Gradient-Free Methods for Convex Stochastic Optimization with Functional Constraints: Vector Oracles, Acceleration, and Dual Reductions*

Vadim D. Abronin[†2], Alexander V. Gasnikov[‡1,2,3], and Darina M. Dvinskikh[§4]

[1]Innopolis University, 1 Universitetskaya St., Innopolis, Republic of Tatarstan, 420500, Russian Federation
[2]Moscow Institute of Physics and Technology, 9 Institutskiy per., Dolgoprudny, Moscow Region, 141701, Russian Federation
[3]Kharkevich Institute for Information Transmission Problems of the Russian Academy of Sciences, 19 Bolshoy Karetny per., bldg. 1, Moscow, 127051, Russian Federation
[4]HSE University, 11 Pokrovsky Boulevard, Moscow, 109028, Russian Federation

Abstract

We consider a d-dimensional convex stochastic problem with m functional constraints in which the algorithm observes only noisy function values (the zeroth-order, gradient-free setting). The basic primitive is a paired vector query: the oracle returns the objective and all m constraints at two nearby points $x \pm \gamma u$, evaluated on one and the same random sample, so that the data noise cancels in the difference. Feasibility is handled by shifted smoothing: every smoothed constraint is lowered by an explicit upper bound on its smoothing bias, which preserves the feasible set of the original problem and the Slater margin. To obtain an ε-accurate solution ($\varepsilon > 0$), batched Mirror–Prox uses $\mathcal{O}(d/\varepsilon^2)$ zeroth-order vector queries in both the smooth and Lipschitz nonsmooth regimes; averaging the noise of the constraint values (measured in the ℓ_∞ norm paired with the dual variable) adds the factor $1 + \log(2m)$. We propose the accelerated method PB–ACGD-S (phase-banked ACGD with sliding), in which all random queries of an outer phase are grouped into pre-drawn independent packets — banks — and can therefore be issued in parallel. Its sequential depth — the number of oracle rounds that must be issued one after another — is $\mathcal{O}(\varepsilon^{-1/2})$ in the smooth regime and $\mathcal{O}(d^{1/4}/\varepsilon)$ after nonsmooth smoothing when the oracle domain permits the chosen radius, with total vector work $\widetilde{\mathcal{O}}(d/\varepsilon^2)$, where $\widetilde{\mathcal{O}}$ hides logarithmic factors and fixed problem scales. We also give a Vaidya dual route for small m and a specialization to known affine constraints. The lower bounds match the leading dependence on d and ε in the paired-second-moment class, i.e., the oracle class in which both points of a pair are evaluated on one sample and the random Lipschitz modulus has a bounded second moment.

Keywords: stochastic optimization; zeroth-order oracle; functional constraints; randomized smoothing; Mirror–Prox; accelerated methods; Vaidya method.

*The research was supported by Russian Science Foundation (project No. 21-71-30005-), https://rscf.ru/en/project/21-71-30005/.

[†]abronin.vd@phystech.edu
[‡]gasnikov@yandex.ru
[§]dmdvinskikh@hse.ru

1 Introduction

We study stochastic convex programs with functional constraints using only noisy function values:

$$\min_{\substack{x\in X\subset\mathbb{R}^d \\ X \text{ compact and convex}}} \underbrace{f_0(x) := \mathbb{E}[F_0(x,\xi)]}_{\text{expected objective}} \quad \text{subject to} \quad \underbrace{f_i(x) := \mathbb{E}[F_i(x,\xi)] \le 0}_{m \text{ expected-value constraints}}, \qquad i = 1,\dots,m. \tag{1}$$

Here $d \ge 1$ and $m \ge 1$ are integers, x is the d-dimensional decision vector, ξ is the random sample, and $F_0, F_1, \dots, F_m$ are observable integrands with finite expectations. For target accuracy $\varepsilon > 0$, an expected ε-solution has objective residual and maximum positive constraint violation at most ε; the exact criterion is given in Definition 1.

First-order stochastic methods for functional constraints include switching, cooperative, primal-dual, parameter-free, and sampled-constraint schemes [Bayandina et al., 2018, Lan, Zhou, 2020, Boob et al., 2023, Deng et al., 2024, Rajoriya et al., 2025, Sanyal et al., 2025]. Existing zeroth-order constrained methods cover smooth stochastic and composite models [Nguyen, Balasubramanian, 2023, Li et al., 2022]. In componentwise accounting, the SZO-ConEx method of Nguyen and Balasubramanian [Nguyen, Balasubramanian, 2023] (stochastic zeroth-order constraint extrapolation) has order $\mathcal{O}((m+1)d/\varepsilon^2)$ for convex problems. Gradient-free saddle methods give a related route through monotone variational inequalities [Beznosikov et al., 2020]. Safe and long-term constraint models impose different feasibility or feedback requirements [Usmanova et al., 2019, Yu, Neely, 2016].

Complexity depends on what one oracle call returns. One *paired vector query* uses the same sample at two perturbed points and returns $\big(F(x+\gamma u,\xi), F(x-\gamma u,\xi)\big)$, where $F = (F_0,\dots,F_m)$, u is a random unit direction, and $\gamma > 0$ is the smoothing radius (Definition 2 below states this model formally). The same direction estimates every row of the constraint Jacobian. Thus the number of vector rounds need not grow linearly with m, although each returned vector contains $m+1$ values and has linear storage and processing cost. A direct componentwise implementation uses $m+1$ scalar calls per vector value; sampling one constraint at a time defines a different oracle model and requires a different analysis.

Ordinary randomized smoothing may destroy feasibility. We instead shift each smoothed constraint by an explicit bias bound, which preserves the feasible set and the quantitative Slater margin. Batched stochastic Mirror Prox then has paired zeroth-order vector complexity $\mathcal{O}(d/\varepsilon^2)$ in both the smooth and Lipschitz nonsmooth regimes. Averaging finite-second-moment noise in the dual ℓ_∞ block contributes the factor $1+\log(2m)$; a dimension-free $1/r$ batching law is false for this geometry without stronger assumptions, where r is the batch size.

To reduce sequential depth — the number of oracle rounds that must be issued one after another because later query points depend on earlier answers (the query-accounting convention at the end of Section 2 makes this precise) — we place fresh parallel *sample banks* inside the ACGD-S (accelerated constrained gradient descent with sliding) recurrence of Zhang and Lan [Zhang, Lan, 2025]. A bank is a packet of independent oracle queries drawn at an already known point before the inner loop starts; Definition 5 makes this formal. The resulting PB-ACGD-S (phase-banked ACGD-S) method has sequential depth $\mathcal{O}(\varepsilon^{-1/2})$ in the smooth regime and $\mathcal{O}(d^{1/4}/\varepsilon)$ after nonsmooth smoothing when the available oracle neighborhood does not restrict the smoothing radius. Its leading vector work is $\widetilde{\mathcal{O}}(d/\varepsilon^2)$, where $\widetilde{\mathcal{O}}$ suppresses logarithmic factors and fixed problem scales. For small m, we also study a regularized dual reduction with inexact Vaidya cuts. Known affine constraints admit a separate specialization that distinguishes objective-value calls from matrix products. Lower-bound reductions recover the leading dependence on d and ε in the paired-second-moment class, while the experiments report vector work and sequential depth separately.

2 Problem, accuracy criteria, and oracle models

Throughout, $\widetilde{\mathcal{O}}(\cdot)$ suppresses logarithmic factors in accuracy, confidence, dimension, and explicitly displayed problem-scale ratios, but never suppresses polynomial factors in the number of constraints m. Write $[m] := \{1, \dots, m\}$. We use the shorthand

$$\kappa_m := 1 + \log(2m) \tag{2}$$

for the logarithmic averaging factor that appears when finite-second-moment ℓ_∞ noise is averaged. Its origin is geometric: when r independent centered vectors are averaged in a Euclidean norm, the second moment of the average decreases as $1/r$ with no extra factor, whereas in the ℓ_∞ norm on $\mathbb{R}^m$ the same $1/r$ law holds only up to a factor of order $\log m$ (Lemma 3 below). In Banach-space language this reflects the type-2 constant of order $\sqrt{\log m}$ of ℓ_∞^m; we do not use that terminology further and simply call κ_m the ℓ_∞-*averaging factor*.

2.1 Basic assumptions

Retain the compact convex set X from the introduction and let $\bar{\gamma} > 0$. Write $B_2^d := \{v \in \mathbb{R}^d : \|v\|_2 \le 1\}$ for the Euclidean unit ball and

$$D_X := \sup_{x,z \in X} \|x - z\|_2 < \infty, \qquad X_{\bar{\gamma}} := X + \bar{\gamma} B_2^d$$

for a known oracle neighborhood. The neighborhood is needed for a simple reason: the finite-difference points $x \pm \gamma u$ and the smoothing points $x + \gamma v$ leave X by a distance up to γ, so the oracle must be defined not only on X but on its $\bar{\gamma}$-enlargement. All finite-difference queries use radii $0 < \gamma \le \bar{\gamma}$. To include the boundary choice $\gamma = \bar{\gamma}$ without a hidden differentiability issue, we adopt the convention that the oracle and the regularity assumptions below extend to an open neighborhood of $X_{\bar{\gamma}}$. If only the closed neighborhood is available, every displayed boundary choice may instead be replaced by an arbitrarily smaller radius.

Assumption 1 (Convexity and pathwise Lipschitz continuity)**.** For every $i \in \{0, \dots, m\}$ and almost every ξ, the mapping $F_i(\cdot, \xi)$ is convex on $X_{\bar{\gamma}}$ and

$$|F_i(x, \xi) - F_i(z, \xi)| \le M_i \|x - z\|_2, \qquad x, z \in X_{\bar{\gamma}}. \tag{3}$$

The constants M_i are deterministic and known upper bounds.

The pathwise form of (3) is stronger than Lipschitz continuity of the expectation. It is exactly what keeps the two-point estimator variance independent of the smoothing radius when the same sample is used at the two perturbed points.

Assumption 2 (Paired-value second-moment alternative)**.** As an alternative to Assumption 1, for almost every ξ let $F_i(\cdot, \xi)$ be convex on $X_{\bar{\gamma}}$ and admit a random Lipschitz modulus $\mathsf{M}_i(\xi)$,

$$|F_i(x, \xi) - F_i(z, \xi)| \le \mathsf{M}_i(\xi) \|x - z\|_2, \qquad \mathbb{E}\mathsf{M}_i(\xi)^2 \le \overline{M}_i^2.$$

All paired evaluations use the same draw ξ. We assume the expectations defining f_i are finite.

Under Assumption 2, Jensen's inequality gives the same smoothing-bias bounds with M_i replaced by $\widetilde{M}_i := \mathbb{E}\mathsf{M}_i \le \overline{M}_i$. The corresponding dual radius and localized second-moment aggregate are defined after the Slater-based dual localization (14) in Section 3. This alternative is also the class used for the lower-bound comparison; no minimax claim for the stricter deterministic pathwise constants is inferred from that comparison. In all subsequent bias and localization formulas, M_i denotes the deterministic bound from Assumption 1 when that assumption is used, and denotes $\widetilde{M}_i$ in the alternative paired-moment branch. Thus the displayed definition of R below is well-defined in either branch. Second-moment estimates in the alternative branch retain the random moduli M_i. In either branch, set $M_g := \max_{1 \le i \le m} M_i$.

Assumption 3 (Optional smooth regime)**.** In the smooth regime, $F_i(\cdot, \xi)$ is differentiable and has an L_i-Lipschitz gradient almost surely on $X_{\bar{\gamma}}$. We set $L_g := \max_{1 \le i \le m} L_i$.

Assumption 4 (Quantitative Slater condition)**.** There exist a known point $x^{\mathrm{s}} \in X$ and a known margin $\rho > 0$ such that

$$f_i(x^{\mathrm{s}}) \le -\rho, \qquad i \in [m]. \tag{4}$$

Assumption 4 is used to construct an a priori bounded dual domain. It may be replaced by any other explicit multiplier bound. Let $x^\star$ be an optimal solution of (1), and let $f_0^\star = f_0(x^\star)$.

Definition 1 (Accuracy)**.** Set $g(x) := (f_1(x), \dots, f_m(x))$. A random point $\widehat{x} \in X$ is an expected ε-solution in the maximum-violation metric if

$$\mathbb{E}[f_0(\widehat{x}) - f_0^\star] \le \varepsilon, \qquad \mathbb{E}\,\|[g(\widehat{x})]_+\|_\infty \le \varepsilon. \tag{5}$$

The objective residual is not replaced by its absolute value: an infeasible point can have an objective value below $f_0^\star$. For high-probability guarantees, $\beta \in (0, 1)$ denotes the failure probability and both inequalities are required on an event of probability at least $1 - \beta$.

The ℓ_∞ violation is natural for a dual ℓ_1 ball and prevents an artificial $\sqrt{m}$ factor. An ℓ_2 violation can be used instead, but then either the dual geometry or aggregate Lipschitz constants generally contain an explicit dependence on m.

2.2 Three oracle models

The query model is part of the theorem, not an implementation detail.

Definition 2 (Vector-value two-point oracle)**.** Given two points $x^+, x^- \in X_{\bar{\gamma}}$, the common-randomness vector oracle draws one sample ξ and returns

$$\mathcal{O}_{\mathrm{vec}}(x^+, x^-; \xi) = \big(F(x^+, \xi), F(x^-, \xi)\big), \qquad F = (F_0, \dots, F_m). \tag{6}$$

Each vector $F(x, \xi) \in \mathbb{R}^{m+1}$ is counted as one vector-value call. Thus (6) costs two vector calls and outputs $2(m + 1)$ scalars.

