu_x, \mu_y\}$: the mixed terms cancel by symmetry of the cross-Hessian. The Minty point z^* is in fact the unique zero of F , as required by the phrase “choosing the zero element of $F(z^*)$ ” in Section 5. The argument proceeds in three steps. First, the Minty condition (2) together with the $\bar{M}$ -Lipschitz property of Definition 5.1 implies that z^* solves the Stampacchia variational inequality, $\langle \zeta^*, z - z^* \rangle \geq 0$ for all $z \in \text{dom } \partial \Psi$ and every selection $\zeta^* \in F(z^*)$: for $z_t = z^* + t(z - z^*)$ the Minty condition gives $\langle \zeta_t, z_t - z^* \rangle \geq 0$ with $\zeta_t \in F(z_t)$, and the Lipschitz bound gives $|\langle \zeta_t - \zeta^*, z - z^* \rangle| \leq \bar{M} t \|z - z^*\|^2$; dividing by t and letting $t \downarrow 0$ yields the claim. Second, since $\text{dom } \partial \Psi$ is convex, $z = z^* - t \zeta^*$ is admissible for small $t > 0$, and $\langle \zeta^*, z - z^* \rangle = -t \|\zeta^*\|^2 \geq 0$ forces $\zeta^* = 0$; hence $0 \in F(z^*)$. Third, strong monotonicity gives uniqueness: if $0 \in F(z_1) \cap F(z_2)$ then $\mu \|z_1 - z_2\|^2 \leq \langle 0 - 0, z_1 - z_2 \rangle = 0$. Moreover the *two-sided PL* condition holds with $\nu = \frac{1}{2\mu}$: for $P(z) = \Psi(z) - \Psi^*$, $P(z) \leq \frac{1}{2\mu_x} \text{dist}(0, \partial_x \Psi(z))^2$ and $-P(z) \leq \frac{1}{2\mu_y} \text{dist}(0, \partial_y \Psi(z))^2$ by the PL inequality for strongly convex minimization applied blockwise, hence $|P(z)| \leq \frac{1}{2\mu} \text{dist}(0, \partial \Psi(z))^2$. Note that P is sign-indefinite, which is precisely why a descent argument on Ψ cannot work and the distance-based argument below is needed.

Theorem 5.3 (*R-linear convergence*). Consider Algorithm 1 with exact gradients and $\alpha_i^k = \beta_i^k = \gamma_i^k = 0$. Under Assumptions 2.1 and Definition 5.1, for $\eta \leq \frac{\mu}{16\bar{M}^2}$ and all $k \geq 0$,

$$\|z_k - z^*\|^2 \leq (1 + \eta\mu)^{-k} \|z_0 - z^*\|^2. \quad (28)$$

Moreover, with $L_\Phi := \frac{1}{\eta} + \bar{M}$,

$$\text{dist}(0, \partial\Psi(z_k))^2 \leq 8L_\Phi^2(1 + \eta\mu)^{-(k-1)} \|z_0 - z^*\|^2, \quad (29)$$

and if the two-sided PL condition holds with constant ν ,

$$|\Psi(z_k) - \Psi^*| \leq 8\nu L_\Phi^2(1 + \eta\mu)^{-(k-1)} \|z_0 - z^*\|^2. \quad (30)$$

Proof. Write $u_k = \|z_k - z^*\|^2$. The proximal first-order conditions give, with $u \in \partial f(x_{k+1})$, $v \in \partial g(y_{k+1})$,

$$\frac{1}{\eta} \Delta_{k+1}^x = -(\nabla_x H(z_k) + u) = -(\nabla_x H(z_{k+1}) + u) + B_k^x,$$

$$B_k^x = \nabla_x H(z_{k+1}) - \nabla_x H(z_k), \quad \|B_k^x\| \leq \bar{M} \|\Delta_{k+1}\|,$$

$$\frac{1}{\eta} \Delta_{k+1}^y = \nabla_y H(x_{k+1}, y_k) - v = (\nabla_y H(z_{k+1}) - v) - B_k^y,$$

$$B_k^y = \nabla_y H(z_{k+1}) - \nabla_y H(x_{k+1}, y_k), \quad \|B_k^y\| \leq \bar{M} \|\Delta_{k+1}\|.$$

Set $B_k = (B_k^x, -B_k^y)$, $\|B_k\| \leq 2\bar{M} \|\Delta_{k+1}\|$. The three-point identity yields

$$u_{k+1} - u_k + \|\Delta_{k+1}\|^2 = -2\eta \langle F(z_{k+1}), z_{k+1} - z^* \rangle + 2\eta \langle B_k, z_{k+1} - z^* \rangle. \quad (31)$$

Strong monotonicity applied to z_{k+1} and z^* (choosing the zero element of $F(z^*)$, see Remark 5.2) gives $\langle F(z_{k+1}), z_{k+1} - z^* \rangle \geq \mu u_{k+1}$. Young's inequality ($ab \leq \frac{a^2}{2t} + \frac{tb^2}{2}$ with $t = 2$) gives $2\eta \langle B_k, z_{k+1} - z^* \rangle \leq 4\eta \bar{M} \|\Delta_{k+1}\| \sqrt{u_{k+1}} \leq \frac{1}{4} \|\Delta_{k+1}\|^2 + 16\eta^2 \bar{M}^2 u_{k+1}$. Substituting into (31) and rearranging,

$$u_{k+1} - u_k + \frac{3}{4} \|\Delta_{k+1}\|^2 \leq (-2\eta\mu + 16\eta^2 \bar{M}^2) u_{k+1} \leq -\eta\mu u_{k+1},$$

since $\eta \leq \frac{\mu}{16\bar{M}^2}$. (we drop the nonnegative term $\frac{3}{4} \|\Delta_{k+1}\|^2$ from the left-hand side). Hence $(1 + \eta\mu)u_{k+1} \leq u_k$, which is (28). For (29): the residual vector built from the first-order subgradients satisfies $\|w_k\| \leq (\frac{1}{\eta} + \bar{M}) \|\Delta_k\|$ blockwise (the argument of Lemma 4.7 with $S = 0$, $\alpha = 0$), so $\text{dist}(0, \partial\Psi(z_k))^2 \leq 2L_\Phi^2 \|\Delta_k\|^2 \leq 2L_\Phi^2(2u_k + 2u_{k-1}) \leq 8L_\Phi^2(1 + \eta\mu)^{-(k-1)} u_0$. Finally (30) is the two-sided PL inequality applied to (29). $\square$

Remark 5.4 (Rate and regime). With $\eta = \frac{\mu}{16\bar{M}^2}$ the per-step factor is $(1 + \frac{\mu}{16\bar{M}^2})^{-1} = 1 - \Theta(\kappa^{-2})$, $\kappa = \bar{M}/\mu$ — the canonical dependence for descent-ascent-type schemes (extragradient improves this to κ^{-1} at the price of two operator evaluations per step); the rate itself is classical — the contribution is the proximal block-wise certificate with explicit constants under composite regularizers. The constant $\frac{1}{16}$ is not optimized.

A stochastic linear rate is also available under strong monotonicity, for the epoch SARAH variant, at the price of a fourth-power condition-number dependence.

Proposition 5.5 (Stochastic linear rate under strong monotonicity). Let F be μ -strongly monotone and $\bar{M}$ -Lipschitz, let each H_i be L -smooth, and set $L' := \max\{L, \bar{M}\}$. Run the epoch SARAH variant of Proposition 4.10 with zero inertia, epoch length m and minibatch $b = m$, where

$$\eta \leq \frac{\mu}{1024 L'^2}, \quad m = \left\lceil \frac{4 \ln 2}{\eta\mu} \right\rceil. \quad (32)$$

Then, with $z_{k,m}$ the last iterate of the k -th epoch,

$$\mathbb{E} \|z_{k,m} - z^*\|^2 \leq \left(\frac{7}{8}\right)^k \|z_{0,0} - z^*\|^2. \quad (33)$$

In particular $\mathbb{E} \|z_{k,m} - z^*\|^2 \leq \varepsilon^2$ after $k = \mathcal{O}(\log(D/\varepsilon))$ epochs, i.e. $\mathcal{O}((N + L^4/\mu^4) \log(D/\varepsilon))$ evaluations.

Proof. Within one epoch write $u_t = \|z_{k,t} - z^*\|^2$, $\delta_t = z_{k,t} - z_{k,t-1}$, $S_t = \nabla H(z_{k,t-1}) - v_t$. The strong-monotonicity anchor (the argument of Theorem 5.3, with the biased error S_t treated by Young instead of unbiasedness, and $\|B_t\| \leq 3\bar{M}\|\delta_t\|$ from Proposition 4.10) gives, using $\eta \leq \mu/(36\bar{M}^2)$ and $\eta\mu \leq 1$,

$$\left(1 + \frac{3\eta\mu}{2}\right)u_{t+1} + \frac{1}{4}\|\delta_{t+1}\|^2 \leq u_t + \frac{\eta\mu}{4}u_{t-1} + \frac{8\eta}{\mu}\|S_t\|^2. \quad (34)$$

Summing (34) over $t = 1, \dots, m$ with weights r^{m-t} , $r := \frac{1+\eta\mu/2}{1+3\eta\mu/2} \leq 1 - \frac{\eta\mu}{4}$ (as $(1 - \frac{x}{4})(1 + \frac{3x}{2}) = 1 + \frac{5x}{8} - \frac{3x^2}{8} \geq 1 + \frac{x}{2}$ for $x = \eta\mu \leq \frac{1}{16}$), telescopes: the coefficient of each u_s on the left exceeds that on the right by a factor $1 + \Omega(\eta\mu)$, so

$$u_{k,m} \leq r^m u_{k,0} + \frac{16\eta}{\mu} \sum_{t=1}^m \mathbb{E} \|S_t\|^2. \quad (35)$$

Next, $\mathbb{E} \|S_t\|^2 \leq \frac{L^2}{b} \sum_{s \leq t} \mathbb{E} \|\delta_s\|^2$ [29] and summing (34) without weights, using $m\eta\mu \leq 4 \ln 2 + 1$ and $u_t \leq 2u_{k,0} + \frac{8\eta}{\mu} \sum_s \mathbb{E} \|S_s\|^2$ (a crude two-term recursion bound), gives $\sum_t \mathbb{E} \|\delta_t\|^2 \leq 12u_{k,0} + \frac{64\eta}{\mu} \sum_t \mathbb{E} \|S_t\|^2$. Hence, with $b = m$ and $\eta \leq \mu/(128L^2)$ (implied by (32)),

$$\sum_{t=1}^m \mathbb{E} \|S_t\|^2 \leq L^2 \left[12u_{k,0} + \frac{64\eta}{\mu} \sum_t \mathbb{E} \|S_t\|^2 \right] \implies \sum_{t=1}^m \mathbb{E} \|S_t\|^2 \leq 24L^2 u_{k,0}.$$

Substituting into (35),

$$u_{k,m} \leq \left(r^m + \frac{384\eta L^2}{\mu}\right)u_{k,0} \leq \left(\frac{1}{2} + \frac{384}{1024}\right)u_{k,0} \leq \frac{7}{8} u_{k,0},$$

using $m = \lceil 4 \ln 2 / (\eta\mu) \rceil$ for $r^m \leq e^{-\eta\mu m/4} \leq \frac{1}{2}$ and $\eta L^2/\mu \leq 1/1024$. Iterating over epochs gives (33); the cost statement follows since each epoch uses $N + mb = N + m^2$ evaluations and $m = \mathcal{O}(L^2/\mu^2)$. $\square$

For quadratic games — the class on which the numerical experiments of Section 7 are run — the inertial certificate covers the *full practical range* of inertia parameters and certifies acceleration, via an exact spectral analysis of the linear recursion. Since the resolvent bounds below degenerate as the corner $\alpha = 0$, $\eta\lambda = 2$ is approached, we state the uniformity over compact subsets of the open parameter set.