Definition 3 (Componentwise scalar oracle)**.** A componentwise call returns one scalar $F_i(x, \xi)$ for a specified index i. If the oracle exposes a common sample identifier, a full vector value can be reproduced exactly by querying all $m + 1$ components with that identifier. Without such an interface, querying all components still costs $m + 1$ scalar calls but may induce a different joint noise law. In either direct implementation of an algorithm that evaluates every component at every queried point,

$$Q_{\mathrm{cmp}} = (m + 1) Q_{\mathrm{vec}}. \tag{7}$$

This is an accounting identity for direct simulation, not an information-theoretic lower bound for every componentwise algorithm.

Definition 4 (Sampled-constraint oracle)**.** A sampled-constraint oracle draws an index I and returns information only about F_I. Its cost can be independent of m, but the algorithm observes a different random object from Definitions 2–3. Rates in this model require a sampling distribution and usually an exact-penalty, regularity, or finite-sum argument; see, for example, recent first-order methods [Rajoriya et al., 2025, Sanyal et al., 2025].

Even when Q_{vec} has no explicit linear factor in m, writing, storing, and processing the returned vector costs at least linearly many arithmetic operations in m. The localized dual radius and the noise parameters can also depend on m. Accordingly, we use the literal statement "no linear multiplicative dependence on m in the number of vector oracle rounds." The physical accounting units are collected in Table 1.

Query-accounting convention. Four units are used throughout. A *vector-value call* is one evaluation of the full vector $F(x,\xi) \in \mathbb{R}^{m+1}$. A *round* (synonym: *stage*) is a group of vector-value calls whose query points are known simultaneously and which can therefore be issued in parallel. The *adaptive (sequential) depth* N_{seq} is the number of consecutive rounds: the query points of the next round may depend on the answers of the previous one, so rounds cannot be merged. The *work* Q_{vec} is the total number of vector-value calls over all rounds. In the language of parallel computing, depth is the running time on an idealized parallel machine, and work is the total operation count. Table 1 fixes the cost of the recurring objects in these units.

2.3 Noise in function values

The baseline two-point model is the same-sample, or canceling-noise, model in (6). Suppose instead that an observation is $F_i(x,\xi) + \zeta_i$, where ζ_i has zero mean and variance at most σ_{val}^2, and independent noises are used at $x + \gamma u$ and $x - \gamma u$ for a unit direction u. The central difference then contains the additional second-moment term

$$\frac{d^2 \sigma_{\mathrm{val}}^2}{2\gamma^2}. \tag{8}$$

This term diverges as $\gamma \downarrow 0$ and is added explicitly to the variance bounds below. Non-canceling value noise is classical in bandit and zeroth-order optimization [Flaxman et al., 2005, Bach, Perchet, 2016, Akhavan et al., 2020]; our independent-noise corollary is an upper bound for the paired central-difference construction, not a new minimax claim.

3 Shifted smoothing and simultaneous vector estimation

3.1 Bias-corrected smoothing

Let v be uniform on the Euclidean unit ball B_2^d. Define

$$f_{i,\gamma}(x) := \mathbb{E}_v[f_i(x + \gamma v)] = \mathbb{E}_{v,\xi}[F_i(x + \gamma v, \xi)]. \tag{9}$$

For the nonsmooth regime let $b_i(\gamma) = \gamma M_i$. Under Assumption 3, the sharper choice $b_i(\gamma) = L_i \gamma^2/2$ is valid. The shifted smoothed constraints are

$$h_{i,\gamma}(x) := f_{i,\gamma}(x) - b_i(\gamma), \qquad h_\gamma = (h_{1,\gamma}, \dots, h_{m,\gamma}). \tag{10}$$

Lemma 1 (Smoothing and feasibility preservation)**.** *Under Assumption 1, for every $x, z \in X$,*

$$f_i(x) \le f_{i,\gamma}(x) \le f_i(x) + b_i(\gamma), \qquad \|\nabla f_{i,\gamma}(x) - \nabla f_{i,\gamma}(z)\|_2 \le L_{i,\gamma} \|x - z\|_2, \tag{11}$$

where one may take $L_{i,\gamma} \le c_{\mathrm{sm}} \sqrt{d}\, M_i/\gamma$ with a universal constant c_{sm}. Under Assumption 3, one may instead take $L_{i,\gamma} \le L_i$ and $b_i(\gamma) = L_i \gamma^2/2$. Every point feasible for (1) is feasible for

$$\min_{x \in X} f_{0,\gamma}(x) \quad \text{subject to} \quad h_{i,\gamma}(x) \le 0, \quad i \in [m]. \tag{12}$$

In particular, the Slater point in (4) satisfies $h_{i,\gamma}(x^{\mathrm{s}}) \le -\rho$.

The last statement is the reason for the shift. Without it, an active original constraint can become positive after smoothing, and the optimal solution of the original problem need not be feasible for the surrogate.

3.2 A priori dual radius and restricted gap

Let $x^\star_\gamma$ solve (12), set $f^\star_{0,\gamma} := f_{0,\gamma}(x^\star_\gamma)$, and let $y^\star_\gamma$ be an optimal multiplier. Strong duality and the Slater point give

$$\rho \left\|y^\star_\gamma\right\|_1 \le f_{0,\gamma}(x^{\mathrm{s}}) - f^\star_{0,\gamma} \le M_0 D_X. \tag{13}$$

We therefore set

$$R := 1 + \frac{M_0 D_X}{\rho}, \qquad Y_R := \{y \in \mathbb{R}^m_+ : \|y\|_1 \le R\}. \tag{14}$$

Set $M_R := M_0 + R M_g$. Under Assumption 2, also set

$$\overline{M}^2_R := \sup_{y \in Y_R} \mathbb{E}\left(\mathsf{M}_0 + \sum_{i=1}^m y_i \mathsf{M}_i\right)^2 < \infty. \tag{15}$$

The finiteness follows from $m < \infty$, the coordinatewise second-moment bounds, and $\|y\|_1 \le R$. Lemmas 1–2 and the Mirror–Prox bounds extend to the paired-moment branch by replacing the deterministic aggregate M^2_R with $\overline{M}^2_R$. The margin of one between R and the multiplier bound is convenient for converting a saddle gap into feasibility.

Remark 1 (Normalization of the dual margin). The additive constant 1 in (14) is a normalization that fixes the exchange rate between the restricted gap and the violation metric in Theorem 1. More generally, one may set $R = \kappa + M_0 D_X/\rho$ for any $\kappa > 0$; the same proof gives $\|[h_\gamma(\widehat{x})]_+\|_\infty \le \mathrm{Gap}_{R,\gamma}(\widehat{z})/\kappa$. Rescaling the constraints rescales their Lipschitz constants, the Slater margin, the multipliers, and the chosen violation units consistently.

Define

$$\Phi_\gamma(x, y) := f_{0,\gamma}(x) + \langle y, h_\gamma(x)\rangle, \qquad (x, y) \in X \times Y_R, \tag{16}$$

and the restricted gap at $\widehat{z} = (\widehat{x}, \widehat{y})$ by

$$\mathrm{Gap}_{R,\gamma}(\widehat{z}) := \sup_{x \in X,\, y \in Y_R} \left\{\Phi_\gamma(\widehat{x}, y) - \Phi_\gamma(x, \widehat{y})\right\}. \tag{17}$$

A supremum over the unbounded orthant would be infinite at every infeasible $\widehat{x}$, so the restriction is essential.

Theorem 1 (Gap transfer to the original problem)**.** *For every $\widehat{z} = (\widehat{x}, \widehat{y}) \in X \times Y_R$,*

$$f_0(\widehat{x}) - f^\star_0 \le \mathrm{Gap}_{R,\gamma}(\widehat{z}) + b_0(\gamma), \qquad \|[g(\widehat{x})]_+\|_\infty \le \mathrm{Gap}_{R,\gamma}(\widehat{z}) + b_g(\gamma). \tag{18}$$

Here $b_g(\gamma) := \max_{i \in [m]} b_i(\gamma)$. The same inequalities hold after taking expectations. In particular, if $\mathbb{E}\,\mathrm{Gap}_{R,\gamma}(\widehat{z}) \le \varepsilon/3$, $b_0(\gamma) \le \varepsilon/3$, and $b_g(\gamma) \le \varepsilon/3$, then $\mathbb{E}[f_0(\widehat{x}) - f^\star_0] \le 2\varepsilon/3 \le \varepsilon$ and $\mathbb{E}\|[g(\widehat{x})]_+\|_\infty \le 2\varepsilon/3 \le \varepsilon$.

3.3 Localized primal certificate

For later use define, for $x \in X$,

$$\mathcal{C}_{R,\gamma}(x) := \sup_{y \in Y_R} \left\{f_{0,\gamma}(x) - f^\star_{0,\gamma} + \langle y, h_\gamma(x)\rangle\right\}. \tag{19}$$

The unit margin in (14) gives

$$f_{0,\gamma}(x) - f^\star_{0,\gamma} \le \mathcal{C}_{R,\gamma}(x), \qquad \|[h_\gamma(x)]_+\|_\infty \le \mathcal{C}_{R,\gamma}(x). \tag{20}$$

Indeed, the first inequality uses $y = 0$; for the second, if j maximizes $[h_{j,\gamma}(x)]_+$ and e_j is the jth coordinate vector, then $y^\star_\gamma + e_j \in Y_R$ and saddle optimality yields the claim. The

name *certificate* is deliberate: by (20) a single scalar quantity simultaneously certifies both a small objective residual and a small maximum constraint violation, so the whole accelerated analysis below can be carried out in terms of this one quantity. For any $\widehat{z} = (\widehat{x}, \widehat{y}) \in X \times Y_R$,

$$\mathcal{C}_{R,\gamma}(\widehat{x}) \le \mathrm{Gap}_{R,\gamma}(\widehat{z}), \tag{21}$$

because $\inf_{x \in X} \Phi_\gamma(x, \widehat{y}) \le \Phi_\gamma(x^\star_\gamma, \widehat{y}) \le f^\star_{0,\gamma}$.

3.4 One random direction estimates the full Jacobian

Let u be uniform on the Euclidean unit sphere S_2^{d-1}. With the same sample ξ at the two perturbed points, define for every component

$$\widehat{\nabla} f_{i,\gamma}(x; u, \xi) := \frac{d}{2\gamma}\big(F_i(x + \gamma u, \xi) - F_i(x - \gamma u, \xi)\big)u. \tag{22}$$

The two vector calls in Definition 2 produce (22) for all $i = 0, \dots, m$ simultaneously.

Lemma 2 (Unbiasedness and a vectorized second moment)**.** *Under Assumption 1, for every $x \in X$,*

$$\mathbb{E}_{u,\xi}[\widehat{\nabla} f_{i,\gamma}(x; u, \xi)] = \nabla f_{i,\gamma}(x). \tag{23}$$

Moreover, for every $y \in Y_R$,

$$\mathbb{E}\left\| \widehat{\nabla} f_{0,\gamma}(x; u, \xi) + \sum_{i=1}^{m} y_i \widehat{\nabla} f_{i,\gamma}(x; u, \xi) \right\|_2^2 \le c_{\mathrm{zo}}\, d\,(M_0 + \textstyle\sum_i y_i M_i)^2 \le c_{\mathrm{zo}}\, d M_R^2, \tag{24}$$

where one may take $c_{\mathrm{zo}} \le 2$ for $d \ge 2$ (and $c_{\mathrm{zo}} = 1$ in dimension $d = 1$). Thus estimating all constraint gradients does not multiply the number of vector query rounds by m.

To estimate the dual component $-h_\gamma(x)$ of the saddle operator, draw v uniformly from B_2^d and use

$$\widehat{h}_\gamma(x; v, \xi) = F_{1:m}(x + \gamma v, \xi) - b_{1:m}(\gamma), \tag{25}$$

where the slice notation collects the constraint components: $F_{1:m} := (F_1, \dots, F_m)$ and $b_{1:m}(\gamma) := (b_1(\gamma), \dots, b_m(\gamma))$.

Assumption 5 (Dual-value noise on a query domain)**.** Fix a deterministic center set $\mathcal{S}$ and an oracle domain $\mathcal{D}$ such that $\mathcal{S} + \bar{\gamma} B_2^d \subseteq \mathcal{D}$. Let $\mathcal{F}$ denote the algorithmic history before a fresh dual-value sample is drawn. Uniformly over $\mathcal{F}$-measurable $x \in \mathcal{S}$ and admissible smoothing radii,

$$\mathbb{E}\left[\left\| \widehat{h}_\gamma(x; v, \xi) - h_\gamma(x) \right\|_\infty^2 \mid \mathcal{F} \right] \le \sigma_h^2. \tag{26}$$

The fresh sample is conditionally unbiased for $h_\gamma(x)$. For Mirror–Prox we take $\mathcal{S} = X$ and $\mathcal{D} = X_{\bar{\gamma}}$; for PB–ACGD-S the two sets are specified in Assumption 6. This moment assumption is independent of pathwise Lipschitz continuity: an additive random offset cancels in a common-randomness central difference but does not cancel in the absolute constraint-value block.

This bound may include both data noise and randomization by v. It is an ℓ_∞ quantity and can therefore contain a logarithmic dependence on m for sub-Gaussian coordinate noise, but it need not contain a multiplicative factor m.

The next conditional ℓ_∞-averaging estimate is the step needed to pass from the one-sample moment in Assumption 5 to a mini-batch. In contrast with a Hilbert norm, a dimension-free division of an arbitrary finite ℓ_∞ second moment by the batch size is not valid.

Lemma 3 (Conditional ℓ_∞ batching)**.** *Let $\mathcal{H}$ be a sigma-field and let $\xi_1, \dots, \xi_r \in \mathbb{R}^m$ be conditionally independent given $\mathcal{H}$, conditionally centered, and satisfy $v_j := \mathbb{E}[\|\xi_j\|_\infty^2 \mid \mathcal{H}] < \infty$. Set $\bar{\xi}_r := r^{-1} \sum_{j=1}^r \xi_j$. There is a universal constant C_{bt} such that, almost surely,*

$$\mathbb{E}[\|\bar{\xi}_r\|_\infty^2 \mid \mathcal{H}] \le \frac{C_{\mathrm{bt}} \kappa_m}{r^2} \sum_{j=1}^r v_j \le \frac{C_{\mathrm{bt}} \kappa_m}{r} \max_{1 \le j \le r} v_j. \tag{27}$$

The proof in Appendix A uses conditional symmetrization followed by the Rademacher maximal inequality over the m coordinates. Thus Lemma 3 uses only the original finite-second-moment hypothesis; no sub-Gaussian tail assumption is being inserted into the general Mirror–Prox theorem.

Lemma 4 (A checkable sub-Gaussian sufficient condition)**.** *Suppose that, conditionally on the algorithmic history and on the smoothing draw v, the centered coordinate noises*

$$Z_i = F_i(x + \gamma v, \xi) - f_i(x + \gamma v), \qquad i \in [m],$$

are sub-Gaussian with a common variance proxy σ^2 (independence across coordinates is not required). Under Assumption 1, uniformly for $\gamma \le \bar{\gamma}$,

$$\mathbb{E}\left[\|\widehat{h}_\gamma(x; v, \xi) - h_\gamma(x)\|_\infty^2 \mid \mathcal{F}\right] \le C\sigma^2 \log(2m) + 8\bar{\gamma}^2 M_g^2. \tag{28}$$

Consequently Assumption 5 holds with the right-hand side of (28) as a valid choice of σ_h^2. Moreover, for an independent batch of r such samples at a conditionally fixed query point,

$$\mathbb{E}\left[\left\|\frac{1}{r}\sum_{j=1}^r (\widehat{h}_{\gamma,j} - h_\gamma)\right\|_\infty^2 \middle| \mathcal{F}\right] \le \frac{C \log(2m)}{r}\left(\sigma^2 + \bar{\gamma}^2 M_g^2\right). \tag{29}$$

Indeed, the standard maximal inequality for sub-Gaussian coordinates gives $\mathbb{E} \max_i |Z_i|^2 \le C\sigma^2 \log(2m)$. The remaining randomness comes from v; Lipschitz continuity gives $|f_i(x + \gamma v) - \mathbb{E}_v f_i(x + \gamma v)| \le 2\gamma M_i$, and $(a+b)^2 \le 2a^2 + 2b^2$ yields (28). For (29), the same bounded centered smoothing term is sub-Gaussian, averaging divides its coordinatewise variance proxy by r, and the maximal inequality is then applied once to the averaged coordinates. This sharper specialization avoids paying one logarithm in σ_h^2 and a second through the generic ℓ_∞-averaging bound of Lemma 3.

Assumption 6 (Accelerated query-domain availability)**.** For the PB–ACGD-S block, define the center set and its oracle neighborhood by

$$\mathcal{S}_{\mathrm{ac}} := X + 2D_X B_2^d, \qquad X_{\mathrm{ac},\bar{\gamma}} := \mathcal{S}_{\mathrm{ac}} + \bar{\gamma} B_2^d. \tag{30}$$

The oracle is defined on $X_{\mathrm{ac},\bar{\gamma}}$, and Assumption 1 holds there with the same constants. Whenever Assumption 3 is invoked, its gradient-Lipschitz bounds also hold there with the same L_i. The accelerated corollaries use this deterministic pathwise branch. Extending them to the paired-second-moment branch would require replacing both primal and Jacobian-action terms in (59) by suitable localized random-modulus moments; that extension is not claimed here. Assumption 5 holds with $\mathcal{S} = \mathcal{S}_{\mathrm{ac}}$ and $\mathcal{D} = X_{\mathrm{ac},\bar{\gamma}}$. This requirement is stronger than the Mirror–Prox neighborhood because accelerated centers may lie outside X before the inner points are projected back. It is an oracle-domain assumption, not a consequence of convexity on X.

4 An unconditional batched stochastic Mirror–Prox method

4.1 Geometry and operator constants

Algorithm 1 may use any product distance-generating function that is one-strongly convex in a chosen product norm and has a finite radius from its initial center. For the explicit vector-round bounds, equip X with the Euclidean norm and a one-strongly convex function ω_X. Let V_X be its Bregman divergence, choose $x_1 \in X$, and define

$$D^2_{\omega,X} := \sup_{x\in X} V_X(x_1, x) < \infty.$$

Identify Y_R with the augmented simplex $\widetilde{Y}_R = \{\widetilde{y} = (y_0, y) \in \mathbb{R}^{m+1}_+ : \sum_{j=0}^m y_j = R\}$ by setting $y_0 = R - \|y\|_1$. Let $\widetilde{y}_1 = (R/(m+1), \dots, R/(m+1))$ and use the scaled entropy $\omega_Y(\widetilde{y}) = R\sum_{j=0}^m y_j \log(y_j/R)$. Writing $p = \widetilde{y}/R$ and $q = \widetilde{y}'/R$, the Bregman divergence is $V_Y(\widetilde{y}, \widetilde{y}') = R^2\mathrm{KL}(q\|p) \geq \frac{1}{2}\|\widetilde{y}' - \widetilde{y}\|_1^2$ by Pinsker, so this prox is one-strongly convex in the convention $V \geq \|\cdot\|^2/2$. Its radius from $\widetilde{y}_1$ satisfies

$$D^2_{\omega,Y} := \sup_{\widetilde{y}\in\widetilde{Y}_R} V_Y(\widetilde{y}_1, \widetilde{y}) \leq R^2 \log(m+1). \tag{31}$$

This center-based radius is the quantity needed in the Mirror–Prox proof; the pairwise entropy Bregman diameter over boundary points is infinite. Set $z_1 := (x_1, \widetilde{y}_1)$, $V_Z(z, z') := V_X(x, x') + V_Y(\widetilde{y}, \widetilde{y}')$, and $D^2_\omega = D^2_{\omega,X} + D^2_{\omega,Y}$. Suppressing the slack coordinate in the notation, use the product norm

$$\|(x, \widetilde{y})\|_Z^2 = \|x\|_2^2 + \|\widetilde{y}\|_1^2, \qquad \big\|(s, \widetilde{t})\big\|_{Z,*}^2 = \|s\|_2^2 + \big\|\widetilde{t}\big\|_\infty^2. \tag{32}$$

The dual part of the saddle operator is embedded as $(0, -h_\gamma(x))$, so its augmented dual norm is still $\|h_\gamma(x)\|_\infty$. The entropy prox step is also explicit. If the current augmented vector is $\widetilde{y} = (y_0, y_1, \dots, y_m)$ and the augmented dual coefficient is $a = (a_0, \dots, a_m)$, then

$$\widetilde{y}_j^+ = R\frac{\widetilde{y}_j \exp(-a_j/R)}{\sum_{\ell=0}^m \widetilde{y}_\ell \exp(-a_\ell/R)}, \qquad j = 0, \dots, m. \tag{33}$$

For a saddle-operator step, $a_0 = 0$ and $a_i = -\eta\widehat{h}_{i,\gamma}$ for $i \geq 1$. Algorithm 1 performs (33) in the augmented simplex and then returns the physical multiplier by dropping the slack coordinate y_0.

The monotone saddle operator associated with (16) is

$$G_\gamma(x, y) = \begin{pmatrix} \nabla f_{0,\gamma}(x) + Jh_\gamma(x)^\top y \\ -h_\gamma(x) \end{pmatrix}. \tag{34}$$

Here and below $Jh_\gamma(x) \in \mathbb{R}^{m\times d}$ denotes the Jacobian matrix of h_γ at x, whose ith row is $\nabla h_{i,\gamma}(x)^\top$. Let $L_{g,\gamma} = \max_i L_{i,\gamma}$. A direct product-norm estimate gives

$$\big\|G_\gamma(z) - G_\gamma(z')\big\|_{Z,*} \leq L_\gamma \big\|z - z'\big\|_Z, \qquad L_\gamma \leq c_L\big(L_{0,\gamma} + RL_{g,\gamma} + M_g\big), \tag{35}$$

with a universal c_L. In the nonsmooth regime, this becomes

$$L_\gamma \leq c\left(\frac{\sqrt{d}\,(M_0 + RM_g)}{\gamma} + M_g\right). \tag{36}$$

Let $\widehat{G}_\gamma(z)$ be the unbiased estimator formed from (22) and (25), and let $\widehat{G}_{\gamma,r}(z)$ average r conditionally independent copies at a conditionally fixed query point. The Euclidean primal

error averages with the Hilbert-space $1/r$ factor, whereas Lemma 3 must be used for the dual error. To keep both moment branches visible, write $\mathfrak{M}_R^2 = M_R^2$ under Assumption 1 and $\mathfrak{M}_R^2 = \overline{M}_R^2$ under its paired-second-moment alternative in the Mirror–Prox formulas and sharpness comparison. Combining (24), (26), and (27), we may take a batch-variance proxy satisfying

$$\mathbb{E}\left[\left\|\widehat{G}_{\gamma,r}(z) - G_\gamma(z)\right\|_{Z,*}^2 \mid \mathcal{F}\right] \le \frac{\Sigma_\gamma^2}{r}, \qquad \Sigma_\gamma^2 \le C_\Sigma\big(d\mathfrak{M}_R^2 + \kappa_m \sigma_h^2\big) \tag{37}$$

under common-randomness two-point feedback. Thus the one-step bound (26) alone is not being divided by r without a geometry factor. Under Lemma 4, the term $\kappa_m\sigma_h^2$ in this generic finite-moment proxy may instead be replaced by the sharper $C\log(2m)(\sigma^2 + \bar{\gamma}^2 M_g^2)$ from (29). With independent additive value noise, both the central-difference and absolute-value contributions described in Corollary 9 must be added.

4.2 Algorithm

For a vector a and a base point z, let

$$\mathrm{Prox}_z(a) = \operatorname*{arg\,min}_{w \in X \times Y_R} \big\{\langle a, w\rangle + V_Z(z, w)\big\}.$$

At each oracle point, $\widehat{G}_{\gamma,r}$ denotes the average of r independent copies of $\widehat{G}_\gamma$, as above.

4.3 Oracle accounting within Mirror–Prox

All physical cost units used throughout the paper are summarized once in Table 1; the oracle definitions above specify the information returned, while this table fixes the accounting convention. Thus N Mirror–Prox iterations have strict adaptive depth $2N$ up to initialization,

Table 1. Oracle-accounting units used in theory and figure captions (see the query-accounting convention at the end of Section 2). The second column is the contribution of the object to the work Q_{vec}; the third column is its contribution to the sequential depth: objects within one stage are issued in parallel.

Object	Vector-value calls	Adaptive stage
One full value $F(x,\xi)$	1	—
One common-randomness two-point gradient sample	2	one sampled pair
One dual-value sample $\widehat{h}_\gamma$	1	same operator stage
One saddle-operator sample	3	one operator stage
One Mirror–Prox iteration with batch r	$6r$	two sequential extragradient stages

while the paper reports N as iteration depth when constant factors are suppressed. Direct componentwise simulation multiplies the scalar output count by $m+1$; it does not change this adaptive-stage count.

Each stochastic operator sample uses a constant number of full vector-value calls: two for the central finite difference and one for the smoothed constraint value. Mirror–Prox evaluates the operator twice per iteration. These fixed constants are suppressed below.

Theorem 2 (Expected restricted-gap bound)**.** *Suppose* (35) *holds and* $D_\omega^2 > 0$ *(the zero-radius domain is trivial). At each extragradient stage, conditionally on the history before*

Algorithm 1 Batched vector-oracle stochastic Mirror–Prox

Input: Prox center $z_1 = (x_1, \widetilde{y}_1)$ defined above, smoothing radius γ, batch size r, step size η, and iteration count N.

1: **for** $k = 1, \dots, N$ **do**
2: Draw a fresh batch and form $\widehat{G}_{\gamma,r}(z_k)$.
3: $w_k \leftarrow \mathrm{Prox}_{z_k}\left(\eta \widehat{G}_{\gamma,r}(z_k)\right)$.
4: Draw an independent batch and form $\widehat{G}_{\gamma,r}(w_k)$.
5: $z_{k+1} \leftarrow \mathrm{Prox}_{z_k}\left(\eta \widehat{G}_{\gamma,r}(w_k)\right)$.
6: **end for**

Output: $\bar{z}_N \leftarrow N^{-1} \sum_{k=1}^{N} w_k$.

that stage, let the fresh batch consist of r conditionally independent unbiased samples, and assume that their average satisfies the conditional batch bound in (37) for some Σ_γ; the two stage batches are conditionally independent. For the concrete common-randomness estimator, Assumption 1 (or its paired-second-moment alternative), Assumption 5 with $\mathcal{S} = X$ and $\mathcal{D} = X_{\bar{\gamma}}$, and the preceding estimator lemmas supply this bound. There are universal constants $C_0, C_1, C_2, c_\eta > 0$ such that, when $L_\gamma + \Sigma_\gamma > 0$, Algorithm 1, with

$$\eta = \min\left\{ \frac{1}{2L_\gamma}, c_\eta \sqrt{\frac{D_\omega^2 r}{\Sigma_\gamma^2 N}} \right\}, \tag{38}$$

where an entry with zero denominator is omitted, satisfies

$$\mathbb{E}\,\mathrm{Gap}_{R,\gamma}(\bar{z}_N) \le C_1 \frac{L_\gamma D_\omega^2}{N} + C_2 \sqrt{\frac{\Sigma_\gamma^2 D_\omega^2}{Nr}}. \tag{39}$$

If $L_\gamma = \Sigma_\gamma = 0$, the unsimplified estimate $\mathbb{E}\,\mathrm{Gap}_{R,\gamma}(\bar{z}_N) \le C_0 D_\omega^2/(\eta N)$ holds for every $\eta > 0$; in particular, for any target value $\varepsilon_s > 0$, choosing $\eta \ge C_0 D_\omega^2/(N\varepsilon_s)$ gives gap at most ε_s. This is the convention for the otherwise empty minimum in (38).