Theorem 5.6 (Inertia on quadratic games: the practical range). *Let $\Psi(x, y) = \frac{1}{2}x^\top Ax + x^\top By - \frac{1}{2}y^\top Cy$ with $A \succeq \mu I$, $C \succeq \mu I$, $\max\{\|A\|, \|C\|\} \leq L_0$, and $f \equiv g \equiv 0$. Consider Algorithm 1 with $\beta = \gamma = 0$, $\alpha_1 = \alpha_2 = \alpha$, exact gradients, and $\eta \leq 2/L_0$. Then:*

- (i) **Decoupled case** $B = 0$. *For every $\alpha \in (0, 1)$ the iterates converge R -linearly: in the eigenbases of A and C the error satisfies a scalar recursion with characteristic polynomial $r^2 - (1 + \alpha - \eta\lambda)r + \alpha = 0$, $\lambda \in \text{spec}(A) \cup \text{spec}(C)$, whose two roots lie strictly inside the unit disk by the Jury conditions $1 - (1 + \alpha - \eta\lambda) + \alpha = \eta\lambda > 0$, $1 + (1 + \alpha - \eta\lambda) + \alpha = 2 + 2\alpha - \eta\lambda > 0$ (using $\eta\lambda \leq \eta L_0 \leq 2$) together with $\alpha < 1$. The rate is $\max_\lambda |r(\lambda)| < 1$, equal to $\sqrt{\alpha}$ in the complex-root region.*

(ii) **Small coupling.** For $B \neq 0$ the joint recursion $\xi_{k+1} = T(B)\xi_k$, with ξ_k stacking (e_k, e_{k-1}) , $e_k = z_k - z^*$, satisfies $\|T(B) - T(0)\| \leq 3\eta\|B\|$. Let $\rho_0 = \max_\lambda |r(\lambda)|$ from (i). Then the perturbative conclusion — R -linear convergence with rate at most $(1 + \rho_0)/2$ for $\|B\| \leq \epsilon_B$ — holds with constants ϵ_B that are uniform on every compact subset of the open parameter set $\mathcal{Q} := \{(\alpha, \eta\lambda) : \alpha \in (0, 1), 0 < \eta\lambda < 2(1 + \alpha)\}$; explicit values for the experimental instance are computed in Remark 5.7. The statement is not uniform up to the boundary point $\alpha = 0, \eta\lambda = 2$, where $\rho_0 \uparrow 1$ and the perturbation window closes; the tested instance is certified separately by exact spectral computation (Remark 5.7).

(iii) **Certified acceleration (each under its own optimal tuning).** The non-inertial variant, optimally tuned ($\alpha = 0, \eta = 2/(L_0 + \mu)$), has rate $1 - \Theta(\kappa^{-1})$; the inertial variant, optimally tuned (the classical heavy-ball choice of α and η [31]), has rate $\sqrt{\alpha^*} = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} = 1 - \Theta(\kappa^{-1/2})$, $\kappa = L_0/\mu$. In particular $\alpha = 0.05$, used in Section 7, lies in the certified range on the tested instances.

Proof. (i) The recursion for an eigencomponent with curvature λ is $e_{k+1} = (1 + \alpha - \eta\lambda)e_k - \alpha e_{k-1}$, with characteristic polynomial as stated; the three Jury conditions for $r^2 + a_1r + a_2$ are $|a_2| < 1$, $1 + a_1 + a_2 > 0$, $1 - a_1 + a_2 > 0$, i.e. $\alpha < 1$, $\eta\lambda > 0$, and $\eta\lambda < 2(1 + \alpha)$, the last holding since $\eta\lambda \leq 2$. Hence $\max(|r_+|, |r_-|) < 1$; for complex conjugate roots $|r_\pm| = \sqrt{\text{product}} = \sqrt{\alpha}$. (ii) The fresh- x coupling enters only through the terms $-\eta B e_k^y$ and $+\eta B^\top e_{k+1}^x$, so $T(B) = T(0) + \eta E(B)$ with $\|E(B)\| \leq 3\|B\|$ (using $\alpha < 1$). For the uniformity we use compactness: on every compact $\mathcal{Q}' \subset \mathcal{Q}$ the spectral radius ρ_0 of $T(0)$ is a continuous function bounded away from 1 (at the discriminant locus the double root equals $\sqrt{\alpha} \leq \sqrt{1 - \delta} < 1$ for $\alpha \leq 1 - \delta$), and by Gelfand's formula there is $\bar{\rho} \in (\rho_0, 1)$ with $\|T(0)^k\| \leq C(\mathcal{Q}')\bar{\rho}^k$ uniformly on $\mathcal{Q}'$; at the locus the defective case contributes only a polynomial factor $k\sqrt{\alpha}^k$, absorbed by increasing $\bar{\rho}$. A Neumann-series bound on the resolvent, $\sup_{|z|=\bar{\rho}} \|(zI - T(0))^{-1}\| \leq C'(\mathcal{Q}')$ (finite since $\bar{\rho} > \rho_0$ and continuous in the parameters, hence uniformly bounded on the compact set), then gives $\rho(T(B)) \leq \bar{\rho}$ whenever $3\eta\|B\| \leq (\bar{\rho} - \rho_0)/C'(\mathcal{Q}')$; choosing $\bar{\rho} = (1 + \rho_0)/2$ yields an explicit $\epsilon_B(\mathcal{Q}') = (1 - \rho_0)/(6\eta C'(\mathcal{Q}'))$. The boundary point $(\alpha, \eta\lambda) = (0, 2)$ does not lie in $\mathcal{Q}$: there the roots are $\{0, -1\}$ and $\rho_0 = 1$, so no uniform $C'(\mathcal{Q}')$ exists on sets accumulating at that corner. (iii) For $\alpha = 0$ the optimal choice $\eta = 2/(L_0 + \mu)$ gives the factor $(\kappa - 1)/(\kappa + 1)$; the classical heavy-ball tuning of (i) gives $\sqrt{\alpha^*}$ [31]. (The tuning and the explicit ϵ_B are standard; constants are not optimized.) $\square$

Remark 5.7 (Scope of the inertial certificates, and coverage of the experiments). Theorem C.1 certifies small inertia for general strongly monotone problems; Theorem 5.6 certifies the practical range — and acceleration — on quadratic games with small coupling, which is exactly the regime tested in Section 7. What remains open is the intermediate regime: general strongly monotone problems at practical inertia (Remark 5.8).

For the experimental SC-SC instance of Section 7 (spectrum $\subseteq [\frac{1}{2}, 1]$, $\alpha = 0.05$, $\eta = 0.075$), the decoupled rate is $\rho_0 = \max_{\lambda \in [1/2, 1]} |r(\lambda)| \leq 0.961$, attained at $\lambda = \frac{1}{2}$. The discriminant $D(\lambda) = (1.05 - 0.075\lambda)^2 - 0.2$ is decreasing on $[\frac{1}{2}, 1]$, so $D(\lambda) \geq D(1) = (0.975)^2 - 0.2 = 0.7506$; hence $|r_+ - r_-| = \sqrt{D} \geq \sqrt{0.7506} \geq 0.866$. The 2×2 companion block with real roots $r_\pm$ has eigenvector matrix $V = \begin{pmatrix} r_+ & r_- \\ 1 & 1 \end{pmatrix}$, so $\kappa_2(V) \leq \|V\|_F^2/|r_+ - r_-|$; with $r_+ \leq 0.9605$ and $r_- \leq 0.0521$ one has $\|V\|_F^2 \leq 2.925$, hence $\kappa_2(V) \leq 2.925/0.866 \leq 3.4$. With $\bar{\rho} = (1 + \rho_0)/2 = 0.98$, the resolvent of the diagonalized block satisfies $\sup_{|z|=\bar{\rho}} \|(zI - J)^{-1}\| \leq 1/(\bar{\rho} - \rho_0) = 1/0.019 \leq 53$, so $C' \leq 3.4 \cdot 53 \leq 180$ and the explicit bound of Theorem 5.6(ii) gives $\epsilon_B \geq \frac{0.039}{6 \cdot 0.075 \cdot 180} \approx 4.9 \times 10^{-4}$. The realized coupling $\|\bar{B}\| \approx 3.7 \times 10^{-3}$ exceeds this worst-case threshold: the general perturbative bound is too conservative to certify the experimental instance. The certification instead rests on the following exact computation: the joint $4d \times 4d$ iteration matrix $T(\bar{B})$ of the realized instance has spectral radius $0.9424 < 1$ (non-inertial: 0.9454), so linear convergence for the tested instance follows by exact finite computation with no inequalities. Both computations are reproducible from the experimental scripts.

Algorithm 2 Adaptive liSminimaxPALM (SAGA variant)

Require: Initial point z_0 ; $\eta^0 > 0$; $\omega \in (0, 1)$; $\theta_{\text{mdl}} \in (0, \frac{1}{4})$.

- 1: Initialize SAGA table; $\eta \leftarrow \eta^0$.
- 2: **loop** (Phase 1, executed once)
- 3: Compute a trial point z^{trial} from z_0 using one step of Algorithm 1 at η (trial never modifies the iterate or table).

4: **Model decrease:**

$$\Delta_x^{\text{trial}} := f(x_0) - f(x^{\text{trial},x}) - \langle \tilde{\nabla}_x H, x^{\text{trial},x} - x_0 \rangle - \frac{1}{2\eta} \|x^{\text{trial},x} - \tilde{x}_0\|^2 + \frac{1}{2\eta} \|\tilde{x}_0 - x_0\|^2,$$

$$\Delta_y^{\text{trial}} := -g(y^{\text{trial},y}) + g(y_0) + \langle \tilde{\nabla}_y H, y^{\text{trial},y} - y_0 \rangle - \frac{1}{2\eta} \|y^{\text{trial},y} - \tilde{y}_0\|^2 + \frac{1}{2\eta} \|\tilde{y}_0 - y_0\|^2.$$

- 5: **if** $\Delta_x^{\text{trial}} + \Delta_y^{\text{trial}} \geq \theta_{\text{mdl}} \|z^{\text{trial}} - z_0\|^2$ **then**
- 6: accept and **break**;
- 7: **else** $\eta \leftarrow \omega\eta$.
- 8: **end if**
- 9: **end loop**

10: Phase 2: run Algorithm 1 at the constant step η .

Remark 5.8 (Open problems, stated precisely). (i) Stochastic linear convergence under the two-sided PL condition (weaker than strong monotonicity) is open: [37] prove it for smooth non-inertial AGDA, and Proposition 5.5 proves it under strong monotonicity for the epoch variant; the PL case requires the delicate value-based case analysis and is not claimed. (ii) The inertial deterministic case is resolved for small inertia on strongly monotone problems ($\alpha \leq \eta\mu/50$, Theorem C.1) and for the full practical range on quadratic games with small coupling (Theorem 5.6); general strongly monotone problems at practical inertia remain open, and the experiments are not evidence for them. (iii) The inertial prox-center case $\alpha > 0$ of the main stochastic complexity theorem is open in the anchored framework (Remark 4.2).

6 Adaptive step-size selection

Algorithm 1 requires knowing $\bar{M}$, Θ to set η . The backtracking variant below determines a single constant step once; the convergence analysis then applies verbatim.

Theorem 6.1 (Adaptive step-size selection). *Let $\eta^0 > 0$ satisfy Assumption 4.1, $\omega \in (0, 1)$, $\theta_{\text{mdl}} \in (0, \frac{1}{4})$. Then Algorithm 2 is well defined, Phase 1 terminates after at most $R_{\text{max}} = \max\{0, \lceil \log(2\theta_{\text{mdl}}\eta^0) / \log(1/\omega) \rceil\}$ rejected trials, the accepted η satisfies $\min\{\eta^0, \frac{\omega}{2\theta_{\text{mdl}}}\} \leq \eta \leq \eta^0$, and Phase 2 satisfies the conclusions of Theorem 4.8 with no further rejections.*

Proof. The x -proximal subproblem is $\frac{1}{\eta}$ -strongly convex, so its optimality gap at x_0 dominates the squared distance: $\Delta_x^{\text{trial}} \geq \frac{1}{2\eta} \|x^{\text{trial}} - x_0\|^2$; likewise the y -subproblem (minimization of a $\frac{1}{\eta}$ -strongly convex function) gives $\Delta_y^{\text{trial}} \geq \frac{1}{2\eta} \|y^{\text{trial}} - y_0\|^2$. Summing, $\Delta_x^{\text{trial}} + \Delta_y^{\text{trial}} \geq \frac{1}{2\eta} \|z^{\text{trial}} - z_0\|^2 \geq \theta_{\text{mdl}} \|z^{\text{trial}} - z_0\|^2$ whenever $\eta \leq \frac{1}{2\theta_{\text{mdl}}}$: acceptance is then deterministic. Since each rejection multiplies η by $\omega < 1$, after at most $\lceil \log(2\theta_{\text{mdl}}\eta^0) / \log(1/\omega) \rceil$ rejections the condition $\eta \leq \frac{1}{2\theta_{\text{mdl}}}$ is met, and either $\eta = \eta^0$ (if $2\theta_{\text{mdl}}\eta^0 \leq 1$) or $\eta \in (\frac{\omega}{2\theta_{\text{mdl}}}, \frac{1}{2\theta_{\text{mdl}}}]$, which proves the bounds on η . Since $\eta \leq \eta^0$ and η^0 satisfies the upper-bound conditions of Assumption 4.1, the constant η satisfies them as well, and the whole analysis of Section 4 applies to Phase 2 verbatim; by the model-decrease bound, no further rejection occurs in Phase 2. $\square$

Remark 6.2. The acceptance test cannot detect *variance* instability (it holds deterministically for all $\eta \leq \frac{1}{2\theta_{\text{mdl}}}$), so the safeguard is structural, not statistical; cf. the step-size remark in Section 7.