4.4 Parameter choice and oracle cost

For a target smoothed gap ε_s, it is sufficient to choose

$$N \ge \max\left\{1, C \frac{L_\gamma D_\omega^2}{\varepsilon_s}\right\}, \qquad Nr \ge C \frac{\Sigma_\gamma^2 D_\omega^2}{\varepsilon_s^2} \tag{40}$$

for an expected smoothed gap at most ε_s, using the preceding step-size convention in the simultaneous zero case. The sequential depth and number of vector calls obey

$$N_{\mathrm{seq}} = \mathcal{O}\left(1 + \frac{L_\gamma D_\omega^2}{\varepsilon_s}\right), \qquad Q_{\mathrm{vec}} = \mathcal{O}\left(1 + \frac{L_\gamma D_\omega^2}{\varepsilon_s} + \frac{\Sigma_\gamma^2 D_\omega^2}{\varepsilon_s^2}\right). \tag{41}$$

By (21), the same output also satisfies $\mathbb{E}\mathcal{C}_{R,\gamma}(\bar{x}_N) \le \mathbb{E}\,\mathrm{Gap}_{R,\gamma}(\bar{z}_N)$. Thus Corollaries 1–2 pass through Theorem 1 without any forward reference.

Theorem 2 is a direct application of the stochastic Mirror–Prox analysis [Juditsky et al., 2011], rather than an informal substitution of a finite-difference vector into a deterministic proof. It allows arbitrary prox geometry and counts the batch size explicitly.

Corollary 1 (Smooth fully stochastic vector-oracle complexity)**.** *Assume Assumptions 1, 3, 4, and 5. Let* $L_{\max} = \max_{0\le i\le m} L_i$ *and choose*

$$\gamma = \min\left\{\bar{\gamma}, \sqrt{\frac{2\varepsilon}{3L_{\max}}}\right\}, \qquad \varepsilon_s = \frac{\varepsilon}{3}, \tag{42}$$

with the second entry omitted when $L_{\max} = 0$*. Then Algorithm 1 returns an expected* ε*-solution of (1) with*

$$N_{\mathrm{seq}} = \mathcal{O}\left(1 + \frac{(L_0 + RL_g + M_g)D_\omega^2}{\varepsilon}\right), \tag{43}$$

$$Q_{\mathrm{vec}} = \mathcal{O}\left(1 + \frac{(L_0 + RL_g + M_g)D_\omega^2}{\varepsilon} + \frac{(dM_R^2 + \kappa_m\sigma_h^2)D_\omega^2}{\varepsilon^2}\right). \tag{44}$$

Thus the primal two-point contribution is $\mathcal{O}(dM_R^2D_\omega^2/\varepsilon^2)$*, while the general finite-moment dual contribution is* $\mathcal{O}(\kappa_m\sigma_h^2D_\omega^2/\varepsilon^2)$*. Under Lemma 4, the latter can be sharpened to* $\mathcal{O}((\sigma^2 + \bar{\gamma}^2M_g^2)\log(2m)D_\omega^2/\varepsilon^2)$*. In the componentwise model, the same implementation has* $Q_{\mathrm{cmp}} = (m+1)Q_{\mathrm{vec}}$*.*

Corollary 2 (Lipschitz nonsmooth complexity)**.** *Under Assumptions 1, 4, and 5, let* $M_{\max} = \max\{M_0, M_g\}$ *and choose*

$$\gamma = \min\left\{\bar{\gamma}, \frac{\varepsilon}{3M_{\max}}\right\}, \qquad \varepsilon_s = \frac{\varepsilon}{3}. \tag{45}$$

The second entry is omitted when $M_{\max} = 0$*; with this convention,* $\gamma^{-1} = \max\{\bar{\gamma}^{-1}, 3M_{\max}/\varepsilon\}$ *in all cases. Then Algorithm 1 returns an expected* ε*-solution and*

$$N_{\mathrm{seq}} = \mathcal{O}\left(1 + \frac{\sqrt{d}\,M_RD_\omega^2}{\varepsilon}\max\left\{\frac{1}{\bar{\gamma}}, \frac{3M_{\max}}{\varepsilon}\right\} + \frac{M_gD_\omega^2}{\varepsilon}\right), \tag{46}$$

$$Q_{\mathrm{vec}} = \mathcal{O}\left(1 + \frac{\sqrt{d}\,M_RD_\omega^2}{\varepsilon}\max\left\{\frac{1}{\bar{\gamma}}, \frac{3M_{\max}}{\varepsilon}\right\} + \frac{(dM_R^2 + \kappa_m\sigma_h^2)D_\omega^2}{\varepsilon^2} + \frac{M_gD_\omega^2}{\varepsilon}\right). \tag{47}$$

These are the general bounds, including the regime where the oracle-neighborhood radius is the active restriction. If $M_{\max} > 0$ *and* $\varepsilon \le 3M_{\max}\bar{\gamma}$*, then* $\gamma = \varepsilon/(3M_{\max})$ *and the first terms reduce, respectively, to* $\mathcal{O}(\sqrt{d}\,M_RM_{\max}D_\omega^2/\varepsilon^2)$*; hence the primal stochastic order is* $\mathcal{O}(dM_R^2D_\omega^2/\varepsilon^2)$*. The general finite-moment dual contribution is* $\mathcal{O}(\kappa_m\sigma_h^2D_\omega^2/\varepsilon^2)$*, or the sharper* $\mathcal{O}((\sigma^2 + \bar{\gamma}^2M_g^2)\log(2m)D_\omega^2/\varepsilon^2)$ *under Lemma 4.*

Independent additive measurement noise produces the familiar γ^{-2} central-difference penalty and, after nonsmooth smoothing, an ε^{-4} upper-bound term. Corollary 9 in the Appendix records the precise statement; no matching minimax lower bound is claimed for that noise model.

In the small-accuracy branch of Corollary 2, the primal two-point term has the same d/ε^2 leading order as unconstrained stochastic zeroth-order convex optimization, while the finite-moment dual-value term is displayed separately with its κ_m factor. The price of nonsmoothness is visible in the sequential depth: there $L_\gamma = \mathcal{O}(\sqrt{d}/\varepsilon)$ up to problem scales, and a nonaccelerated variational-inequality method therefore needs $\mathcal{O}(\sqrt{d}/\varepsilon^2)$ rounds. The acceleration analysis below removes the remaining sequential-depth gap by inserting fresh vector-value noise directly into the ACGD-S sliding filtration; no uniform stochastic inexact-model assumption is used.

5 Acceleration via phase-banked stochastic ACGD-S

The acronym ACGD-S stands for *accelerated constrained gradient descent with sliding* [Zhang, Lan, 2025]; the prefix PB (*phase-banked*) refers to the phase-wise banking of samples proposed here.

Sliding. The device of Zhang and Lan separates the two parts of the saddle operator by cost: the "expensive" smooth part (the objective gradient) is evaluated once per outer phase, at a frozen query center $\underline{x}_t$, whereas the "cheap" bilinear part (the constraint Jacobian multiplied by the current multiplier) is re-evaluated many times inside the inner loop of the same phase. The inner loop slides along the dual variable while the center stays frozen — hence the name. The parameter $r_{\mathrm{sl}} > 0$ (the sliding scale) sets the relative weight of the primal and dual metrics in the product prox function (50) and thereby controls the number of inner steps per phase. Throughout this section we call one inner iteration, indexed by the pair (t, s), a *slot*; the underline in $\underline{x}_t$ marks the query center of phase t, as opposed to the phase output x_t.

Definition 5 (Sample bank)**.** Let $\underline{x}_t$ be the query center of phase t, defined in (54), and let S_t be the number of its inner slots, defined in (52). The *bank* of phase t is the family of independent mini-batches

$$\{A_{t,s},\, B_{t,s} :\ s = 1, \dots, S_t\} \quad (\text{plus, for } t \geq 2, \text{ one additional batch } A^{-}_{t,1}),$$

each consisting of b independent oracle realizations, with two properties. (i) All queries of the bank are issued at the single, already known point $\underline{x}_t$ (for $A^{-}_{t,1}$, at $\underline{x}_{t-1}$), immediately after that point has been computed and before the inner loop starts; they are therefore physically parallel and occupy $O(1)$ rounds. (ii) The batch assigned to slot s is used ("revealed") for the first time only at that slot, so with respect to the inner-loop filtration it remains fresh and conditionally unbiased.

The batches have fixed roles. Batch $A_{t,s}$ serves the primal step (estimates of the objective gradient and of the Jacobian), and the independent batch $B_{t,s}$ serves the dual step (a second Jacobian estimate and a value of $h_\gamma(\underline{x}_t)$). The batch $A^{-}_{t,1}$ at the previous center is called the *reserve*; it is needed by a single term — the cross-phase extrapolation in the first slot, (56) — and adds only a constant factor to the cost of the phase.

5.1 Why a uniform stochastic inexact model is the wrong object

Before introducing the banks into the recurrence, we explain why one cannot simply substitute mini-batch estimates into the deterministic ACGD-S proof. For the accelerated block, let $L_{A,\gamma} > 0$ be a deterministic localized smoothness majorant satisfying

$$L_{A,\gamma} \geq L_{0,\gamma} + R L_{g,\gamma}. \tag{48}$$

Any positive majorant is valid for the fixed-horizon theorem. For the target-accuracy tuning below, it is raised, if necessary, by the explicit target-dependent floor (62), applied after $\mathcal{D}_A$ has been defined. This avoids both zero and arbitrarily small denominators in the sliding schedule. For smooth deterministic function constraints, Zhang and Lan [Zhang, Lan, 2025] obtain the accelerated outer weights

$$\omega_t = t, \qquad \tau_t = \frac{t-1}{2}, \qquad \eta_t = \frac{L_{A,\gamma}}{\tau_{t+1}} = \frac{2L_{A,\gamma}}{t}, \tag{49}$$

inside ACGD and then solve each linearized constrained descent problem by the primal–dual sliding routine ACGD-S. The deterministic ACGD-S analysis gives an $O(L_{A,\gamma}/N^2)$ weighted primal–dual residual; with the schedule below the total number of inner bilinear slots is $K_N =$

$O(N + (\bar{M}_\gamma r_{\mathrm{sl}}/L_{A,\gamma})N^2)$; Theorem 5 and Proposition 3 of Zhang and Lan [Zhang, Lan, 2025] give the precise Euclidean instance. We use the same telescoping argument in the product prox geometry below.

Assumption 6 supplies the larger center and oracle sets needed below. Equation (54) shows that every accelerated center lies in $\mathcal{S}_{\mathrm{ac}}$, so the stated neighborhood is sufficient for each two-point and shifted-value query.

A direct replacement of the exact affine models by a single random batch per outer phase is not sufficient. Consider the following one-dimensional example. Let the modeled function be $q \equiv 0$ on the segment $[-D, D]$, and let a batched estimate of its gradient be the random scalar e_r with $\mathbb{E}e_r = 0$ and $\mathbb{E}e_r^2 \asymp \Sigma^2/r$. The resulting random affine model $x \mapsto e_r x$ then deviates from q by $\sup_{|x|\le D} |e_r x| = D|e_r|$, whose root-mean-square size is of order $D\Sigma/\sqrt{r}$ — of order $1/\sqrt{r}$, not $1/r$. Thus a uniform $O(1/r)$ inexact-model law cannot follow from unbiased batching alone. The construction below avoids this obstruction: randomness is inserted *inside* the ACGD-S sliding proof, where it is handled as a martingale perturbation of the prox inequalities rather than as a pointwise model error.

5.2 Phase-wise sample banks

The plan of this subsection is as follows. We first fix the prox geometry and the deterministic schedule of inner steps — this is exactly the ACGD-S construction of [Zhang, Lan, 2025] rewritten in our notation. We then describe the single departure from the deterministic method: every exact quantity (objective gradient, Jacobian, constraint values) is replaced by the average over a fresh mini-batch from the bank of the phase (Definition 5), with the primal and dual steps receiving *independent* batches.