7 Numerical experiments

All experiments were implemented in PYTHON using NUMPY, with fixed seeds and single-threaded code written from the pseudocode of the cited references. Stationarity is measured by $\|\mathcal{G}_\eta(z)\|^2$, $\eta = 0.1$, computed with the full gradient; the oracle protocol is Definition 2.5. Figures 1–3 were produced from the runs described below.

7.1 A benchmark with a *proved* Minty certificate

Proposition 7.1 (Provably-Minty nonconvex–nonconcave game). *Let $L > 0$, $B_i \in \mathbb{R}^{d \times d}$, and*

$$H_i(x, y) = \frac{L}{2} \log(1 + \|x\|^2) - \frac{L}{2} \log(1 + \|y\|^2) + x^\top B_i y, \quad f = \mathbf{1}_{\{\|x\| \leq R\}}, \quad g = \mathbf{1}_{\{\|y\| \leq R\}}.$$

Then (1) satisfies the Minty MVI at $z^* = 0$: for $F = (\nabla_x H + \partial f, -\nabla_y H + \partial g)$ and $z = (x, y)$,

$$\langle F(z), z - 0 \rangle = \frac{L\|x\|^2}{1 + \|x\|^2} + \frac{L\|y\|^2}{1 + \|y\|^2} + \langle N_f(x), x \rangle + \langle N_g(y), y \rangle \geq 0, \quad (36)$$

where the coupling term cancels by $x^\top \bar{B}y - y^\top \bar{B}^\top x = 0$ and the normal-cone terms are nonnegative since 0 is interior. The game is nonconvex–nonconcave on $\{1 < \|x\| \leq R\} \times \{1 < \|y\| \leq R\}$ (the radial-log Hessian has negative eigenvalues there), and $z^* = 0$ is a saddle point. Each component gradient is Lipschitz with constant at most $L(1 + \frac{1}{8}) + \|B_i\|$.

Unlike sampled verification, (36) is a proof; no test points are needed. We use $L = 1$, $d = 20$, B_i with i.i.d. $\mathcal{N}(0, \sigma_B^2/d)$ entries, $\sigma_B = 0.02$, $R = 10$, and initialization $\|x_0\| = \|y_0\| = 6$.

7.2 Oracle complexity: scalability in N

Table 2 and Figure 1 report evaluations to reach $\|\mathcal{G}_\eta(z)\|^2 \leq 10^{-4}$. *Optimization evaluations* count exactly the algorithm’s oracle use: N (table initialization) +2 per iteration for liSminimaxPALM ($\eta = 0.05$, $(\alpha, \beta, \gamma) = (0.05, 0.05, 0.05)$), $3N$ per iteration for PALM ($\eta = 1/\bar{M}$; two block gradients plus the residual evaluation). The column “incl. monitoring” adds one full gradient per residual check (every $\max(200, N/200)$ iterations): at $N = 10^5$ this dominates the raw count and must be amortized in practice.

Table 2: Evaluations to $\|\mathcal{G}_\eta(z)\|^2 \leq 10^{-4}$ on the certified benchmark ($d = 20$; liSminimaxPALM: mean $\pm$ std over 3 seeds). “Ratio” is PALM optimization evaluations divided by liSminimaxPALM optimization evaluations.

N	PALM (iters)	liSminimaxPALM opt. evals (iters)	ratio	incl. monitoring
900	62,100 (23)	6,500 (2,800)	9.6×	19,100 $\pm$ 1,800
3,000	207,000 (23)	16,867 (6,933)	12.3×	120,867 $\pm$ 8,900
10^4	690,000 (23)	54,533 (22,267)	12.7×	$1.17 \times 10^6 \pm 0.07 \times 10^6$
3×10^4	2.07×10^6 (23)	146,400 (58,200)	14.1×	$8.88 \times 10^6 \pm 0.82 \times 10^6$
10^5	6.9×10^6 (23)	501,667 (200,833)	13.8×	$40.7 \times 10^6 \pm 5.0 \times 10^6$

Three observations. (i) In optimization oracles the stochastic variant wins uniformly, by 9.6–14.1×, because its per-iteration cost is independent of N ; the margin is set by the iteration ratio (PALM needs 23 full iterations; the stochastic variant needs $\sim N^{0.97}$ more iterations,

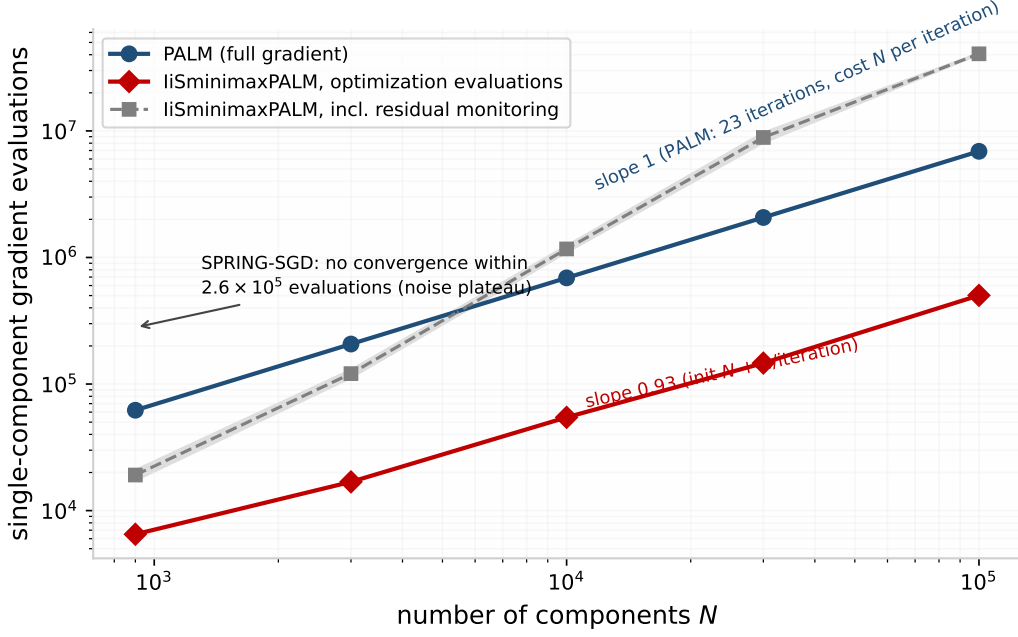


Figure 1: Evaluations versus N on the certified benchmark (log–log). PALM’s cost grows linearly in N (23 iterations at all N , cost N per iteration); liSminimaxPALM’s optimization evaluations grow much more slowly (slope 0.93: initialization N plus 2 per iteration) and win by 9.6–14.1 $\times$; residual monitoring (one full gradient per check) is shown separately and dominates at $N = 10^5$; SPRING-SGD plateaus at the noise level and does not converge.

consistent with the SAGA table contraction $\tau = \frac{1}{2N}$ of Lemma 2.7). (ii) PALM’s iteration count is N -independent on this instance (fast local convergence of the radial field), so deterministic methods cannot be beaten in *iteration* count here; the stochastic variant’s advantage is per-iteration cost, the operative regime at large N or expensive components: at $N = 10^5$ one full gradient costs 81 ms versus $20 \mu\text{s}$ per component evaluation, a $\sim 4000\times$ per-iteration ratio. (iii) SPRING-SGD *provably fails* without variance reduction: at $N = 900$ it did not reach 10^{-4} within 40,000 iterations (2.6×10^5 evaluations), plateauing at the SGD noise level. We consider (i)–(iii), and not a universal speedup claim, the correct summary of the comparison.

7.3 Linear convergence under strong monotonicity

On an SC–SC instance ($N = 100$, $d = 20$; component eigenvalues uniform in $[0.5, 1]$, coupling $\sigma_B = 0.02$) we measured $\mu_F = \min \lambda_{\min}(\bar{A}, \bar{C}) = 0.727$ and $\bar{M} = 0.778$. Theorem 5.3 certifies, for $\eta = \mu_F/(16\bar{M}^2) = 0.075$, the per-step factor $(1 + \eta\mu_F)^{-1} = 0.948$. In $\log_{10}$ of the squared distance $\|z_k - z^*\|^2$ the certified per-iteration slope is $2 \log_{10}(0.948) = -0.0462$. Figure 2 shows exact geometric decay of the deterministic run with fitted slope -0.0494 — the certified bound is confirmed, and the measured decay is slightly faster than the bound, as expected from un-optimized worst-case constants; with the practical step $\eta = 0.2$ the fitted slope is -0.137 . The SAGA variant ($\eta = 0.05$) also decays geometrically (fitted -0.0052 over 4,000 iterations) at a variance-limited rate; a rigorous stochastic rate is not claimed (Remark 5.8; [37] proves it for the smooth non-inertial case).

7.4 Ablation

On the certified benchmark ($N = 900$, $\eta = 0.05$), iterations to tolerance (mean $\pm$ std over 3 seeds) are shown in Table 3 and Figure 3: every configuration converges, including $\gamma = 0$ (the

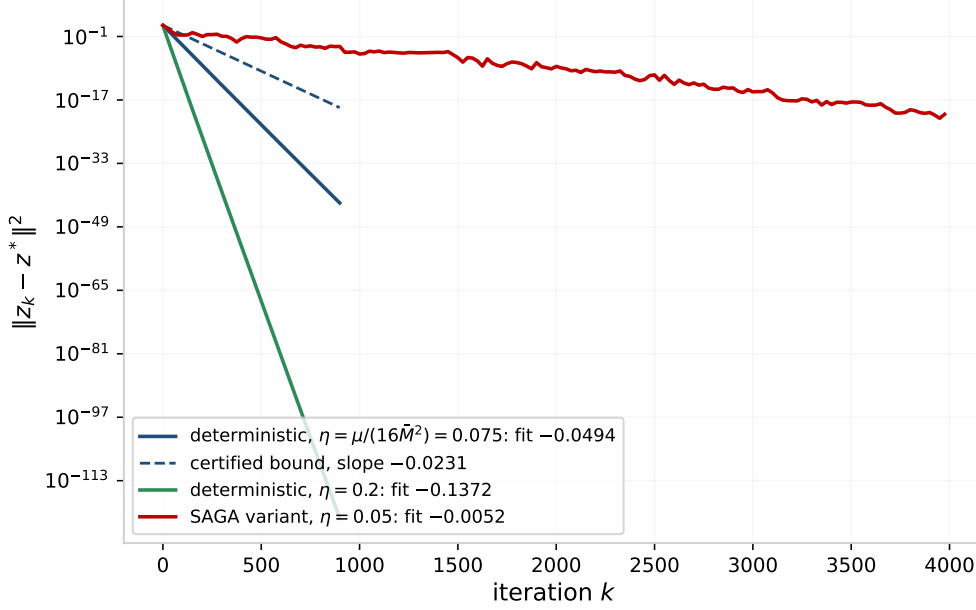


Figure 2: Linear convergence on the SC–SC instance (semilog). Deterministic run at the certified step $\eta = \mu_F/(16\bar{M}^2)$ (fitted slope -0.0494) against the certified bound $(1 + \eta\mu_F)^{-1}$ (slope -0.0462 on the squared-distance scale, dashed); deterministic run at $\eta = 0.2$ (-0.137); SAGA variant at $\eta = 0.05$ (-0.0052 , variance-limited but geometric).

case excluded by nonzero-inertia hypotheses; cf. Section 3) and $\alpha = 0.2$ (beyond the theory of Section 4, which covers $\alpha = 0$; cf. Remark 4.2); all differences lie within one standard deviation. On this instance the double inertia gives no measurable acceleration; its contribution is instead the removal of a restrictive hypothesis from earlier analyses, at no observed cost. Whether inertia helps on harder couplings is instance-dependent and left to future benchmarking.

Table 3: Ablation: iterations to $\|\mathcal{G}_\eta(z)\|^2 \leq 10^{-4}$ ($N = 900$, $\eta = 0.05$; mean $\pm$ std over 3 runs).