For the accelerated block we fix the primal prox to the Euclidean divergence $V_X(x, z) = \frac{1}{2}\|x - z\|_2^2$, matching the outer ACGD-S identity. The dual prox V_Y may remain any one-strongly-convex divergence in the ℓ_1 geometry (in particular the entropy prox used above); the inner proof uses only its three-point inequality and strong convexity. Choose PB prox centers $x_0 \in X$ and $y_0 \in Y_R$ with finite center-based radii

$$D_{\omega,X,\mathrm{ac}}^2 := \sup_{x\in X} V_X(x_0, x), \qquad D_{\omega,Y,\mathrm{ac}}^2 := \sup_{y\in Y_R} V_Y(y_0, y).$$

For the entropy prox, y_0 is the non-slack part of the uniform augmented center $\widetilde{y}_1$ defined above; a Euclidean dual prox may instead use $y_0 = 0$. Fix a sliding scale $r_{\mathrm{sl}} > 0$ and use the scaled product divergence

$$V_A((x, y), (x', y')) = V_X(x, x') + r_{\mathrm{sl}}^{-2} V_Y(y, y'), \qquad \mathcal{D}_A^2 := 3D_{\omega,X,\mathrm{ac}}^2 + r_{\mathrm{sl}}^{-2} D_{\omega,Y,\mathrm{ac}}^2. \tag{50}$$

The fixed numerical factor 3 multiplying the primal radius is a convenience constant inherited from the telescoping proof in Appendix A; it carries no structural meaning. We henceforth take $\mathcal{D}_A > 0$; if this radius is zero, the localized product domain is a singleton and the optimization problem is trivial. Let

$$\bar{M}_\gamma := \sup_{x\in\mathcal{S}_{\mathrm{ac}}} \|Jh_\gamma(x)\|_{2\to\infty} \le M_g, \qquad \Delta := \frac{r_{\mathrm{sl}}}{L_{A,\gamma}}, \tag{51}$$

where $\|A\|_{2\to\infty} := \sup_{\|w\|_2\le 1} \|Aw\|_\infty$ is the operator norm from ℓ_2 to ℓ_∞ (equivalently, the largest Euclidean row norm); it is the norm matched with the primal ℓ_2 and dual ℓ_1 geometries. We choose the deterministic conservative sliding schedule

$$S_t = \max\{1, \lceil \bar{M}_\gamma \Delta t \rceil\}, \qquad \widetilde{M}_t = \frac{S_t}{\Delta t}, \qquad a_{t,s} = \frac{\omega_t}{S_t} = \frac{t}{S_t}, \tag{52}$$

with the usual ACGD-S inner coefficients $\beta^{\mathrm{det}}_{t,s} = \widetilde{M}_t r_{\mathrm{sl}}$, $\gamma^{\mathrm{det}}_{t,s} = \widetilde{M}_t / r_{\mathrm{sl}}$, and $\delta_{t,s} = 1$ (we keep the standard ACGD-S symbols: the slot-indexed dual step coefficients $\gamma_{t,s}$ are unrelated to the smoothing radius γ, which never carries slot indices). For $t \geq 2$ the first-slot extrapolation factor is the standard ratio $\rho_{t,1} = \widetilde{M}_t / \widetilde{M}_{t-1}$; set $\rho_{1,1} = 0$, and set $\rho_{t,s} = 1$ for $s \geq 2$. Since $\widetilde{M}_t \geq \bar{M}_\gamma$, these coefficients satisfy the deterministic intra- and inter-phase coupling inequalities used in the ACGD-S proof. The meaning of the schedule (52) is the following: S_t is the number of inner slots of phase t; it grows linearly in t with slope $\bar{M}_\gamma r_{\mathrm{sl}} / L_{A,\gamma}$, so the more expensive the bilinear coupling is relative to the smoothness of the objective, the more cheap inner steps are taken per one gradient evaluation. The quantity Δ acts as the discretization step of the schedule, $S_t \approx \bar{M}_\gamma \Delta t$, and $\widetilde{M}_t = S_t / (\Delta t)$ is $\bar{M}_\gamma$ corrected for rounding S_t up to an integer.

After the accelerated center $\underline{x}_t$ is known, the entire bank of phase t (Definition 5) is queried in a constant number of parallel oracle stages: for every inner slot $s \leq S_t$, two independent mini-batches $A_{t,s}$ and $B_{t,s}$ of size b, and, for $t \geq 2$, the reserve $A^-_{t,1}$ at $\underline{x}_{t-1}$. Batch $A_{t,s}$ supplies common-randomness two-point estimates of the objective gradient and of all rows of the constraint Jacobian at $\underline{x}_t$ along one shared random direction (Lemma 2) for the primal/extrapolation step. Independently, $B_{t,s}$ supplies a second Jacobian estimate and a fresh estimate of $h_\gamma(\underline{x}_t)$ for the dual step. Because every stored sample is independent and is queried at an already known point, the bank may be obtained physically in parallel and then *revealed* sequentially in the proof without changing its joint law (the deferred-revelation construction in Step 2 of the proof of Theorem 3 in Appendix A).

Here $\widehat{J}_\gamma^{A_{t,s}}$ and $\widehat{J}_\gamma^{B_{t,s}}$ are the matrices whose rows are the mini-batch averages of the constraint estimators in (22); the same superscripts on $\widehat{\nabla} f_{0,\gamma}$ and $\widehat{h}_\gamma$ denote the corresponding mini-batch averages. A stored Jacobian sample is combined with an inner multiplier only after that multiplier is known, so conditional unbiasedness is preserved. The independent dual coefficient is

$$\widehat{q}^{B}_{t,s}(x) = \widehat{h}_\gamma^{B_{t,s}}(\underline{x}_t) + \widehat{J}_\gamma^{B_{t,s}}(\underline{x}_t)(x - \underline{x}_t). \tag{53}$$

The B-bank is revealed only after the primal point of that slot has been formed, so x in (53) is predictable for the dual noise.

For completeness, the stochastic inner recurrence is written explicitly. Let $p_{t,s}$ be the inner primal point and $y_{t,s} \in Y_R$ the inner multiplier. Initialize

$$x_{-1} = x_0, \qquad \underline{x}_0 = x_0, \qquad p_{1,0} = x_0, \qquad y_{1,-1} = y_{1,0} = y_0.$$

After phase t, carry the terminal states exactly as in deterministic ACGD-S:

$$p_{t+1,0} = p_{t,S_t}, \qquad y_{t+1,0} = y_{t,S_t}, \qquad y_{t+1,-1} = y_{t,S_t-1}.$$

(The last expression is well defined also for $S_t = 1$, in which case it equals $y_{t,0}$.) Set $\theta_1 = 0$ and $\theta_t = \omega_t / \omega_{t-1}$ for $t \geq 2$, and form

$$\widetilde{x}_t = x_{t-1} + \theta_t (x_{t-1} - x_{t-2}), \qquad \underline{x}_t = \frac{\tau_t \underline{x}_{t-1} + \widetilde{x}_t}{1 + \tau_t}. \tag{54}$$

For $s \geq 2$ define the predictable extrapolated multiplier $\widetilde{y}_{t,s-1} = y_{t,s-1} + \rho_{t,s}(y_{t,s-1} - y_{t,s-2})$ and set

$$\widehat{r}^{A}_{t,s} = \widehat{\nabla} f_{0,\gamma}^{A_{t,s}}(\underline{x}_t) + [\widehat{J}_\gamma^{A_{t,s}}(\underline{x}_t)]^\top \widetilde{y}_{t,s-1}. \tag{55}$$

At $s = 1$ the standard ACGD-S cross-phase extrapolation is

$$\widehat{r}^{A}_{t,1} = \widehat{\nabla} f_{0,\gamma}^{A_{t,1}}(\underline{x}_t) + [\widehat{J}_\gamma^{A_{t,1}}(\underline{x}_t)]^\top y_{t,0} + \rho_{t,1} [\widehat{J}_\gamma^{A^-_{t,1}}(\underline{x}_{t-1})]^\top (y_{t,0} - y_{t,-1}), \tag{56}$$

where $A^-_{t,1}$ is the independent reserve at the previous center. The final term in (56) is omitted when $t = 1$; equivalently, it is multiplied by $\rho_{1,1} = 0$ without querying an undefined reserve. The primal and dual prox steps are

$$p_{t,s} = \arg\min_{x \in X} \left\{ \langle \widehat{r}^A_{t,s}, x \rangle + \eta_t V_X(x_{t-1}, x) + \beta_{t,s} V_X(p_{t,s-1}, x) \right\}, \tag{57}$$

$$y_{t,s} = \arg\max_{y \in Y_R} \left\{ \langle y, \widehat{q}^B_{t,s}(p_{t,s}) \rangle - \gamma_{t,s} V_Y(y_{t,s-1}, y) \right\}. \tag{58}$$

The phase output is $x_t = S_t^{-1} \sum_{s=1}^{S_t} p_{t,s}$ and $y_t = S_t^{-1} \sum_{s=1}^{S_t} y_{t,s}$. Replacing every bank by exact first-order data recovers the deterministic ACGD-S inner loop. The complete one-slot product-moment proxy is

$$\Sigma^2_{A,\gamma} := c_x \, d M_R^2 + \kappa_m r_{\mathrm{sl}}^2 \Big(c_J \, d D_X^2 M_g^2 + c_h \sigma_h^2 \Big), \tag{59}$$

with universal constants c_x, c_J, c_h. Euclidean primal averaging contributes the factor $1/b$ without κ_m; only the dual ℓ_∞ block carries κ_m/b. Appendix A gives the domain, moment, and telescoping details.

The quantities $\mathcal{D}_A$ and $\Sigma_{A,\gamma}$ used in the displayed choice of λ_N are algorithmic inputs: any deterministic upper bounds valid for the prescribed oracle neighborhood may be substituted, and the theorem then holds with those upper bounds.

Algorithm 2 Phase-banked stochastic ACGD-S (PB-ACGD-S)

Input: PB prox centers $x_0 \in X$ and $y_0 \in Y_R$, phase count N, smoothing radius γ, batch size b, sliding scale r_{sl}, set Y_R, tuned majorant $L_{A,\gamma}$, and bounds $\bar{M}_\gamma$, $\mathcal{D}_A$, and $\Sigma_{A,\gamma}$ from (50), (51), and (59).

Set the coefficients by (49) and (52).
Initialize $x_{-1} = x_0$, $\underline{x}_0 = x_0$, $p_{1,0} = x_0$, and $y_{1,-1} = y_{1,0} = y_0$.
Set $W_N = \sum_{t=1}^N \omega_t$, $K_N = \sum_{t=1}^N S_t$, and $A_N = \sum_{t=1}^N \sum_{s=1}^{S_t} a_{t,s}^2$.
$\lambda_N \leftarrow (\Sigma_{A,\gamma}/\mathcal{D}_A)\sqrt{A_N/b}$.
Set $\beta_{t,s} \leftarrow \beta^{\mathrm{det}}_{t,s} + \lambda_N/a_{t,s}$ and $\gamma_{t,s} \leftarrow \gamma^{\mathrm{det}}_{t,s} + \lambda_N/(a_{t,s} r_{\mathrm{sl}}^2)$.
for $t = 1, \ldots, N$ **do**
    Form $\underline{x}_t$ by (54).
    Query all independent banks $A_{t,s}$ and $B_{t,s}$ in parallel; if $t \geq 2$, also query the reserve $A^-_{t,1}$.
    for $s = 1, \ldots, S_t$ **do**
        if $s = 1$ **then**
            Form $\widehat{r}^A_{t,1}$ by (56).
        else
            Form $\widetilde{y}_{t,s-1}$ and $\widehat{r}^A_{t,s}$ by (55).
        end if
        Compute $p_{t,s}$ by (57).
        Reveal $B_{t,s}$ and form $\widehat{q}^B_{t,s}(p_{t,s})$ by (53).
        Compute $y_{t,s}$ by (58).
    end for
    $x_t \leftarrow S_t^{-1} \sum_{s=1}^{S_t} p_{t,s}$ and $y_t \leftarrow S_t^{-1} \sum_{s=1}^{S_t} y_{t,s}$.
    Carry $p_{t+1,0} \leftarrow p_{t,S_t}$, $y_{t+1,0} \leftarrow y_{t,S_t}$, and $y_{t+1,-1} \leftarrow y_{t,S_t-1}$.
end for

Output: $\bar{z}_N \leftarrow W_N^{-1} \sum_{t=1}^N \omega_t(x_t, y_t)$.

No data-dependent estimate from the current phase is used to set λ_N. The additional terms $\lambda_N/a_{t,s}$ are not an assumption on a stochastic model. Their weighted contribution

is the same constant λ_N at every inner slot. Consequently they telescope across the sliding proof and only add one initial-radius term; at the same time they absorb the iterate-dependent part of the martingale perturbation. This is the mechanism that replaces the invalid uniform $O(1/r)$ model argument. If $\Sigma_{A,\gamma} = 0$, the conditional moment bounds make every bank error zero almost surely; the displayed rule then gives $\lambda_N = 0$ and Algorithm 2 reduces to its exact-bank deterministic analysis. Thus no division by λ_N is used in that degenerate case.

5.3 Certificate used by acceleration

Algorithm 2 is analyzed directly in terms of the localized primal certificate (19); inequalities (20) convert it to the two primal accuracy metrics.

Theorem 3 (Unconditional stochastic ACGD-S under vector-value feedback)**.** *Assume Assumptions 4 and 6. After smoothing, suppose $0 < L_{A,\gamma} < \infty$ is the supplied majorant in (48). Let the algorithm be supplied with deterministic upper bounds $\mathcal{D}_A > 0$ and $\Sigma_{A,\gamma} \geq 0$ satisfying (50) and (59). Run Algorithm 2 with independent common-randomness banks. Then there are universal constants $C_0, C_1, c_Q, c_D > 0$ such that*

$$\mathbb{E}\mathcal{C}_{R,\gamma}(\bar{x}_N) \leq C_0 \frac{L_{A,\gamma}\mathcal{D}_A^2}{N(N+1)} + C_1 \frac{\Sigma_{A,\gamma}\mathcal{D}_A}{\sqrt{bK_N}}. \tag{60}$$

Moreover,

$$K_N \leq c_D \left(N + \frac{\bar{M}_\gamma r_{\mathrm{sl}}}{L_{A,\gamma}} N(N+1) \right), \qquad Q_{\mathrm{vec}} \leq c_Q b K_N, \tag{61}$$

and all bS_t samples of one phase are issued after $\underline{x}_t$ is known but before the next outer center is formed. Hence the strict adaptive vector-query depth is $O(N)$, independent of b and S_t. No certified uniform random affine model is assumed in this theorem.

5.4 Target parameter choice

For a target smoothed primal certificate ε_s, first replace the supplied majorant, if necessary, by

$$L_{A,\gamma} \leftarrow \max\left\{ L_{A,\gamma}, \frac{\varepsilon_s}{\mathcal{D}_A^2} \right\}, \tag{62}$$

and use the raised value in all coefficients. Then one may choose

$$N = \max\left\{ 1, \left\lceil C\sqrt{\frac{L_{A,\gamma}\mathcal{D}_A^2}{\varepsilon_s}} \right\rceil \right\}, \qquad b = \max\left\{ 1, \left\lceil C\frac{\Sigma_{A,\gamma}^2 \mathcal{D}_A^2}{K_N \varepsilon_s^2} \right\rceil \right\}. \tag{63}$$

The order of selection is non-circular: first fix N from the first expression, then compute the deterministic schedule S_t and hence $K_N = \sum_t S_t$, and only then choose b from the second expression. Then

$$N_{\mathrm{seq}} = \mathcal{O}\left(1 + \sqrt{\frac{L_{A,\gamma}\mathcal{D}_A^2}{\varepsilon_s}} \right), \tag{64}$$

$$Q_{\mathrm{vec}} = \mathcal{O}\left(N_{\mathrm{seq}} + \frac{\bar{M}_\gamma r_{\mathrm{sl}} \mathcal{D}_A^2}{\varepsilon_s} + \frac{\Sigma_{A,\gamma}^2 \mathcal{D}_A^2}{\varepsilon_s^2} \right). \tag{65}$$

The next two corollaries allocate constant fractions of ε to smoothing bias, objective error, and constraint violation. Both use the natural majorant $L_{0,\gamma} + RL_{g,\gamma}$ with the target floor in (62).