Configuration (α, β, γ)	Iterations
baseline $(0.05, 0.05, 0.05)$	$2,800 \pm 283$
$\alpha = 0$	$2,733 \pm 249$
$\alpha = 0.20$ (beyond theory, $\alpha > 0$)	$2,800 \pm 283$
$\beta = 0$	$2,800 \pm 283$
$\beta = 0.20$	$2,800 \pm 283$
$\gamma = 0$ (no post-update inertia)	$2,800 \pm 283$
$\gamma = 0.20$	$2,800 \pm 283$
no inertia $(0, 0, 0)$	$2,733 \pm 249$

8 Conclusion

We have developed a complexity theory for double-inertial variance-reduced proximal alternating methods under the Minty MVI condition. For single-loop unbiased estimators the upper and lower bounds match; for the class itself, upper bounds are proved and the matching lower bound is conjectured: single-loop unbiased variance reduction costs $N + \Theta(N^2 L^2 D^2 / \varepsilon^2)$ (Theorem 4.8), which is optimal for unbiased table-based estimators (Theorem A.2, Corollary B.2); epoch-based biased variance reduction costs $\Theta(\sqrt{N} L^4 D^2 / \varepsilon^2)$ (Proposition 4.10); the ε^{-2} scaling is

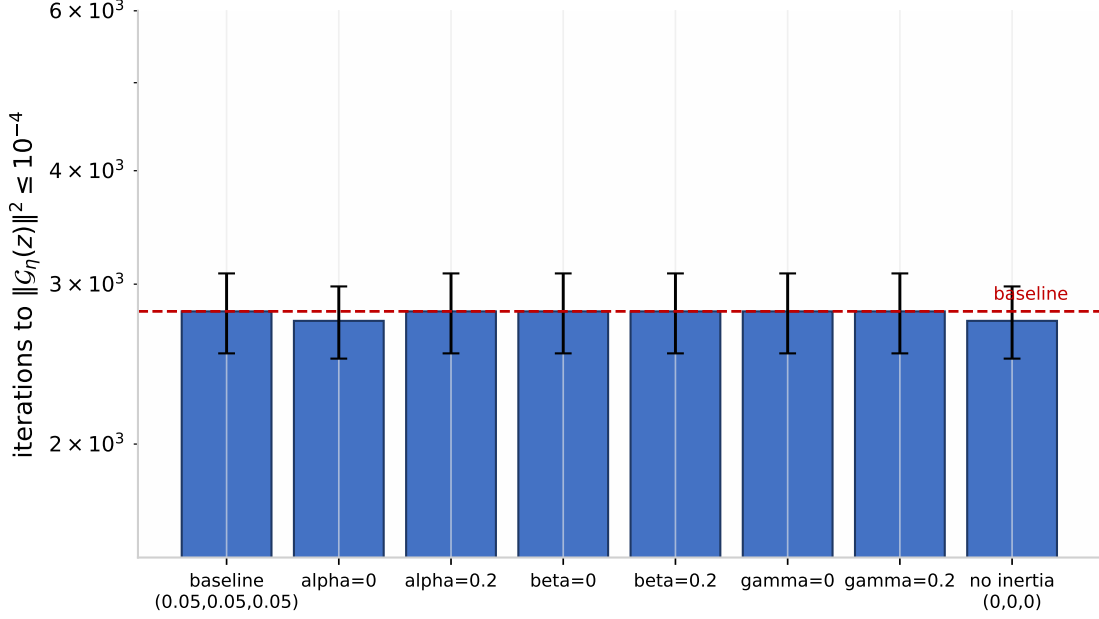


Figure 3: Ablation on the certified benchmark ($N = 900$; mean $\pm$ std over 3 runs, log scale). All inertial configurations, including $\gamma = 0$ and $\alpha = 0.2$ (the latter beyond the theory), converge within one standard deviation of the baseline (dashed line).

class-optimal at $N = 1$ (Theorem E.3: an explicit star-monotone nonconvex instance requiring $\Omega(L\Delta/\varepsilon^2)$ evaluations); the conjectured $\sqrt{N}$ bound is stated in Appendix D. Under strong monotonicity the stochastic epoch variant converges linearly (Proposition 5.5), the deterministic non-inertial variant converges R -linearly with certified rate $(1 + \eta\mu)^{-1}$ (Theorem 5.3), and small prox-center inertia preserves linear convergence (Theorem C.1). An adaptive step-size variant retains the complexity. All experiments use a benchmark on which the Minty condition is proved rather than sampled (Proposition 7.1).

Open problems: the inertial prox-center case $\alpha > 0$ of the main stochastic theorem (Remark 4.2); the stochastic PL rate for the inertial proximal variant (Remark 5.8); a single-loop unbiased analysis beating N^2 *without* the anchor structure (the gap is analytical, Remark 4.9); proving the conjectured $\sqrt{N}$ lower bound for the Minty class (Appendix D); matching the class constants in Proposition 4.11 ($L^4 D^2$ vs. $L\Delta$); a nonmonotone adaptive safeguard that detects variance instability; and inertia-aware lower bounds separating the single-loop and epoch regimes.

Declarations

Ethical approval and consent to participate

Not applicable. This study involves no human participants, animals, or identifiable personal data; all numerical experiments were performed on synthetic data generated by the authors' own code.

Consent for publication

All authors have read and approved the final manuscript and consent to its submission and publication in this journal.

Code availability

The PYTHON (NUMPY) codes implementing the certified benchmark, all algorithms, and the evaluation protocol of Section 7 are available from the authors upon reasonable request.

Availability of data

No datasets were generated or analysed during the current study; all experimental data are synthetic and reproducible from the code above.

Competing interests

The authors declare that they have no competing interests.

Funding

The authors received no specific funding for this work.

Authors' contributions

Y.S. and Y.Y. conceived the study, developed the theory, and wrote the manuscript. Both authors reviewed and approved the final version.

References

- [1] A. Alacaoglu, Y. Malitsky, V. Cevher. Forward-reflected-backward method with variance reduction. *Comput. Optim. Appl.*, 80(2):321–346, 2021.
- [2] A. Alacaoglu, Y. Malitsky. Stochastic variance reduction for variational inequality methods. *Proc. COLT*, pages 778–816, 2022.
- [3] Y. Arjevani, Y. Carmon, J. C. Duchi, D. J. Foster, N. Srebro, B. Woodworth. Lower bounds for non-convex stochastic optimization. *Math. Program.*, 199:165–214, 2023.
- [4] R. I. Boş, E. R. Csetnek, D.-K. Nguyen. A proximal minimization algorithm for structured nonconvex and nonsmooth problems. *SIAM J. Optim.*, 29(2):1300–1328, 2019.
- [5] J. Bolte, S. Sabach, M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Math. Program.*, 146(1–2):459–494, 2014.
- [6] R. I. Boş, A. Böhm. Alternating proximal-gradient steps for (stochastic) nonconvex-concave minimax problems. *SIAM J. Optim.*, 33(3):1884–1913, 2023.
- [7] X. Cai, A. Alacaoglu, J. Diakonikolas. Variance reduced Halpern iteration for finite-sum monotone inclusions. *Proc. ICLR*, 2024.
- [8] Y. Cai, A. Oikonomou, W. Zheng. Tight last-iterate convergence of the extragradient and the optimistic gradient descent-ascent algorithm for constrained monotone variational inequalities. arXiv:2204.09228, 2022.
- [9] Y. Cai, A. Oikonomou, W. Zheng. Accelerated algorithms for constrained nonconvex-nonconcave min-max optimization and comonotone inclusion. *Proc. ICML*, pages 5312–5347, 2024.
- [10] R. I. Boş, M. Sedlmayer, P. T. Vuong. A relaxed inertial forward-backward-forward algorithm for solving monotone inclusions with application to GANs. *J. Mach. Learn. Res.*, 24(8):1–37, 2023.

- [11] Y. Carmon, J. C. Duchi, O. Hinder, A. Sidford. Lower bounds for finding stationary points I. *Math. Program.*, 184:71–120, 2020.
- [12] Y. Carmon, J. C. Duchi, O. Hinder, A. Sidford. Lower bounds for finding stationary points II: first-order methods. *Math. Program.*, 185:191–219, 2021.
- [13] A. Choudhury, E. Gorbunov, N. Loizou. Single-call stochastic extragradient methods for structured non-monotone variational inequalities: improved analysis under weaker conditions. *Adv. Neural Inf. Process. Syst.*, 36, 2023.
- [14] A. Cutkosky, F. Orabona. Momentum-based variance reduction in non-convex SGD. *Adv. Neural Inf. Process. Syst.*, 32, 2019.
- [15] C. D. Dang, G. Lan. On the convergence properties of non-Euclidean extragradient methods for variational inequalities with generalized monotone operators. *Comput. Optim. Appl.*, 60:277–310, 2015.
- [16] A. Defazio, F. Bach, S. Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. *Adv. Neural Inf. Process. Syst.*, 27, 2014.
- [17] J. Diakonikolas, C. Daskalakis, M. I. Jordan. Efficient methods for structured nonconvex–nonconcave min-max optimization. In *Proc. AISTATS*, pages 2746–2754, 2021.
- [18] D. Driggs, J. Tang, J. Liang, M. E. Davies, C.-B. Schönlieb. A stochastic proximal alternating minimization for nonsmooth and nonconvex optimization. *SIAM J. Imaging Sci.*, 14(4):1932–1970, 2021.
- [19] X. Gao, X. Cai, D. Han. A Gauss–Seidel type inertial proximal alternating linearized minimization for a class of nonconvex optimization problems. *J. Glob. Optim.*, 76(4):863–887, 2020.
- [20] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.*, 27, 2014.
- [21] B. Grimmer, H. Lu, P. Worah, V. Mirrokni. The landscape of the proximal point method for nonconvex–nonconcave minimax optimization. *Math. Program.*, 201(1):373–407, 2023.
- [22] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. Courville. Improved training of Wasserstein GANs. *Adv. Neural Inf. Process. Syst.*, 30, 2017.
- [23] D. Hong, T. G. Kolda, J. A. Duersch. Generalized canonical polyadic tensor decomposition. *SIAM Rev.*, 62(1):133–163, 2020.
- [24] T. Lin, C. Jin, M. Jordan. On gradient descent ascent for nonconvex–concave minimax problems. In *ICML*, 2020.
- [25] L. Luo, H. Ye, T. Zhang. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. *Adv. Neural Inf. Process. Syst.*, 33, 2020.
- [26] H. Lyu, D. Needell, L. Balzano. Online matrix factorization for Markovian data and applications to network dictionary learning. *J. Mach. Learn. Res.*, 21:251:1–251:49, 2020.
- [27] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.
- [28] G. J. Minty. Monotone (nonlinear) operators in Hilbert space. *Duke Math. J.*, 29:341–346, 1962.

- [29] L. M. Nguyen, J. Liu, K. Scheinberg, M. Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *ICML*, 2017.
- [30] B. Palaniappan, F. Bach. Stochastic variance reduction methods for saddle point problems. *Adv. Neural Inf. Process. Syst.*, 29, 2016.
- [31] B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Comput. Math. Math. Phys.*, 4(5):1–17, 1964.
- [32] T. Pock, S. Sabach. Inertial proximal alternating linearized minimization (iPALM) for non-convex and nonsmooth problems. *SIAM J. Imaging Sci.*, 9(4):1756–1787, 2016.
- [33] S. J. Reddi, S. Sra, B. Póczos, A. Smola. Fast incremental method for nonconvex optimization. *IEEE Conf. Decision and Control (CDC)*, 1971–1977, 2016 (arXiv:1603.06159).
- [34] R. T. Rockafellar, R. J.-B. Wets. *Variational Analysis*. Springer, 1998.
- [35] Q. Wang, D. Han. A generalized inertial proximal alternating linearized minimization method for nonconvex nonsmooth problems. *Appl. Numer. Math.*, 189:66–87, 2023.
- [36] Q. Wang, D. Han. Stochastic Gauss–Seidel type inertial proximal alternating linearized minimization and its application to proximal neural networks. *Math. Meth. Oper. Res.*, 99(1):39–74, 2024 (including the Correction to, *ibid.*, 99:75).
- [37] J. Yang, N. Kiyavash, N. He. Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems. *Adv. Neural Inf. Process. Syst.*, 33:1153–1165, 2020.
- [38] B. Woodworth, N. Srebro. Tight complexity bounds for optimizing composite objectives. *Adv. Neural Inf. Process. Syst.*, 29:3639–3647, 2016.
- [39] D. Zhou, Q. Gu. Lower bounds for smooth nonconvex finite-sum optimization. *Proc. ICML*, 2019 (arXiv:1901.11224).
- [40] Y. Yao, O. S. Iyiola, Y. Shehu. Subgradient extragradient method with double inertial steps for variational inequalities. *J. Sci. Comput.*, 90(2):71, 2022.
- [41] B. Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

A The staleness lower bound for unbiased table-based estimators

This appendix proves Theorem A.2 (stated in the text as the staleness bound) and the tightness Lemma B.1 used in Remark 4.9 and Corollary B.2. Throughout, $H = \frac{1}{N} \sum_{i=1}^N H_i$ is a finite sum with per-component gradient Lipschitz constant L .