Corollary 3 (Smooth accelerated value-only rate)**.** *Under the assumptions of Theorem 3 and Assumption 3, use* $L_{A,\gamma} = \max\{L_0 + RL_g, \varepsilon_s/\mathcal{D}_A^2\}$*, and*

$$N_{\mathrm{seq}} = \mathcal{O}\left(1 + \sqrt{\frac{(L_0 + RL_g)\mathcal{D}_A^2}{\varepsilon}}\right), \tag{66}$$

$$Q_{\mathrm{vec}} = \mathcal{O}\left(N_{\mathrm{seq}} + \frac{\bar{M}_\gamma r_{\mathrm{sl}} \mathcal{D}_A^2}{\varepsilon} + \frac{[dM_R^2 + \kappa_m r_{\mathrm{sl}}^2 (dD_X^2 M_g^2 + \sigma_h^2)]\mathcal{D}_A^2}{\varepsilon^2}\right). \tag{67}$$

In particular, for fixed problem scales the bank contribution is $\widetilde{\mathcal{O}}(d/\varepsilon^2)$*, where the only displayed logarithm in* m *is* κ_m*, while the adaptive depth is* $\mathcal{O}(\varepsilon^{-1/2})$*.*

Corollary 4 (Nonsmooth accelerated value-only rate)**.** *Under the assumptions of Theorem 3, set* $M_A = \max\{M_0, M_g\}$ *and*

$$\gamma_\varepsilon := \min\left\{\bar{\gamma}, \frac{\varepsilon}{3M_A}\right\}, \tag{68}$$

where the second entry is omitted when $M_A = 0$*, and take* $\gamma = \gamma_\varepsilon$*, as permitted by (45). Then when* $M_A > 0$ *the natural majorant before the target floor satisfies* $L_{A,\gamma} = \mathcal{O}(\sqrt{d}\, M_R/\gamma_\varepsilon)$ *up to the displayed problem scales; when* $M_A = 0$*, the natural majorant vanishes and the target floor is used. In either case the theorem and the preceding parameter choice give*

$$N_{\mathrm{seq}} = \mathcal{O}\left(1 + d^{1/4}\sqrt{\frac{M_R \mathcal{D}_A^2}{\gamma_\varepsilon \varepsilon}}\right),$$

$$Q_{\mathrm{vec}} = \mathcal{O}\left(N_{\mathrm{seq}} + \frac{\bar{M}_\gamma r_{\mathrm{sl}} \mathcal{D}_A^2}{\varepsilon} + \frac{[dM_R^2 + \kappa_m r_{\mathrm{sl}}^2 (dD_X^2 M_g^2 + \sigma_h^2)]\mathcal{D}_A^2}{\varepsilon^2}\right). \tag{69}$$

If $M_A > 0$ *and* $\bar{\gamma} \geq \varepsilon/(3M_A)$*, then the familiar specialization is* $N_{\mathrm{seq}} = \mathcal{O}(d^{1/4}/\varepsilon)$ *up to problem scales and the bank contribution is* $\widetilde{\mathcal{O}}(d/\varepsilon^2)$*. If* $\bar{\gamma}$ *is the active restriction, the displayed general bound instead has adaptive depth* $\mathcal{O}(d^{1/4}/\sqrt{\bar{\gamma}\varepsilon})$ *up to problem scales; no unconditional assertion* $\gamma = \Theta(\varepsilon)$ *is made. The stochastic term comes from the explicit bank variance (59), not from an assumed* $O(1/r)$ *model law.*

Remark 2 (What is new in the stochastic proof). The deterministic $O(1/N^2)$ ACGD-S telescoping remains the outer backbone, but every stochastic coefficient used by an inner prox step is fresh relative to the corresponding inner filtration. The two-bank split is essential: the primal point is formed from $A_{t,s}$ before the independent $B_{t,s}$ sample used in the dual step is revealed. The proof in Appendix A shows that the random perturbations contribute $O(\Sigma_{A,\gamma}\mathcal{D}_A/\sqrt{bK_N})$ to the primal certificate (19) after optimization of the added prox stabilizer. By (20), the same bound controls both smoothed objective residual and maximum shifted-constraint violation. Thus the accelerated depth and stochastic bank term in Corollaries 3 and 4 hold simultaneously in the stated value-only vector oracle model, with κ_m confined to the dual ℓ_∞ block and with the nonsmooth depth specialized to $d^{1/4}/\varepsilon$ only when $\bar{\gamma}$ does not bind.

6 A small-number-of-constraints route via the dual and Vaidya's method

Assume now that m is small, while the primal dimension d may be large. By a *route* we mean a reduction of the original problem to an auxiliary (here, dual) problem followed by a recovery

of the primal solution; the outer complexity of the reduction and the cost of the recovery are accounted for separately. Introduce a regularized primal Lagrangian

$$\mathcal{L}_\mu(x, y) = f_0(x) + \langle y, g(x)\rangle + \frac{\mu}{2}\|x - x_c\|_2^2, \qquad y \in Y, \tag{70}$$

where $\mu > 0$, $x_c \in X$, and $Y \subset \mathbb{R}^m_+$ is a known compact dual localization set containing an optimal regularized multiplier. Define

$$d_\mu(y) = \min_{x\in X} \mathcal{L}_\mu(x, y), \qquad \varphi_\mu(y) = -d_\mu(y), \qquad \varphi^\star_\mu := \min_{y\in Y} \varphi_\mu(y). \tag{71}$$

Because the inner objective is μ-strongly convex, its minimizer $x_\mu(y)$ is unique.

For the localized constrained problem we instantiate

$$R_\mu := 1 + \frac{M_0 D_X + \mu D_X^2/2}{\rho}, \qquad Y = Y_{R_\mu} = \{y \in \mathbb{R}^m_+ : \|y\|_1 \le R_\mu\}. \tag{72}$$

The Slater argument applied to the regularized primal objective bounds an optimal regularized multiplier by $(M_0 D_X + \mu D_X^2/2)/\rho$, so Y_{R_μ} contains one with unit slack. With $c_Y = (R_\mu/(2m))\mathbf{1}$, this body satisfies the Euclidean localization

$$B_{R_\mu/(2m)}(c_Y) \subseteq Y_{R_\mu} \subseteq B_{3R_\mu/2}(c_Y), \tag{73}$$

while its ℓ_1 diameter is at most $2R_\mu$. Hence the Vaidya logarithm below contains the corresponding dependence on m and R_μ; $\widetilde{\mathcal{O}}$ notation does not remove these geometric factors.

Assumption 7 (Value tails for the Vaidya route)**.** Conditional on the history before each fresh validation batch, the coordinate noises $F_i(x, \xi) - f_i(x)$, $i \in [m]$, are sub-Gaussian with proxy at most σ_g^2, uniformly over the queried x. In addition, uniformly over $x \in X$ and $y \in Y_{R_\mu}$, the centered Lagrangian value

$$F_0(x, \xi) + \langle y, F_{1:m}(x, \xi)\rangle - \big(f_0(x) + \langle y, g(x)\rangle\big)$$

is conditionally sub-Gaussian with proxy at most σ_φ^2. Fresh validation batches are independent of the stochastic samples used to construct the inexact inner point.

6.1 Inexact inner solutions produce inexact dual subgradients

For a convex function φ, a vector s is a δ-subgradient at y if

$$\varphi(z) \ge \varphi(y) + \langle s, z - y\rangle - \delta \quad \text{for all } z \in Y.$$

Let $D_Y = \sup_{y,z\in Y} \|y - z\|$ in the norm paired with the constraint-value error.

Theorem 4 (Inner error to dual oracle error)**.** *Suppose $\widetilde{x}(y)$ satisfies*

$$\mathcal{L}_\mu(\widetilde{x}(y), y) - d_\mu(y) \le \delta_{\text{in}}. \tag{74}$$

Then

$$-g(\widetilde{x}(y)) \in \partial_{\delta_{\text{in}}} \varphi_\mu(y). \tag{75}$$

If only an estimate $\widehat{g}$ is available and $\|\widehat{g} - g(\widetilde{x}(y))\|_* \le \tau$, then

$$-\widehat{g} \in \partial_{\delta_{\text{in}} + D_Y \tau} \varphi_\mu(y). \tag{76}$$

Theorem 4 formalizes why the inner problem need not be solved with progressively increasing accuracy. Vaidya's method tolerates an additive oracle error at every cut; the errors do not sum over the cuts [Gladin et al., 2021, Gladin et al., 2022].

Theorem 5 (Vaidya outer complexity with stochastic inner solves)**.** *Assume that Y is a full-dimensional compact convex body with known Euclidean localizations $B_{\rho_Y}(c_Y) \subseteq Y \subseteq B_{R_Y}(c_Y)$ (in particular, (73) for Y_{R_μ}), and let $B_\varphi := \sup_{y,z\in Y} |\varphi_\mu(y) - \varphi_\mu(z)| < \infty$. Fix a dual target ε_D and failure probability β, and let $c > 0$ be a sufficiently small universal constant. At every queried point y, suppose that (i) the cut vector is a δ_{cut}-subgradient with*

$$\delta_{\mathrm{cut}} = \delta_{\mathrm{in}} + D_Y \tau \le c\varepsilon_D, \tag{77}$$

and (ii) an incumbent value estimate $\widehat{\varphi}(y)$ is available with $|\widehat{\varphi}(y) - \varphi_\mu(y)| \le \eta_{\mathrm{val}}$, where $\eta_{\mathrm{val}} \le c\varepsilon_D$. If these properties hold at each cut with conditional failure probability at most β/N_V, then the inexact Vaidya method returns an incumbent $\widehat{y}$ satisfying $\varphi_\mu(\widehat{y}) - \varphi_\mu^\star = \mathcal{O}(\varepsilon_D)$ after

$$N_V = \mathcal{O}\left(m \log\left(1 + \frac{m^{3/2} R_Y B_\varphi}{\rho_Y \varepsilon_D} \right) + m \right) \tag{78}$$

outer cuts, with probability at least $1 - \beta$. The deterministic localization-plus-δ_{cut} estimate used here is Theorem 1 of [Gladin et al., 2022]; the noisy-incumbent correction is given below.

6.2 Noisy incumbent selection

The selection step needed in Theorem 5 is elementary. If y^{best} is the best queried point in true value and $\widehat{y}$ minimizes the noisy estimates, then

$$\varphi_\mu(\widehat{y}) \le \varphi_\mu(y^{\mathrm{best}}) + 2\eta_{\mathrm{val}}. \tag{79}$$

This elementary $2\eta_{\mathrm{val}}$ term is our additional selection argument. In particular, the admissible additive cut error is of order ε_D, not ε_D/N_V.

Suppose a Euclidean strongly convex stochastic zeroth-order inner method reaches (74) with high-probability vector-query complexity

$$Q_{\mathrm{in}}(\delta_{\mathrm{in}}, \beta') = \widetilde{\mathcal{O}}\left(\frac{dM_Y^2}{\mu\delta_{\mathrm{in}}} \log \frac{1}{\beta'} \right), \qquad M_Y := M_0 + \sup_{y\in Y} \sum_i y_i M_i. \tag{80}$$

The logarithm can be obtained from a sub-Gaussian tail bound or from independent repetitions with validation. With $\delta_{\mathrm{in}} = \Theta(\varepsilon_D)$ and $\beta' = \Theta(\beta/N_V)$, the total inner-solve cost is

$$Q_{\mathrm{inner,total}} = \widetilde{\mathcal{O}}\left(m \frac{dM_Y^2}{\mu\varepsilon_D} \right). \tag{81}$$

A replacement of d by a non-Euclidean second-moment factor is valid only when the regularizer is strongly convex in the same primal norm and the dual smoothness estimate is recomputed in the paired geometry; Euclidean quadratic regularization alone does not yield a logarithmic dimension factor.