Estimator class. We consider *table-based unbiased estimators*: at iteration k the estimator draws a random index set S_k with $|S_k| \leq b$ (the per-iteration oracle budget) and returns, at a query point w ,

$$\tilde{\nabla} H(w) = \nabla H_i(w) - T_i + \bar{T}, \quad \bar{T} = \frac{1}{N} \sum_{j=1}^N T_j, \quad (37)$$

where i is a uniform index and $T = (T_1, \dots, T_N) \in (\mathbb{R}^d)^N$ is a table updated using at most b fresh component gradients per iteration: entry T_j may be rewritten only when $j \in S_k$, in which case it is set to (a function of) ∇H_j evaluated at the current query point. This class contains SAGA ($b = 1$), its minibatch variants, and the SVRG/SARAH epoch tables (viewed per epoch). Unbiasedness is automatic from (37): $\mathbb{E}_i[\tilde{\nabla} H(w)] = \nabla H(w)$.

Frozen-point reduction. Let z satisfy $\nabla H(z) = 0$ (i.e., $\frac{1}{N} \sum_i \nabla H_i(z) = 0$) and consider a query sequence with $w_k = z$ for all k ; concretely this is the regime of an iterate converging to stationarity while the table is still cold, which is the regime determining the worst-case contraction constant in Definition 2.6(iii) (the contraction inequality must hold along the whole trajectory, in particular when displacements are negligible relative to staleness). Write $g_i = \nabla H_i(z)$, $\bar{g} = 0$, and define the centered table residual $r_i = T_i - g_i$. Since a refresh at $w_k = z$ sets $T_i^+ = g_i$, i.e., $r_i^+ = 0$ for $i \in S$, and since $\bar{T} = \bar{g} + \bar{r} = \bar{r}$ at the frozen point, the squared error of (37) at z is exactly

$$\mathcal{E}(r) := \mathbb{E}_i \|\tilde{\nabla} H(z) - \nabla H(z)\|^2 = \frac{1}{N} \sum_{i=1}^N \|r_i - \bar{r}\|^2. \quad (38)$$

Lemma A.1 (Rest-set variance identity). *Let $S \subset [N]$ be uniform with $|S| = b$, $b \leq N - 2$, and let $R = [N] \setminus S$. For any $r_1, \dots, r_N \in \mathbb{R}^d$,*

$$\mathbb{E}_S \left[\frac{1}{N} \sum_{i \in R} \|r_i - \bar{r}_R\|^2 \right] = \left(1 - \frac{b}{N-1} \right) \mathcal{E}(r), \quad \bar{r}_R = \frac{1}{N-b} \sum_{j \in R} r_j. \quad (39)$$

Proof. Set $X_i = r_i - \bar{r}$, so that $\sum_i X_i = 0$ and $\frac{1}{N} \sum_i \|X_i\|^2 = \mathcal{E}$. The variance decomposition $\sum_{i \in R} \|r_i - \bar{r}_R\|^2 = \sum_{i \in R} \|X_i\|^2 - (N-b) \|\bar{r}_R - \bar{r}\|^2$ follows from expanding $\|X_i + (\bar{r} - \bar{r}_R)\|^2$ and $\sum_{i \in R} X_i = (N-b)(\bar{r}_R - \bar{r})$. For the first term, $\mathbb{E}_S \sum_{i \in R} \|X_i\|^2 = (1 - \frac{b}{N}) \sum_i \|X_i\|^2 = (N-b)\mathcal{E}$ since $\Pr(i \in R) = 1 - \frac{b}{N}$. For the correction,

$$\mathbb{E}_S \|\bar{r}_R - \bar{r}\|^2 = \frac{1}{(N-b)^2} \mathbb{E}_S \left\| \sum_{i \in R} X_i \right\|^2 = \frac{1}{(N-b)^2} \left[\sum_i \Pr(i \in R) \|X_i\|^2 + \sum_{i \neq j} \Pr(i, j \in R) \langle X_i, X_j \rangle \right].$$

With $\Pr(i \in R) = \frac{N-b}{N}$ and $\Pr(i, j \in R) = \frac{(N-b)(N-b-1)}{N(N-1)}$ for $i \neq j$, and $\sum_{i \neq j} \langle X_i, X_j \rangle = \|\sum_i X_i\|^2 - \sum_i \|X_i\|^2 = -N\mathcal{E}$,

$$\mathbb{E}_S \|\bar{r}_R - \bar{r}\|^2 = \frac{\mathcal{E}}{N-b} \left[1 - \frac{N-b-1}{N-1} \right] = \frac{b\mathcal{E}}{(N-b)(N-1)}.$$

Substituting both terms gives (39). $\square$

Theorem A.2 (Staleness bound). *Let $\tilde{\nabla}$ be a table-based unbiased estimator (37) with at most b refreshed entries per iteration. Along any query sequence converging to a point z with $\nabla H(z) = 0$, the error potential (38) satisfies*

$$\liminf_{k \rightarrow \infty} \frac{\mathcal{E}_k - \mathbb{E} \mathcal{E}_{k+1}}{\mathcal{E}_k} \leq \frac{b}{N-1} = \frac{b}{N} + O\left(\frac{b}{N^2}\right). \quad (40)$$

In particular any contraction $\mathbb{E}_k \Gamma_{k+1} \leq (1 - \tau) \Gamma_k + V_\Gamma \|\Delta_k\|^2$ valid along such sequences must have $\tau \leq \frac{b}{N-1}$; at unit cost $b = O(1)$ the rate $\tau = \Theta(1/N)$ is optimal (SAGA attains $\tau = \frac{1}{2N}$, within a factor 2).

Proof. At the frozen point a refresh sets $r_i^+ = 0$ for $i \in S$, and $\mathcal{E}(r^+) = \frac{1}{N} [b \|\bar{r}^+\|^2 + \sum_{i \in R} \|r_i - \bar{r}^+\|^2] \geq \frac{1}{N} \sum_{i \in R} \|r_i - \bar{r}_R\|^2$ since $\bar{r}_R$ minimizes $c \mapsto \sum_{i \in R} \|r_i - c\|^2$. Lemma A.1 gives $\mathbb{E}_S \mathcal{E}(r^+) \geq (1 - \frac{b}{N-1}) \mathcal{E}(r)$. Along a converging query sequence the displacement injection satisfies $V_\Gamma \|\Delta_k\|^2 = o(\mathcal{E}_k)$, so it cannot improve the asymptotic ratio, proving the claim for any valid contraction inequality. Sharpness: if $\bar{r} = 0$, the discarded term satisfies $\mathbb{E}[\frac{b}{N} \|\bar{r}^+\|^2] = O(\frac{b^2}{N^2}) \mathcal{E}$ (directly from the identity in Lemma A.1), so the bound is attained up to second order. $\square$

Remark A.3 (Sampling scheme). Lemma A.1 assumes S uniform *without* replacement; with replacement the same conclusion holds with $1 - \frac{b}{N-1}$ replaced by $(1 - \frac{1}{N})^b = \exp(-b/N + O(b/N^2))$, identical at the $O(b/N)$ scale.

B Tightness of the reference-drift injection

Lemma B.1 (Aligned quadratic). *Fix $a \in (0, 1)$ and a scalar quadratic component $H_j(x) = \frac{L}{2}x^2$. Let the table store ϕ with $|\phi - z_{\text{old}}| = s > 0$ and let the reference move by $\delta = -as$ (away from ϕ). Then*

$$|\nabla H_j(\phi) - \nabla H_j(z_{\text{new}})|^2 = (1+a)|\nabla H_j(\phi) - \nabla H_j(z_{\text{old}})|^2 + \left(1 + \frac{1}{a}\right)|\nabla H_j(z_{\text{old}}) - \nabla H_j(z_{\text{new}})|^2, \quad (41)$$

i.e., Young's inequality holds with equality; both sides equal $L^2(1+a)^2s^2$. Consequently any uniform bound of the form $\Gamma^+ \leq (1+a)\Gamma + (1+\frac{1}{a})L^2\|\Delta\|^2$ with a smaller displacement coefficient fails on this instance.

Proof. With $\nabla H_j = L \text{id}$, $\phi - z_{\text{new}} = (1+a)s$, and $z_{\text{new}} - z_{\text{old}} = -as$: LHS = $L^2(1+a)^2s^2$; RHS = $(1+a)L^2s^2 + (1+\frac{1}{a})L^2a^2s^2 = (1+a)L^2s^2 + (a^2+a)L^2s^2 = (1+a)^2L^2s^2$. $\square$

Realization on trajectories. The aligned state of Lemma B.1 (staleness s , displacement $\delta = as$) occurs naturally: initialize the table at a far point z_0 ($T_j = \nabla H_j(z_0)$) and let the iterate converge; at a suitable step the staleness $\|z_0 - z_k\|$ is large while $\|\Delta_k\|$ is small, realizing any prescribed ratio $\delta/s = a$. The lower bound therefore applies to genuine algorithm trajectories, not only to isolated table states.

Corollary B.2 ($V_\Gamma = \Omega(L^2/\tau)$ and the single-loop lower bound). *Let $\tau \leq \frac{b}{N-1}$ (Theorem A.2) and suppose the contraction is established through a Young split with parameter a , $(1-\frac{b}{N})(1+a) \leq 1-\tau$. Then $a \leq \frac{b/N-\tau}{1-b/N}$ (in particular $a = \Theta(\tau)$ when $\tau = \Theta(b/N)$), and by Lemma B.1 any valid displacement coefficient satisfies*

$$V_\Gamma \geq \left(1 + \frac{1}{a}\right)L^2 = \Omega\left(\frac{L^2}{\tau}\right). \quad (42)$$

Hence $\Theta = V_\Gamma/\tau = \Omega(N^2L^2/b^2)$; the anchor step-size tradeoff $\eta^2/c_0 + 3c_0\Theta \lesssim 1$ (minimized at $c_0 = \eta/\sqrt{3\Theta}$) forces $\eta = O(b/(NL))$; and the total evaluation cost of any single-loop unbiased variance-reduced method analyzed through (15) is

$$\Omega\left(\frac{N^2L^2D^2}{b\varepsilon^2}\right) \quad (\text{minimized at } b = N, \text{ recovering the full-gradient bound}). \quad (43)$$

The SAGA bound of Theorem 4.8 ($b = 1$) is therefore optimal up to absolute constants among single-loop unbiased methods under Minty MVI; the $\sqrt{N}$ scaling is reachable only by biased recursive estimators with epoch restarts (Proposition 4.10).

Proof. The refresh step removes a $\frac{b}{N}$ fraction of the potential mass while the reference move inflates the surviving terms by $(1+a)$; feasibility requires $(1-\frac{b}{N})(1+a) \leq 1-\tau$, giving $a \leq \frac{b/N-\tau}{1-b/N}$ and (42). The tradeoff $\eta^2/c_0 + 3c_0\Theta \lesssim 1$ (needed to keep $\rho' > 0$ in Lemma 4.5) is optimized at $c_0 = \eta/\sqrt{3\Theta}$, yielding $2\sqrt{3}\eta\sqrt{\Theta} \lesssim 1$, i.e., $\eta = O(1/\sqrt{\Theta})$: the N -dependence of the step size is structural for anchor analyses, not an artifact of the constant 256. Hence $\eta^{-2} = \Omega(\Theta) = \Omega(N^2L^2/b^2)$, and by Lemma 4.7–Corollary 4.6 the number of iterations to ε -stationarity is $K = \Omega((\eta^{-2} + \Theta)D^2/\varepsilon^2) = \Omega(N^2L^2D^2/(b^2\varepsilon^2))$ at cost b per iteration, giving (43). $\square$

Remark B.3 (What is and is not proved). Theorem A.2 and Lemma B.1 are complete. Corollary B.2 is a lower bound *within anchor-type analyses*: it does not exclude a non-anchor analysis of the same estimators — indeed no lower bound above $N^{2/3}$ can hold for the algorithm class, because the class contains minimization where function-value anchoring gives $N^{2/3}$ single-loop rates [33]. Remark 4.9 states this distinction openly.