Under Assumption 7, draw a fresh validation batch $\{\xi_j^{\mathrm{val}}\}_{j=1}^r$ after computing $\widetilde{x}(y)$ and form

$$\widehat{g}(y) = \frac{1}{r} \sum_{j=1}^r F_{1:m}(\widetilde{x}(y), \xi_j^{\mathrm{val}}), \tag{82}$$

$$\widehat{\varphi}(y) = -\left[\frac{1}{r} \sum_{j=1}^r \left(F_0(\widetilde{x}(y), \xi_j^{\mathrm{val}}) + \left\langle y, F_{1:m}(\widetilde{x}(y), \xi_j^{\mathrm{val}}) \right\rangle \right) + \frac{\mu}{2} \|\widetilde{x}(y) - x_c\|_2^2 \right]. \tag{83}$$

The deterministic inner gap implies $\varphi_\mu(y) - \delta_{\text{in}} \le -\mathcal{L}_\mu(\widetilde{x}(y), y) \le \varphi_\mu(y)$. Consequently, with conditional failure probability at most β/N_V, the cut and incumbent requirements are met by taking, up to universal constants,

$$r_g = \mathcal{O}\left(\frac{\sigma_g^2 D_Y^2}{\varepsilon_D^2} \log \frac{mN_V}{\beta}\right), \qquad r_\varphi = \mathcal{O}\left(\frac{\sigma_\varphi^2}{\varepsilon_D^2} \log \frac{N_V}{\beta}\right). \tag{84}$$

with $\delta_{\text{in}} = \Theta(\varepsilon_D)$. The same full-vector validation batch may be reused for both estimators; take its size $r \ge \max\{r_g, r_\varphi\}$. The two accuracy events may be dependent, but a union bound with a constant fraction of β/N_V assigned to each still gives total conditional failure probability at most β/N_V. Combining validation and inner solves gives the dual-target complexity

$$Q_{\text{Vaidya}} = \widetilde{\mathcal{O}}\left[m\left(\frac{dM_Y^2}{\mu\varepsilon_D} + \frac{\sigma_g^2 D_Y^2 + \sigma_\varphi^2}{\varepsilon_D^2}\right)\right]. \tag{85}$$

For Y_{R_μ}, substitution of $D_Y \le 2R_\mu$, $\rho_Y = R_\mu/(2m)$, and $R_Y \le 3R_\mu/2$ in (78) gives explicitly

$$N_V = \mathcal{O}\left(m\left[1 + \log\left(1 + \frac{3m^{5/2}B_\varphi}{\varepsilon_D}\right)\right]\right). \tag{86}$$

Thus the m-dependence hidden by $\widetilde{\mathcal{O}}(m)$ contains this logarithmic geometric factor. If the inner algorithm has sequential depth $N_{\text{in}}(\varepsilon_D)$, the end-to-end depth is $\widetilde{\mathcal{O}}(mN_{\text{in}}(\varepsilon_D))$, not merely $\widetilde{\mathcal{O}}(m)$. The validation batches are parallel within each cut. Thus Theorem 5 is a low-dimensional *dual optimization* route; end-to-end primal complexity is stated only after a recovery relation such as Corollary 5 is substituted.

6.3 Regularization bias

If the original objective is merely convex, the added term in (70) changes the primal objective by at most

$$\frac{\mu}{2} D_X^2. \tag{87}$$

Hence preserving an ε primal target requires $\mu \lesssim \varepsilon/D_X^2$. This substitution must be made in (80); regularization is not free.

6.4 Dual optimization is not the same as primal recovery

For this recovery paragraph, equip the dual space with the Euclidean norm and let Π_Y denote Euclidean projection onto Y. Uniqueness of $x_\mu(y)$ and Danskin's theorem imply differentiability of φ_μ with $\nabla\varphi_\mu(y) = -g(x_\mu(y))$. Assume this gradient is L_D-Lipschitz in Euclidean norm, where the least admissible constant may be zero. If

$$\|g(x) - g(z)\|_* \le \overline{M}_g \|x - z\|$$

in Euclidean norm, strong convexity of the inner objective gives $L_D \le \overline{M}_g^2/\mu$; in the affine case $g(x) = Ax - b$, this becomes $L_D \le \|A\|^2/\mu$. Let

$$\widehat{L}_D > 0, \qquad \widehat{L}_D \ge L_D, \qquad \mathcal{G}_{\widehat{L}_D}(y) = \widehat{L}_D\left(y - \Pi_Y\left(y - \frac{1}{\widehat{L}_D}\nabla\varphi_\mu(y)\right)\right) \tag{88}$$

be the projected-gradient mapping formed with any strictly positive smoothness majorant. This convention also covers the degenerate case $L_D = 0$, for which division by the least Lipschitz constant would be undefined. Smooth convexity gives

$$\left\|\mathcal{G}_{\widehat{L}_D}(y)\right\|^2 \le 2\widehat{L}_D(\varphi_\mu(y) - \varphi_\mu^\star). \tag{89}$$

Corollary 5 (Accuracy required for primal recovery). *Let $\mathcal{E}_{\mathrm{primal}} \geq 0$ be a primal error metric declared by the user of the reduction (for example, a Euclidean distance to the regularized primal minimizer, or an objective-residual metric, whenever the inequality (90) below can be verified for that choice). Suppose a recovery mapping satisfies*

$$\mathcal{E}_{\mathrm{primal}}(x_\mu(y)) \leq C_{\mathrm{rec}} \left\| \mathcal{G}_{\widehat{L}_D}(y) \right\|. \tag{90}$$

If $C_{\mathrm{rec}} > 0$, then a primal error at most ε is guaranteed by the dual target

$$\varepsilon_D \leq \frac{\varepsilon^2}{2\widehat{L}_D C_{\mathrm{rec}}^2}. \tag{91}$$

If $C_{\mathrm{rec}} = 0$ and $\mathcal{E}_{\mathrm{primal}} \geq 0$, the recovery inequality already forces zero primal recovery error, so no dual-gap restriction is needed. If the least smoothness constant is $L_D = 0$, the preceding statement is read with any declared $\widehat{L}_D > 0$ rather than by substituting zero in the denominator. Only under a stronger linear recovery inequality $\mathcal{E}_{\mathrm{primal}}(x_\mu(y)) \leq C(\varphi_\mu(y) - \varphi_\mu^\star)$ may one set $\varepsilon_D = \Theta(\varepsilon)$. If one takes a positive majorant $\widehat{L}_D = \mathcal{O}(1/\mu)$ and the regularization bias forces $\mu = \Theta(\varepsilon)$, the generic gradient-mapping rule can require $\varepsilon_D = \mathcal{O}(\varepsilon^3)$.

Thus the nonaccumulation property of Vaidya's method means that the inner oracle needs the same order of accuracy as the *dual* target at every cut. It does not imply that the dual target equals the desired primal target. In the nondegenerate case $C_{\mathrm{rec}} > 0$ and $\overline{M}_g > 0$, take $\widehat{L}_D = \overline{M}_g^2/\mu$. Then the generic gradient-mapping recovery and the regularization choice $\mu = \Theta(\varepsilon/D_X^2)$ give

$$\varepsilon_D = \mathcal{O}\left(\frac{\varepsilon^3}{\overline{M}_g^2 C_{\mathrm{rec}}^2 D_X^2} \right).$$

Substitution into (85) makes the inner-solve term scale as $\widetilde{\mathcal{O}}(\varepsilon^{-4})$ and the validation-value term as $\widetilde{\mathcal{O}}(\varepsilon^{-6})$ after normalization of the displayed problem scales. Thus low dual dimension alone is not claimed to improve the full primal stochastic complexity; a linear error-bound recovery permits the more favorable choice $\varepsilon_D = \Theta(\varepsilon)$.

7 Affine constraints and bilinear stochastic sliding

Consider the special case

$$\min_{x \in X} f_0(x) \quad \text{subject to} \quad Ax \leq b, \tag{92}$$

where $A \in \mathbb{R}^{m \times d}$ and $b \in \mathbb{R}^m$ are known exactly. On a bounded dual domain Y_R, this is the bilinear saddle problem

$$\min_{x \in X} \max_{y \in Y_R} f_0(x) + \langle y, Ax - b \rangle. \tag{93}$$

The objective oracle and products with $A, A^\top$ can now be scheduled at different frequencies. Fix the Euclidean initial point $z^0 = (x^0, y^0) \in X \times Y_R$ and let $D_A^2 := \sup_{z \in X \times Y_R} \|z - z^0\|_2^2$. Let $\|A\|$ be the Euclidean operator norm, and let the stochastic zeroth-order objective-gradient estimator have variance at most $\sigma_{\mathrm{zo}}^2 = \mathcal{O}(dM_0^2)$ under common-randomness feedback. Appendix A records the precise correspondence with Section K.2 and Corollary K.9 of Nguyen and Gasnikov [Nguyen, Gasnikov, 2026] and gives a self-contained proof for this degenerate bilinear specialization.

Theorem 6 (Affine stochastic sliding). *Fix $0 < \gamma \leq \bar{\gamma}$ and a smoothed-gap target $\varepsilon_s > 0$. Suppose $\nabla f_{0,\gamma}$ is $L_{0,\gamma}$-Lipschitz and the objective-only estimator from Lemma 2 is conditionally unbiased with variance at most σ_{zo}^2. Under the correspondence recorded in the Appendix,*

stochastic sliding applied to $f_{0,\gamma}(x)+\langle y, Ax-b\rangle$ *attains expected restricted saddle gap at most* ε_s *using*

$$Q_f = \mathcal{O}\left(\sqrt{\frac{L_{0,\gamma}D_A^2}{\varepsilon_s}} + \frac{\sigma_{\mathrm{zo}}^2 D_A^2}{\varepsilon_s^2}\right), \qquad Q_A = \mathcal{O}\left(\frac{\|A\|\, D_A^2}{\varepsilon_s}\right). \tag{94}$$

Here Q_f *counts zeroth-order objective-value calls, whereas* Q_A *counts exact products with* A *or* $A^\top$*. The two-point estimator targets* $\nabla f_{0,\gamma}$*; the transfer to the original objective is stated next.*

Corollary 6 (Smooth rate for the original affine problem)**.** *For an* L_0*-smooth objective, run Theorem 6 with smoothed-gap target* $\varepsilon/2$ *and choose*

$$0 < \gamma \le \min\left\{\bar{\gamma}, \frac{1}{\sqrt{6}}\sqrt{\frac{\varepsilon}{L_0}}\right\} \tag{95}$$

with the second entry omitted when $L_0 = 0$*. This convention allows any* $0 < \gamma \le \bar{\gamma}$ *in the zero-curvature case and avoids division by* L_0*. The method obtains expected original saddle gap at most* ε *with*

$$Q_f = \mathcal{O}\left(\sqrt{\frac{L_0 D_A^2}{\varepsilon}} + \frac{dM_0^2 D_A^2}{\varepsilon^2}\right), \qquad Q_A = \mathcal{O}\left(\frac{\|A\|\, D_A^2}{\varepsilon}\right). \tag{96}$$

Corollary 7 (Nonsmooth rate for the original affine problem)**.** *For an* M_0*-Lipschitz objective, run the theorem with a constant-fraction smoothed-gap target and set*

$$\gamma_{f,\varepsilon} = \min\left\{\bar{\gamma}, \frac{\varepsilon}{3M_0}\right\}, \tag{97}$$

where the second entry is omitted when $M_0 = 0$*. The nonsmooth rates are*

$$Q_f = \mathcal{O}\left(d^{1/4}\sqrt{\frac{M_0 D_A^2}{\gamma_{f,\varepsilon}\varepsilon}} + \frac{dM_0^2 D_A^2}{\varepsilon^2}\right), \qquad Q_A = \mathcal{O}\left(\frac{\|A\|\, D_A^2}{\varepsilon}\right). \tag{98}$$

If $M_0 > 0$ *and* $\varepsilon \le 3M_0\bar{\gamma}$*, then* $\gamma_{f,\varepsilon} = \varepsilon/(3M_0)$ *and*

$$Q_f = \mathcal{O}\left(\frac{d^{1/4}M_0\sqrt{D_A^2}}{\varepsilon} + \frac{dM_0^2 D_A^2}{\varepsilon^2}\right), \qquad Q_A = \mathcal{O}\left(\frac{\|A\|\, D_A^2}{\varepsilon}\right). \tag{99}$$

If $M_0 = 0$*, the objective terms vanish in the nonsmooth case. If* $\bar{\gamma}$ *binds, the general rate (98) applies. Appendix A proves the smoothing transfers and these degenerate conventions.*

Remark 3 (Componentwise comparison for the smooth affine specialization). The classical smooth lower bound and Proposition 1 match the two objective terms in their stated oracle classes. The coupling term matches Ouyang–Xu [Ouyang, Xu, 2021] only after balancing the weighted product geometry. Indeed, with $R_x = \sup_{x\in X}\|x-x^0\|_2$ and $R_y = \sup_{y\in Y_R}\|y-y^0\|_2$, the weighted norm gives $D_A^2(\alpha) = \alpha R_x^2 + \alpha^{-1}R_y^2$. For $R_xR_y > 0$, choosing $\alpha = R_y/R_x$ yields $D_A^2(\alpha) = 2R_xR_y$ and the balanced coupling bound $\mathcal{O}(\|A\|R_xR_y/\varepsilon)$. If either radius vanishes, the corresponding bilinear variation is degenerate. These are componentwise comparisons, not a joint minimax claim.

The $d^{1/4}/\varepsilon$ term in (99) is sequential work, whereas d/ε^2 is stochastic sample complexity. The result assumes that A and b are known and matrix products are exact; stochastic affine constraints require a separate estimator analysis.

8 Geometry, oracle lower bounds, and sharpness

8.1 General geometry factor

The Euclidean factor d is not universal. Let the primal norm be $\|\cdot\|_p$, $p \in [1,2]$, let q be its dual exponent, and let M_2 denote a uniform Euclidean Lipschitz bound for the sampled function. The smoothing estimator analyzed in [Gasnikov et al., 2022] satisfies a bound of the form

$$\mathbb{E}\left\|\widehat{\nabla} f_\gamma(x)\right\|_q^2 \le C\,\chi_p(d) M_2^2, \qquad \chi_p(d) = \mathcal{O}\left(\min\{q, 1+\log d\}\, d^{2/q}\right). \tag{100}$$

For $p = q = 2$, $\chi_2(d) = \mathcal{O}(d)$. For $p = 1$, $q = \infty$, one gets $\chi_1(d) = \mathcal{O}(1+\log d)$. Replacing (37) by this bound replaces the leading d in the total vector-query formulas by $\chi_p(d)$, provided the prox radius and operator Lipschitz constants are measured in the matching norms.