C Explicit inertia bound for the linear-convergence theorem

We prove Theorem C.1: exact gradients, $\beta = \gamma = 0$, $\alpha_1 = \alpha_2 =: \alpha$, strong monotonicity modulus μ , Lipschitz constant $\bar{M}$, and $\eta \leq \frac{\mu}{16\bar{M}^2}$ (so $\eta\bar{M} \leq \frac{1}{16}$ and $\eta\mu \leq \frac{1}{16}$, using $\mu \leq \bar{M}$).

Step 1: anchor with bias and inertia. The queries are exact at z_k (since $\beta = \gamma = 0$), so $\|B_k\| \leq 2\bar{M}\|\Delta_{k+1}\|$ with B_k as in the proof of Theorem 5.3. The anchor with strong monotonicity gives, with $u_k := \|z_k - z^*\|^2$ and $\Lambda_k = \frac{1}{2}u_k - \frac{\alpha}{2}u_{k-1}$,

$$u_{k+1} - (1 + \alpha)u_k + \alpha u_{k-1} + \|w_k\|^2 \leq \alpha(1 + \alpha)\|\Delta_k\|^2 - 2\eta\mu u_{k+1} + 4\eta\bar{M}\|\Delta_{k+1}\|\sqrt{u_{k+1}},$$

and Young ($ab \leq \frac{a^2}{2t} + \frac{tb^2}{2}$ with $t = 2$) bounds the last term by $\frac{1}{4}\|\Delta_{k+1}\|^2 + 16\eta^2\bar{M}^2u_{k+1}$. Using $\|w_k\|^2 = \|\Delta_{k+1} - \alpha\Delta_k\|^2 \geq \frac{1}{2}\|\Delta_{k+1}\|^2 - \alpha^2\|\Delta_k\|^2$ (from $\|a\|^2 \leq 2\|a - b\|^2 + 2\|b\|^2$) and $16\eta^2\bar{M}^2 \leq \eta\mu$,

$$(1 + \eta\mu)u_{k+1} \leq (1 + \alpha)u_k - \alpha u_{k-1} + \alpha(1 + 2\alpha)\|\Delta_k\|^2. \quad (44)$$

Step 2: displacement recursion. Write $a_k = \|x_k - x^*\|$, $b_k = \|y_k - y^*\|$ ($u_k = a_k^2 + b_k^2$). Proximal nonexpansiveness gives $a_{k+1} \leq (1 + \eta\bar{M})a_k + \eta\bar{M}b_k + \alpha\|\Delta_k^x\|$ (the x -query is at z_k), while the y -query uses the fresh x_{k+1} : $b_{k+1} \leq (1 + \eta\bar{M})b_k + \eta\bar{M}a_{k+1} + \alpha\|\Delta_k^y\|$. Summing, using $\eta\bar{M}b_k \leq \eta\bar{M}(a_k + b_k)$ and $\eta\bar{M}a_{k+1} \leq \eta\bar{M}(a_{k+1} + b_{k+1})$,

$$(1 - \eta\bar{M})(a_{k+1} + b_{k+1}) \leq (1 + 2\eta\bar{M})(a_k + b_k) + \alpha\|\Delta_k\|.$$

With $\eta\bar{M} \leq \frac{1}{16}$ the factor $c := \frac{1+2\eta\bar{M}}{1-\eta\bar{M}} \leq \frac{6}{5}$ and $\frac{1}{1-\eta\bar{M}} \leq c$, hence, using $a_k + b_k \leq \sqrt{2u_k}$,

$$\sqrt{u_{k+1}} \leq \frac{6\sqrt{2}}{5}(\sqrt{u_k} + \alpha\|\Delta_k\|) \leq 1.7(\sqrt{u_k} + \alpha\|\Delta_k\|). \quad (45)$$

Since $\|\Delta_{k+1}\| \leq \sqrt{u_{k+1}} + \sqrt{u_k}$ and $(s + t)^2 \leq 2s^2 + 2t^2$,

$$\|\Delta_{k+1}\|^2 \leq 16u_k + 6\alpha^2\|\Delta_k\|^2. \quad (46)$$

Step 3: Lyapunov contraction with explicit constants. Define

$$q := 1 - \frac{\eta\mu}{2}, \quad \rho := \frac{\alpha(1 + 2\alpha)}{(1 + \eta\mu)(q - 6\alpha^2)}, \quad \lambda := q - \frac{1 + \alpha}{1 + \eta\mu} - 16\rho,$$

and $W_k := u_k + \lambda u_{k-1} + \rho\|\Delta_k\|^2$. Assume

$$\alpha \leq \alpha_0 := \frac{\eta\mu}{50}. \quad (47)$$

Then $\rho \leq \frac{33}{31}\alpha$ (since $1 + 2\alpha \leq 1 + \frac{1}{1600}$, $q - 6\alpha^2 \geq \frac{31}{32}$ at $\eta\mu \leq \frac{1}{16}$, $\alpha \leq \frac{1}{3200}$), and $\lambda \geq \frac{\eta\mu}{70} > 0$: indeed $\frac{1+\alpha}{1+\eta\mu} \leq (1 + \alpha)(1 - \eta\mu + (\eta\mu)^2) \leq 1 + \alpha - \frac{7\eta\mu}{8}$ (as $2(\eta\mu)^2 \leq \frac{\eta\mu}{8}$), so

$$\lambda \geq \frac{3\eta\mu}{8} - 16 \cdot \frac{33}{31}\alpha - \alpha \geq \eta\mu\left(\frac{3}{8} - \frac{17.03}{50}\right) \geq \frac{\eta\mu}{70} > 0.$$

Adding (44) (divided by $1 + \eta\mu$), λu_k , and ρ times (46): the coefficient of u_{k-1} is $-\frac{\alpha}{1+\eta\mu} \leq 0$ (dropped); the coefficient of u_k is $\frac{1+\alpha}{1+\eta\mu} + \lambda + 16\rho = q$ by definition of λ ; the coefficient of $\|\Delta_k\|^2$ is $\frac{\alpha(1+2\alpha)}{1+\eta\mu} + 6\rho\alpha^2 = \rho(q - 6\alpha^2) + 6\rho\alpha^2 = \rho q$ by definition of ρ . Hence

$$W_{k+1} \leq q W_k. \quad (48)$$

Theorem C.1 (Inertial linear rate). *Under the assumptions of Theorem 5.3, with prox-center inertia $0 \leq \alpha \leq \alpha_0 = \eta\mu/50$ (and $\beta = \gamma = 0$), the iterates of Algorithm 1 satisfy*

$$\|z_k - z^*\|^2 \leq \frac{33}{32} \left(1 + \frac{\eta\mu}{2}\right)^{-k} \|z_0 - z^*\|^2 \quad (k \geq 0, z_{-1} = z_0).$$

Proof. $u_k \leq W_k \leq q^k W_0$ with $W_0 = (1 + \lambda)u_0 \leq \frac{33}{32}u_0$ (as $\lambda \leq \frac{3\eta\mu}{8} \leq \frac{1}{42} < \frac{1}{32}$), and $q = 1 - \frac{\eta\mu}{2} \leq (1 + \frac{\eta\mu}{2})^{-1}$ since $(1 - x)(1 + x) = 1 - x^2 \leq 1$. $\square$

Remark C.2 (Scope and conservatism). (i) The certified range is tiny: on the SC–SC instance of Section 7 ($\mu_F \approx 0.73$, $\bar{M} \approx 0.78$, $\eta = \mu_F/(16\bar{M}^2)$), $\alpha_0 \approx 1.1 \times 10^{-3}$, four orders of magnitude below the value $\alpha = 0.05$ used in the experiments. Theorem C.1 certifies that *small* inertia preserves R -linear convergence; it does not cover practical parameter values. (ii) The rate $1 - \eta\mu/2$ is a factor 2 slower than the non-inertial $(1 + \eta\mu)^{-1} \approx 1 - \eta\mu$; this reflects the crude forcing budget, not a fundamental limitation. (iii) Extension to $\beta, \gamma \neq 0$: the bias bound becomes $\|B_k\| \leq 3\bar{M}(\|\Delta_{k+1}\| + \|\Delta_k\|)$, contributing the forcing $6\eta\bar{M}\|\Delta_k\|\sqrt{u_k} \leq \frac{1}{4}\|\Delta_k\|^2 + 36\eta^2\bar{M}^2u_k$ (Young, $t = 2$); both terms are absorbed by shrinking α_0 and adjusting ρ by absolute factors. The constants remain explicit and computable; the full double-inertial statement with numerical constants is *not* claimed.

D The embedding lemma, and the status of the $\sqrt{N}$ lower bound

This appendix makes the lower-bound side of the paper precise: what is proved (lower bounds through the convex embedding), what is not (the $\sqrt{N}$ optimality), and exactly what remains to be constructed.

Lemma D.1 (Embedding: convex sums are Minty). *Let $h : \mathbb{R}^d \rightarrow \mathbb{R}$ be convex and C^1 with a unique global minimizer x^* , let $h = \frac{1}{N} \sum_{i=1}^N h_i$ be any decomposition with h_i L -smooth, and consider (1) with $H(x, y) = h(x)$, $g \equiv 0$, and f proper l.s.c. convex with $0 \in \partial f(x^*)$. Then $z^* = (x^*, y^*)$ satisfies the Minty condition (2) for any y^* : for all $z = (x, y)$ and $u \in \partial f(x)$,*

$$\langle F(z), z - z^* \rangle = \langle \nabla h(x) + u, x - x^* \rangle \geq 0,$$

because convexity gives $\langle \nabla h(x), x - x^* \rangle \geq h(x) - h(x^*) \geq 0$ (as $\nabla h(x^*) = 0$) and monotonicity of ∂f with $0 \in \partial f(x^*)$ gives $\langle u, x - x^* \rangle \geq 0$. No per-component convexity is required: the Minty condition constrains only the summed operator.

Lemma D.2 (Convex quadratic zero-chains: Minty, and carrying the convex lower bounds). *The quadratic zero-chain f_T of [11] — a strictly convex quadratic with a unique global minimizer — satisfies the hypotheses of Lemma D.1, hence its embedding is Minty for every exact finite-sum decomposition. The lower bounds that apply to f_T are the convex ε -stationarity bounds of [11, Section 3], which are weaker than ε^{-2} ; the $\Omega(L\Delta/\varepsilon^2)$ bound of [11] is a nonconvex result and does not apply to f_T (Remark D.4).*

Remark D.3 (Verification status of Lemma D.2). The qualitative properties used here (strict convexity of the quadratic chain, its unique minimizer, the $\|\nabla f_T(x)\| > \varepsilon$ property, and the randomized lower bound with its dimension requirement) are the content of [11, Sections 3–4]. The Minty membership is verified by Lemma D.1. The applicable lower bounds are those for convex stationarity, now stated quantitatively in Proposition 4.11 (dimension constants inherited from [38, Theorem 7]).

What survives and what does not. Lemma D.1 gives a *free* reduction: to prove any lower bound for the Minty class it suffices to exhibit a hard function that is convex with a unique minimizer. The $\Omega(L\Delta/\varepsilon^2)$ bound of Lemma D.2 is of this form, which proves the ε^{-2} optimality in Proposition 4.11.

The $\sqrt{N}$ bound does *not* transfer this way. The finite-sum lower bound $\Omega(\sqrt{N}L\Delta/\varepsilon^2)$ of [3] is built on the *nonconvex* zero-chain of [11] (the Ψ – Φ chain), and the Minty condition is neither stated nor expected to hold for it: star-monotonicity $\langle \nabla f(x), x - x^* \rangle \geq 0$ is a global radial condition that fails in general for functions with multiple near-stationary regions, and the small quadratic regularizer added in [3] enforces only local strong convexity at the minimizer, not global star-monotonicity. We therefore do not claim that the construction of [3] lies in the Minty class.

Remark D.4 (Convex hard instances are provably insufficient). Three lower-bound families surround Conjecture D.6, and none reaches it. (i) [39] prove $\Omega(\frac{\Delta}{\varepsilon^2} \min\{n^{3/4}\sqrt{\sigma L}, \sqrt{n}L\})$ for $(-\sigma, L)$ -weakly-convex objectives with $\sigma > 0$; the hard objective is not star-monotone. (ii) [3] reach $\Omega(\sqrt{n}L\Delta/\varepsilon^2)$ on the nonconvex Ψ – Φ chain of [11, 12] (Section 4.2). (iii) for convex L -smooth finite sums, accelerated variance reduction [38] finds ε -stationary points in

$$O\left(N + \sqrt{NLR^2/\varepsilon}\right) \quad (49)$$

evaluations, where $R = \text{dist}(x_0, X^*)$: indeed $\|\nabla f(x)\|^2 \leq 2L(f(x) - f^*)$, so ε -stationarity follows from εR -optimality, achievable in $O(N + \sqrt{NLR^2/(\varepsilon R)})$ evaluations (lower bound [38, Theorem 7], transferred in Proposition 4.11). On instances with $\Delta = \Theta(LR^2)$, (49) is $O(N + \sqrt{N}(L\Delta)^{1/4}\varepsilon^{-1/2})$, which is below the conjectured $\sqrt{N}L\Delta/\varepsilon^2$ whenever $\varepsilon < (L\Delta)^{1/2}$; hence *no convex instance can witness Conjecture D.6*. The hard instance required in (P1) must be genuinely nonconvex yet star-monotone — the radial-log family of Proposition 7.1 is the template — and the construction must be validated in the regime $d \geq cN(L\Delta/\varepsilon^2)^2$, where the dimension-free bound (49) is not in force.

Remark D.5 (Deterministic information splitting is insufficient). Suppose the chain is split across components deterministically: $f = \frac{1}{N} \sum_i f_i$ with f_i depending only on a coordinate group G_i (e.g. the tridiagonal splitting of [38], where each f_i involves one group). A query to f_i then reveals the chain coordinates in G_i — a group of size m — so T coordinate discoveries require at least T/m queries, i.e. a lower bound $\Omega(T/m)$. To match Conjecture D.6, which demands $\Omega(\sqrt{N}T)$, one would need $m \leq 1/\sqrt{N} < 1$: impossible. The $\sqrt{N}$ amplification therefore cannot come from any deterministic coordinate split; it requires *probabilistic* hiding in which each query reveals a full new coordinate with probability $O(1/\sqrt{N})$, exactly the mechanism of [3]. This also explains why the convex route stops at $\sqrt{N}\varepsilon^{-1/2}$ (Proposition 4.11): deterministic splitting is the only hiding available to convex constructions with exact sums.

Conjecture D.6 ($\sqrt{N}$ lower bound for the Minty class). There exists $c > 0$ such that in dimension $d \geq cN(L\Delta/\varepsilon^2)^2$, any randomized first-order method for the class of Definition 1.1 requires $\Omega(\sqrt{N}L\Delta/\varepsilon^2)$ single-component evaluations to output a point with $\mathbb{E} \text{dist}(0, \partial\Psi) \leq \varepsilon$.

Construction program (route to Conjecture D.6). Lemma D.1 simplifies the task relative to [3]: per-realization Minty is automatic as soon as the *sum* is convex with a unique minimizer, so the program is:

(P1) **Base.** A star-monotone but genuinely nonconvex zero-chain f with the gap property $\|\nabla f(x)\| > \varepsilon$ on undiscovered coordinates. Convex chains are provably insufficient (Remark D.4); the radial-log field of Proposition 7.1 shows how star-monotonicity is enforced without convexity, and is the candidate carrier of the chain structure.

- (P2) **Information-hiding decomposition.** An exact decomposition $f = \frac{1}{N} \sum_i f_i$ with $\max_i L_i \leq L$, such that a single component query reveals $O(1/\sqrt{N})$ new chain coordinates in expectation — the “probabilistic zero-chain” mechanism of [3]. Deterministic hiding is provably insufficient (Remark D.5), so this step genuinely requires the probabilistic construction; the truncated chain of Appendix E satisfies the properties its analysis uses (gap, zero-chain, smoothness, star-monotonicity), and transferring it requires verifying that the argument of [3] depends on the chain only through these properties — the one remaining verification, which needs their paper in hand.
- (P3) **Transfer.** Minty of every realization is free by Lemma D.1; the estimator-concentration argument of [3, 41] then yields the $\sqrt{N}$ amplification unchanged.

Item (P2) is a concrete open problem and the only step separating Proposition 4.10 from a fully matching lower bound.

E A star-monotone nonconvex hard instance: the $N = 1$ lower bound

This appendix proves that the ε^{-2} scaling is optimal for the Minty class already at $N = 1$, via an explicit instance that is *simultaneously* star-monotone and nonconvex. The construction modifies the hard instance of [11] by a quadratic centering with two smooth truncations. All constants are explicit; the only inherited constant is c_2 in the randomized dimension requirement [11, Thm. 2, Sec. 5].

Chain functions and truncation. Let Ψ and Φ be as in [11, Lemma 1]. Their established bounds $0 \leq \Psi < e$, $\sup \Psi' = \sqrt{54/e}$ (attained at $(2t - 1)^2 = 3/2$; in particular Ψ'' vanishes there), $0 < \Phi < \sqrt{2\pi e}$, $0 \leq \Phi' \leq \sqrt{e}$, and $\sup |\Psi''| \leq 33$ (a direct single-variable computation), give the first-order Lipschitz constant of the (untruncated) chain

$$\ell_1 = 2e^{3/2} + 2\sqrt{\frac{54}{e}}\sqrt{2\pi e} = 2e^{3/2} + 2\sqrt{108\pi} \leq 45.81. \quad (50)$$

For $M = 10$ let $\chi(t) = \sigma(M + 1 - t)$ with $\sigma(s) = (1 + \exp(\frac{1}{s} - \frac{1}{1-s}))^{-1}$ for $s \in (0, 1)$, $\sigma \equiv 1$ on $[1, \infty)$, $\sigma \equiv 0$ on $(-\infty, 0]$: a C^∞ decreasing cutoff on $[M, M + 1]$ with

$$|\chi'| \leq 6 \quad \text{and} \quad |\chi''| \leq 40, \quad (51)$$

$\chi \equiv 1$ on $(-\infty, M]$, $\chi \equiv 0$ on $[M + 1, \infty)$. Set

$$\tilde{\Psi}(t) = \Psi(t)\chi(t) + e(1 - \chi(t)), \quad \tilde{\Phi}'(t) = \Phi'(t)\chi(|t|), \quad (52)$$

with $\tilde{\Phi}(t) = \int_{-\infty}^t \tilde{\Phi}'(s) ds$. Then $\tilde{\Psi}$ is C^∞ , nondecreasing ($\tilde{\Psi}' = \Psi'\chi + (\Psi - e)\chi' \geq 0$), satisfies $\tilde{\Psi} = \Psi$ on $(-\infty, M]$, $\tilde{\Psi} \equiv e$ on $[M + 1, \infty)$, and $\Psi \leq \tilde{\Psi} \leq e$ pointwise. The cutoff $\chi(|t|)$ is 1 on $[-M, M]$ and vanishes for $|t| \geq M + 1$, so $\tilde{\Phi} = \Phi$ on $[-M, M]$ and $\tilde{\Phi}' \equiv 0$ for $|t| \geq M + 1$; moreover $0 \leq \tilde{\Phi}' \leq \sqrt{e}$ and

$$\begin{aligned} |\tilde{\Psi}''| &= |\Psi''\chi + 2\Psi'\chi' + (\Psi - e)\chi''| \leq 33 + 2\sqrt{\frac{54}{e}} \cdot 6 + e \cdot 40 \leq 200, \\ |\tilde{\Phi}''| &= |\Phi''\chi(|t|) + \Phi'\chi'(|t|)\text{sign}(t)| \leq \sup |\Phi''| + 6\sqrt{e} = 1 + 6\sqrt{e} \leq 12. \end{aligned} \quad (53)$$

where the chain rule is applied separately for $t > 0$ and $t < 0$ (i.e., $\frac{d}{dt}\chi(|t|) = \chi'(|t|)\text{sign}(t)$ with $\chi' \leq 0$), and $|\Phi''(t)| = |t|\Phi'(t) \leq 1$ with equality at $t = 1$. (There is no $\Phi\chi''$ term in $\tilde{\Phi}''$ since χ multiplies Φ' , not Φ .)

Smoothness of the truncated chain. A row of $\nabla^2 \bar{f}_T^{\text{tr}}$ contains two off-diagonal entries of magnitude $|\tilde{\Psi}'\tilde{\Phi}'| \leq (\sqrt{54/e} + e \cdot 6)\sqrt{e} \leq 34.2$ each, one term $|\tilde{\Phi}''\tilde{\Psi}| \leq 12e$, one term $|\tilde{\Psi}''\tilde{\Phi}| \leq 200\sqrt{2\pi e}$, and the first row the additional term $|\tilde{\Phi}''\tilde{\Psi}(1)| \leq 12$; hence

$$\bar{f}_T^{\text{tr}} \text{ is } \ell^{\text{tr}}\text{-smooth with } \ell^{\text{tr}} := 2 \cdot 34.2 + 12e + 200\sqrt{2\pi e} + 12 \leq 10^3. \quad (54)$$

The instance. Let $L_f, \varepsilon, \Delta > 0$ satisfy $L_f \Delta \geq 10^9 \varepsilon^2$. Set

$$f(x) = a \bar{f}_T^{\text{tr}}(bx) + \frac{L_f}{32} \|x - x^T\|^2, \quad \bar{f}_T^{\text{tr}}(y) = -\tilde{\Psi}(1)\tilde{\Phi}(y_1) + \sum_{i=2}^T [\tilde{\Psi}(-y_{i-1})\tilde{\Phi}(-y_i) - \tilde{\Psi}(y_{i-1})\tilde{\Phi}(y_i)], \quad (55)$$

with

$$a = \frac{16 \varepsilon^2 \ell^{\text{tr}}}{15 L_f}, \quad b = \frac{15 L_f}{16 \varepsilon \ell^{\text{tr}}}, \quad x^T = \frac{M+2}{b} \mathbf{1} = \frac{192 \varepsilon \ell^{\text{tr}}}{15 L_f} \mathbf{1}. \quad (56)$$

Then $ab = \varepsilon$ and $ab^2 \ell^{\text{tr}} = \frac{15}{16} L_f$, so f is L_f -smooth; with $T = \lfloor 10^{-9} L_f \Delta / \varepsilon^2 \rfloor$ and $d = T$:

Lemma E.1 (Exact global star-monotonicity). $\nabla f(x^T) = 0$ and $\langle \nabla f(x), x - x^T \rangle \geq 0$ for all $x \in \mathbb{R}^d$; hence the embedded instance $(H(x, y) = f(x), g \equiv 0, \text{ any exact decomposition})$ satisfies Minty MVI at (x^T, y^*) by Lemma D.1.

Proof. Every summand of $\partial \bar{f}_T^{\text{tr}} / \partial y_j$ is nonpositive ($\tilde{\Psi}, \tilde{\Psi}', \tilde{\Phi}, \tilde{\Phi}' \geq 0$), hence ∇f is the sum of a componentwise nonpositive vector and $\frac{L_f}{16}(x - x^T)$. At x^T , $bx^T = (M+2)\mathbf{1}$ has all coordinates beyond the truncations: $\tilde{\Phi}'(\pm(M+2)) = 0$ (as $\chi(M+2) = 0$) and $\tilde{\Psi}'(\pm(M+2)) = 0$ (as $\tilde{\Psi} \equiv e$ beyond $M+1$), so all summands of each partial vanish and $\nabla f(x^T) = 0$. For $x \in \mathbb{R}^d$ and each coordinate j : if $x_j \leq x_j^T$ then both $\nabla_j f$ and $x_j - x_j^T$ are ≤ 0 ; if $x_j > x_j^T$ then $bx_j > M+1$ and $\partial \bar{f}_T^{\text{tr}} / \partial y_j = 0$ exactly — the bracket- j terms contain $\tilde{\Phi}'(\pm bx_j) = 0$, and the bracket- $(j+1)$ terms contain $\tilde{\Psi}'(\pm bx_j) = 0$, independently of the neighboring coordinates — so $\nabla_j f(x_j - x_j^T) = \frac{L_f}{16}(x_j - x_j^T)^2 \geq 0$. Summing gives the claim. $\square$

Lemma E.2 (Gap property of the truncated chain, self-contained). For every $y \in \mathbb{R}^d$ with $y_T = 0$, $\|\nabla \bar{f}_T^{\text{tr}}(y)\| > 1$.

Proof. Every summand of each partial $\partial \bar{f}_T^{\text{tr}} / \partial y_j$ is nonpositive (Lemma E.1), so $\nabla \bar{f}_T^{\text{tr}}$ is componentwise nonpositive. We use: $\tilde{\Phi}'(0) = \sqrt{e}$ and $\tilde{\Psi}'(0) = 0$; $\tilde{\Psi} \geq \Psi$ pointwise with $\tilde{\Psi} = \Psi$ on $[-M, M]$ and $\tilde{\Psi}(t) \geq \Psi(10) > 2.71$ for $t \geq 10$; $\tilde{\Psi}'(t) = 0$ for $t \geq M+1$; $\tilde{\Phi} = \Phi$ on $[-M, M]$ with Φ increasing and $\Phi'(t) = \Phi'(-t)$; Ψ increasing and Ψ' increasing on $[\frac{1}{2}, \frac{1+\sqrt{3/2}}{2}]$ (its maximum is attained at $(2t-1)^2 = 3/2$). Suppose, for contradiction, that $\max_j |\partial \bar{f}_T^{\text{tr}} / \partial y_j| \leq 1$.

Frontier. Let p be the last index with $y_p \neq 0$ (if none, $\partial_1 = -\tilde{\Psi}(1)\tilde{\Phi}'(0) = -\sqrt{e}$). Since $y_{p+1} = 0$ and $\tilde{\Psi}'(0) = 0$,

$$-\frac{\partial \bar{f}_T^{\text{tr}}}{\partial y_{p+1}} = [\tilde{\Psi}(-y_p) + \tilde{\Psi}(y_p)]\sqrt{e} \geq \Psi(|y_p|)\sqrt{e},$$

so $|y_p| < \tau_0$, where $\Psi(\tau_0) = e^{-1/2}$, i.e. $\tau_0 = \frac{1}{2}(1 + \sqrt{2/3}) < 0.909$.

Downward step. Let $|y_j| < 0.91$ for some $2 \leq j \leq p$ (true for $j = p$). We show that either the assumed bound is violated or $|y_{j-1}| \leq \tau_1$, where τ_1 is the unique solution of $\Psi'(\tau_1) = 1/\Phi(-\tau_0)$ on $[\frac{1}{2}, t_1]$ (Ψ' is increasing there): Φ is the chain function $\Phi(t) = \sqrt{e} \int_{-\infty}^t e^{-s^2/2} ds$ of [11, Lemma 1], with the prefactor $\sqrt{e}$, so $\Phi(-\tau_0) \approx 0.752$, $1/\Phi(-\tau_0) \approx 1.330$, and since $\Psi'(0.74) \approx 1.28 < 1.330 < 1.59 \approx \Psi'(0.75)$ the solution exists with $\tau_1 \approx 0.742$. Dropping nonnegative terms,

$$-\frac{\partial \bar{f}_T^{\text{tr}}}{\partial y_{j-1}} \geq \tilde{\Psi}'(-y_{j-1})\tilde{\Phi}(-y_j) + \tilde{\Psi}'(y_{j-1})\tilde{\Phi}(y_j) \geq \Phi(-|y_j|)\Psi'(|y_{j-1}|) \geq \Phi(-\tau_0)\Psi'(|y_{j-1}|),$$

the middle inequality using $\min(\tilde{\Phi}(-y_j), \tilde{\Phi}(y_j)) = \Phi(-|y_j|) \geq \Phi(-\tau_0)$ and $\tilde{\Psi}'(-y_{j-1}) + \tilde{\Psi}'(y_{j-1}) = \Psi'(|y_{j-1}|)$ for $|y_{j-1}| \in [\frac{1}{2}, 0.91]$. Two bands of $|y_{j-1}| > 0.91$ are excluded: $(0.91, t_1]$ with $t_1 = \frac{1+\sqrt{3/2}}{2}$ (where Ψ' is increasing) by the display, since $\Psi'(0.91) \Phi(-\tau_0) \approx 3.35 > 1$; and $[t_1, \infty)$ by one of the two terms of $-\partial_j$, namely $-\partial_j \geq \tilde{\Psi}(|y_{j-1}|) \tilde{\Phi}'(|y_j|) \geq \Psi(t_1) \Phi'(\tau_0) \approx 1.523 > 1$ (using $\tilde{\Psi} \geq \Psi$, which covers the belt $(M, M+1]$ and the dead zone, where $\tilde{\Psi} \geq \Psi(10) > 2.71$). Hence $|y_{j-1}| \leq 0.91$, and then $|y_{j-1}| > \tau_1$ would give $-\partial_{j-1} > 1$ (as Ψ' is increasing on $[\frac{1}{2}, t_1] \ni \tau_1$): excluded. So $|y_{j-1}| \leq \tau_1$.

Termination. Iterating the step down to $j = 2$ gives $|y_1| \leq \tau_1$, and then

$$-\frac{\partial \bar{f}_T^{\text{tr}}}{\partial y_1} = \tilde{\Psi}(1)\tilde{\Phi}'(y_1) + \tilde{\Psi}'(-y_1)\tilde{\Phi}(-y_2) + \tilde{\Psi}'(y_1)\tilde{\Phi}(y_2) \geq \Phi'(\tau_1) > 1,$$

since $\Phi'(\tau_1) = \sqrt{e} e^{-\tau_1^2/2} \approx 1.25$ at $\tau_1 \approx 0.742$. Contradiction. $\square$

Theorem E.3 (ε^{-2} lower bound for the Minty class at $N = 1$). *The function f of (55) satisfies: (i) f is L_f -smooth and nonconvex: at $y = (0.76, M+2, 0, \dots, 0)$ the leading 2×2 block of $\nabla^2 \bar{f}_T^{\text{tr}}(y)$ is $\text{diag}(H_{11}, H_{22})$ with $H_{12} = 0$ and*

$$H_{11} = -\tilde{\Phi}''(0.76)\tilde{\Psi}(1) - \tilde{\Psi}''(-0.76)\tilde{\Phi}(-(M+2)) - \tilde{\Psi}''(0.76)\tilde{\Phi}(M+2) \leq 12 - 32 \cdot 4.1 \leq -100,$$

(a direct computation gives $\tilde{\Psi}''(0.76) = \Psi''(0.76) \geq 32$, as $0.76 < M$ and $\chi \equiv 1$ near 0.76 ; $\tilde{\Phi}(M+2) \approx 4.133 \geq 4.1$; and $\tilde{\Psi}''(-0.76) = 0$ as Ψ is flat on $(-\infty, \frac{1}{2}]$), so $\lambda_{\min}(\nabla^2 f) \leq \frac{15L_f}{16\ell^{\text{tr}}}(-100) + \frac{L_f}{16} \leq -\frac{L_f}{32} < 0$; (ii) f is star-monotone at x^T (Lemma E.1), so the embedded instance belongs to the Minty class; (iii) $\|\nabla f(x)\| > \varepsilon$ for every x whose chain profile is undiscovered; (iv) $f(0) - \inf f \leq \Delta$. Consequently any zero-respecting first-order method requires at least $T = \lceil 10^{-9} L_f \Delta / \varepsilon^2 \rceil$ gradient evaluations to find an ε -stationary point of a Minty-class instance, and the same $\Omega(L_f \Delta / \varepsilon^2)$ bound holds for any randomized method in dimension $d \geq c_2 T^2 \log T$ [11, Thm. 2], with c_2 the (underived) numerical constant therein.

Proof. Smoothness and nonconvexity. The chain part is $ab^2\ell^{\text{tr}} = \frac{15}{16}L_f$ -smooth by (54) and the centering adds $\frac{L_f}{16}$, giving L_f -smoothness. For claim (i) we bound $\lambda_{\min}$ below by that of the leading 2×2 principal submatrix (Cauchy interlacing). With $y = (0.76, M+2, 0, \dots, 0)$: $H_{12} = 0$ because $\tilde{\Phi}'(\pm(M+2)) = 0$ and $\tilde{\Psi}'(-0.76) = 0$; $H_{22} = 0$ because $\tilde{\Psi}(-0.76) = 0$, $\tilde{\Phi}''(\pm(M+2)) = 0$ and $\tilde{\Psi}''(\pm(M+2)) = 0$; and $H_{11} = -\tilde{\Phi}''(0.76) - \tilde{\Psi}''(0.76)\tilde{\Phi}(M+2) \leq 12 - 32 \cdot 4.1 \leq -100$.

Value gap. For any y ,

$$\bar{f}_T^{\text{tr}}(y) \geq -\sqrt{2\pi e} - (T-1)e\sqrt{2\pi e} \geq -e^{3/2}\sqrt{2\pi}T,$$

since $-\tilde{\Psi}(1)\tilde{\Phi}(y_1) \geq -\sqrt{2\pi e}$ and each bracket is $\geq -e\sqrt{2\pi e}$; while $\bar{f}_T^{\text{tr}}(0) = -\tilde{\Psi}(1)\tilde{\Phi}(0) = -\sqrt{e\pi/2}$. Hence $\bar{f}_T^{\text{tr}}(0) - \inf \bar{f}_T^{\text{tr}} \leq e^{3/2}\sqrt{2\pi}T$. Since $\inf f \geq a \inf \bar{f}_T^{\text{tr}}$ (the centering is nonnegative),

$$\begin{aligned} f(0) - \inf f &\leq a e^{3/2}\sqrt{2\pi}T + \frac{L_f}{32}\|x^T\|^2 \\ &= \left(\frac{16\ell^{\text{tr}}e^{3/2}\sqrt{2\pi}}{15} + \frac{144}{32}\left(\frac{16}{15}\right)^2(\ell^{\text{tr}})^2 \right) \frac{\varepsilon^2 T}{L_f} \leq \frac{5.2 \times 10^6 \varepsilon^2 T}{L_f} \leq 0.0052 \Delta, \end{aligned}$$

using $\ell^{\text{tr}} = 10^3$ and $T \leq 10^{-9} L_f \Delta / \varepsilon^2$.

Gap property. Wherever the chain profile is undiscovered, $\|\nabla \bar{f}_T^{\text{tr}}\| > 1$ by Lemma E.2 (whose proof is self-contained and covers the truncation belt), hence $\|ab \nabla \bar{f}_T^{\text{tr}}(bx)\| > ab = \varepsilon$; both this vector and $\frac{L_f}{16}(x - x^T)$ are componentwise nonpositive on undiscovered coordinates (where $x_j = 0 < x_j^T$), so they add rather than cancel and $\|\nabla f(x)\| > \varepsilon$. If instead all active coordinates

lie beyond the truncation, the chain gradient vanishes and $\|\nabla f(x)\| \geq \frac{L_f}{16}(M+2)b^{-1} = \frac{12.8\varepsilon\ell^{\text{tr}}}{16} \geq 800\varepsilon$.

Lower bound. Zero-respecting algorithms reveal at most one chain coordinate per query [11, Observation 2] and the gap property forces all T coordinates, giving T queries. The randomized statement follows from [11, Thm. 2]: the rotation preserves star-monotonicity ($\langle U\nabla f(U^\top x), x - Ux^T \rangle = \langle \nabla f(y), y - x^T \rangle \geq 0$), the gap, smoothness, and value gap. $\square$

Remark E.4 (Explicit versus inherited constants). Every constant in Theorem E.3 is explicit: the scaling (56) with $\ell^{\text{tr}} = 10^3$; the budget constant 10^{-9} ; the centering $L_f/16$ with $x^T = \frac{192\varepsilon\ell^{\text{tr}}}{15L_f}\mathbf{1}$; the cutoff bounds (51) and truncation bounds (53); the nonconvexity margin $L_f/32$ witnessed at $(0.76, M+2, 0, \dots, 0)$ with $\Psi''(0.76) \geq 32$ and $\tilde{\Phi}(M+2) \geq 4.1$; and the tension margin 800ε . The single inherited constant is c_2 in the randomized dimension requirement $d \geq c_2 T^2 \log T$ [11, Thm. 2, Sec. 5]. The budget constant 10^{-9} is not optimized: it is dominated by the crude truncation-belt terms entering ℓ^{tr} (the $(\Psi - e)\chi''$ term) and by the $(M+2)^2$ centering contribution; a smaller M , a cutoff with smaller belt derivatives, and a rebalanced centering would improve it by orders of magnitude. The qualitative conclusion, an $\Omega(L_f\Delta/\varepsilon^2)$ lower bound on a Minty-class instance, is unaffected.

Remark E.5 (Status and extension to $\sqrt{N}$). Theorem E.3 is, to our knowledge, the first $\Omega(L_f\Delta/\varepsilon^2)$ lower bound proved on a star-monotone (Minty-satisfying) nonconvex instance; it makes the ε^{-2} dependence of Theorem 4.8 and Proposition 4.10 class-optimal at $N = 1$. The remaining step toward Conjecture D.6 is the information-hiding decomposition of (55) across N components (items (P2)–(P3) of Appendix D): the centering decomposes as $\sum_i \frac{L_f}{32N} \|x - x^T\|^2$ and the truncated chain as edge pieces with the per-component radial sign structure of Lemma E.1 preserved.