Randomization on the ℓ_1 sphere gives the alternative estimator [Akhavan et al., 2022]

$$\widehat{\nabla}^{(1)} f_\gamma(x;\zeta,\xi) = \frac{d}{2\gamma}\big(F(x+\gamma\zeta,\xi) - F(x-\gamma\zeta,\xi)\big)\operatorname{sign}(\zeta), \qquad \zeta \sim \mathrm{Unif}(\partial B_1^d). \tag{101}$$

Here $\mathrm{Unif}(\partial B_1^d)$ means normalized $(d-1)$-dimensional Hausdorff measure on the ℓ_1 sphere. Equivalently, draw independent signs $s_j \in \{-1,1\}$ and independent $E_j \sim \mathrm{Exp}(1)$, and set $\zeta_j = s_j E_j / (\sum_k E_k)$. This fixes the measure rather than leaving "uniform on the sphere" ambiguous. The estimator is unbiased for ℓ_1-ball smoothing and has favorable second-moment estimates under canceling noise. The constrained results transfer once its bias, smoothness, and second moment are substituted consistently. One must not combine the variance bound of one geometry with the prox radius or smoothing Lipschitz constant of another.

Corollary 8 (Favorable simplex geometry)**.** *Suppose X has an entropy prox with center-based radius $D_{\omega,X}^2 = \mathcal{O}(1+\log d)$, the sample functions satisfy the corresponding Lipschitz assumptions, and an estimator with second-moment factor $\chi(d) = \mathcal{O}(1+\log d)$ is used. Then the leading stochastic term in (41) is*

$$\mathcal{O}\left(\frac{M_R^2(1+\log d) + \kappa_m \sigma_h^2}{\varepsilon^2}\left[D_{\omega,X}^2 + R^2\log(m+1)\right]\right), \tag{102}$$

provided the Lipschitz constants, prox function, smoothing distribution, and dual norm are all taken in the same geometry. The Euclidean variance factor multiplying the paired-difference block is then replaced by $\mathcal{O}(1+\log d)$; under the generic finite-second-moment Assumption 5, the separate dual term retains the ℓ_∞-averaging factor κ_m. Under the stronger coordinate-wise sub-Gaussian condition of Lemma 4, it may instead be replaced by the direct averaged-coordinate specialization stated there. The prox radii remain explicit.

8.2 Oracle lower bounds and sharpness

The upper bounds should be compared only with lower bounds for the same feedback model. The following reductions isolate the objective and constraint information.

Proposition 1 (Hardness carried by the objective oracle)**.** *Consider the subclass of (1) with constant inactive constraints $f_i \equiv -\rho$. Any algorithm that returns an ε-solution on the constrained class also solves unconstrained stochastic two-point zeroth-order optimization of f_0. Under the paired-value second-moment alternative in Assumption 2, Proposition 1 of Duchi et al. [Duchi et al., 2015] applies to adaptive paired observations with a common sample and gives, in Euclidean geometry and under the standard normalization (the hard instances are*

scaled so that their Lipschitz constant is M_0 and the domain radius is D_X), an optimization error of order at least $M_0 D_X \sqrt{d/T}$. Hence T must be of order at least

$$\frac{d M_0^2 D_X^2}{\varepsilon^2}. \tag{103}$$

The hard family in this reduction consists of linear losses and therefore belongs to the smooth subclass as well. For general Lipschitz losses the matching two-point upper bound is due to Shamir [Shamir, 2017].

The bounded pathwise version of this reduction requires an additional adaptive-kernel argument. Appendix A.19 records the precise missing step. Because that step is not proved, no same-class minimax claim is based on the sketch.

Proposition 2 (Hardness carried by the constraint oracle)**.** *Let ϕ be an M_ϕ-Lipschitz stochastic convex function from a hard two-point family on X, shifted so that $\phi(x^{\mathrm{ref}}) = 0$ at a known reference point. On $Z = X \times [-M_\phi D_X, M_\phi D_X + \rho]$, consider*

$$\min_{(x,t)\in Z} t \quad \text{subject to} \quad g(x,t) = \phi(x) - t \le 0. \tag{104}$$

The objective is known and linear, while $(x^{\mathrm{ref}}, M_\phi D_X + \rho)$ is a known Slater point with margin at least ρ. Any random point satisfying expected objective residual and expected positive violation at most ε yields

$$\mathbb{E}[\phi(\widehat{x}) - \phi^\star] \le 2\varepsilon. \tag{105}$$

Taking ϕ from the same hard family yields a lower bound of order $d M_\phi^2 D_X^2/\varepsilon^2$, up to fixed product-domain constants. Hence the standard order-d/ε^2 hardness can be carried by the constraint-value component even when the objective requires no black-box queries.

The Slater margin ρ in this reduction is a fixed positive constant chosen independently of ε and d. Thus the d/ε^2 hardness already occurs for a well-conditioned Slater point; this proposition does not establish the necessity of the additional ρ^{-2} dependence entering M_R^2 in the upper bound.

Proposition 3 (A dual-value-noise lower bound)**.** *There is a one-dimensional feasible subclass with a fixed positive Slater margin for which any method satisfying both expected accuracy inequalities in Definition 1 needs*

$$Q_{\mathrm{vec}} = \Omega\left(\frac{\sigma_h^2}{\varepsilon^2}\right) \tag{106}$$

constraint-value observations under Gaussian additive dual-value noise of variance σ_h^2.

The proof is given in Appendix A.

Propositions 1–2 and 3 establish the optimal leading dependence on d and ε in the paired-value second-moment class of Assumption 2, up to problem scales and logarithmic factors, and show that objective and constraint values can be independently information-limiting. The bounded-support discussion in Appendix A.19 outlines how the same order might transfer to a pathwise-Lipschitz subclass, but its adaptive-kernel step (137) is not proved; hence no formal same-class minimax claim is based on that sketch. Proposition 3 separately shows that the additive σ_h^2/ε^2 order is unavoidable even with a fixed Slater margin. No claim of optimality is made for the complete factor $M_R^2 = (M_0 + R M_g)^2$, the accelerated R-weighted noise term, or a universal multiplicative dependence on m.

For the componentwise oracle, (7) is the exact cost of direct simulation, not a lower bound for every scalar-oracle algorithm. Even certifying a fixed candidate against m independently

noisy scalar constraints has an additive information cost proportional to $m\sigma^2/\varepsilon^2$ in a Gaussian testing model. Thus the general scalar-oracle optimum may lie between additive and multiplicative combinations of m and d.

The classical information-based framework of Nemirovski and Yudin [Nemirovski, Yudin, 1983] provides the baseline notion of oracle complexity used in this comparison. The smooth accelerated depth $\mathcal{O}(\sqrt{L_{A,\gamma}\mathcal{D}_A^2/\varepsilon})$ has the classical first-order dependence on ε. For nonsmooth parallel optimization, the local-oracle lower bounds of Nemirovski [Nemirovski, 1994], Bubeck et al. [Bubeck et al., 2019], and Diakonikolas and Guzmán [Diakonikolas, Guzmán, 2020] show that polynomially many parallel queries do not generally collapse the number of adaptive rounds. The $\mathcal{O}(d^{1/4}/\varepsilon)$ depth in Corollary 4 is the guarantee of the explicit phase-banked stochastic ACGD-S construction after shifted smoothing. No matching parallel lower bound is claimed for this adaptive vector-value oracle. Likewise, the smooth $\mathcal{O}(\varepsilon^{-1/2})$ depth matches the classical serial first-order dependence on ε but is not presented as a parallel value-oracle minimax theorem.

For affine constraints, Ouyang and Xu [Ouyang, Xu, 2021] prove a $1/\varepsilon$ lower complexity for bilinear coupling. As detailed in Remark 3, this matches the order of Q_A only after the weighted product geometry is balanced, turning D_A^2 into order R_xR_y; the raw unbalanced factor $\|A\|D_A^2$ is not itself called sharp. The stochastic objective term contains the rigorous unconstrained two-point lower-bound subclass in the paired second-moment model. Its bounded-pathwise analogue remains the proof sketch in Appendix A.19. The $\widetilde{\mathcal{O}}(m)$ number of Vaidya cuts is the natural cutting-plane dependence on the dual dimension, while the full route also includes the inner cost (85).

8.3 Open optimality gaps

The unresolved multiplier-radius, accelerated-noise, scalar-oracle, and parallel-depth questions are listed in Appendix B. They are scope boundaries, not assumptions used by the main theorems.

Table 2. Sharpness comparisons. The d/ε^2 lower bound is rigorous in the paired-value second-moment model; the bounded pathwise transfer remains a proof sketch. "Direct" means direct componentwise simulation.

Method or quantity	Upper bound	Comparison lower bound	Conclusion
Q_{vec}, Lipschitz	$\mathcal{O}((d\mathfrak{M}_R^2 + \kappa_m\sigma_h^2)D_\omega^2/\varepsilon^2)$	Propositions 1–2	sharp in d, ε in the paired second-moment class; bounded-pathwise transfer is a sketch
Q_{vec}, smooth	$\mathcal{O}((d\mathfrak{M}_R^2 + \kappa_m\sigma_h^2)D_\omega^2/\varepsilon^2)$	linear hard family in [Duchi et al., 2015]	sharp in d, ε; σ_h^2/ε^2 also has Proposition 3; no full R or logarithmic sharpness claim
Q_{cmp}, direct	$(m+1)Q_{\mathrm{vec}}$	additive scalar certification cost	direct simulation is not a universal scalar lower bound
Affine smooth terms	$\sqrt{L_0D_A^2/\varepsilon}$, $dM_0^2D_A^2/\varepsilon^2$; balanced $\|A\|R_xR_y/\varepsilon$	classical smooth, Proposition 1, Ouyang–Xu	objective comparison in paired class; coupling sharp only after balancing; no joint claim
Vaidya cuts	$\mathcal{O}(m[1+\log(1+3m^{5/2}B_\varphi/\varepsilon_D)])$	volumetric cutting-plane dimension dependence	dual-route outer cuts; primal complexity requires recovery substitution
PB–ACGD-S (this paper)	smooth depth $\mathcal{O}(\sqrt{L_{A,\gamma}\mathcal{D}_A^2/\varepsilon})$; nonsmooth depth $\mathcal{O}(d^{1/4}/\varepsilon)$ in the small-accuracy regime	classical serial first-order dependence in the smooth case; no matched lower bound in the nonsmooth vector-value model	unconditional upper bounds; the nonsmooth bound retains $\bar{\gamma}$; no parallel minimax claim

8.4 Relation to other first-order access models

Uniformly optimal parameter-free methods [Deng et al., 2024], sampled-constraint hinge-proximal methods [Rajoriya et al., 2025], cooperative stochastic approximation [Lan, Zhou, 2020], and stochastic SQP [Sanyal et al., 2025] use different information or subproblem oracles. Their rates therefore complement rather than dominate the vector value-oracle bounds above.

9 Numerical experiments

The experiments evaluate Algorithms 1 and 2 on synthetic problems whose exact constrained optimum is known. The complete script, instance data, seeds, raw trajectories, and summary tables are included with the source archive. All synthetic functions are defined on all of $\mathbb{R}^d$, so the enlarged PB-ACGD-S query-domain Assumption 6 creates no simulator-domain restriction in these tests.

9.1 Test problems and oracle

We use $d = 80$, $m = 10$, and $X = \{x : \|x\|_2 \le 1\}$. Let e_1 be the first coordinate vector and set $a = 0.75e_1$. The two objectives are

$$f_0^{\mathrm{sm}}(x) = \frac{1}{2}\|x - a\|_2^2, \qquad f_0^{\mathrm{ns}}(x) = \|x - a\|_2. \tag{107}$$

For this experiment write $g_i = f_i$ for $i \in [m]$. The first constraint is active, $g_1(x) = x_1 - 0.2$. Constraints $2, \dots, m-1$ are random affine functions of coordinates $2{:}d$, normalized to unit slope and shifted by -0.3. The final constraint is

$$g_m(x) = \frac{1}{2}\|x_{2:11}\|_2^2 - 0.05.$$

For both objectives,

$$x^\star = 0.2e_1, \qquad f_{0,\mathrm{sm}}^\star = \tfrac{1}{2}(0.55)^2, \qquad f_{0,\mathrm{ns}}^\star = 0.55. \tag{108}$$

The point $x^{\mathrm{s}} = 0$ has Slater margin $\rho = 0.05$. The exact active multipliers are 0.55 and 1, so the dual domain Y_2 contains a saddle point in both regimes.

A stochastic vector value has the form

$$F_i(x, \xi) = f_i(x) + \sigma_i \epsilon_i \langle q_i, x \rangle, \qquad \epsilon_i \in \{-1, 1\}, \tag{109}$$

where the fixed q_i are unit vectors, $\sigma_0 = 0.008$, and $\sigma_i = 0.003$ for constraints. The same Rademacher sample is used at $x + \gamma u$ and $x - \gamma u$. This gives a pathwise Lipschitz common-randomness oracle. The measurement-noise ablation adds independent Gaussian noise at the two perturbed points.

9.2 Implementation and metrics

We use a weighted Euclidean product prox, which is an admissible instance of Algorithm 1: the primal update is projected onto the unit ball and the dual update onto the nonnegative ℓ_1 ball of radius 2. The dual update is scaled by 20 relative to the primal update (equivalently, the quadratic dual prox weight is $1/20$), which balances primal and constraint units. We use $\gamma = 0.02$ in the smooth problem and $\gamma = 0.035$ in the nonsmooth problem. Batch sizes are $r \in \{1, 8, 32\}$, with 400 Mirror–Prox iterations and thirty independent evaluation seeds. The practical schedules are fixed across all reported evaluation seeds: $\eta_r = \min\{0.10, 0.060\sqrt{r}\}$ in

the smooth experiment and $\eta_r = \min\{0.024, 0.004\sqrt{r}\}$ in the nonsmooth experiment. They are implementation hyperparameters rather than theorem-optimal constants; no per-seed retuning is performed. The archive records the exact values and the synthetic optimum used only for reporting residuals, not for stopping the algorithm. The start is the infeasible unconstrained minimizer $x_1 = a$, $y_1 = 0$. Each operator sample uses three vector values, and each Mirror–Prox iteration uses two independent operator batches; hence iteration k costs $6rk$ vector values. Direct componentwise accounting is $(m + 1)$ times this number.

We report the median and interquartile range of

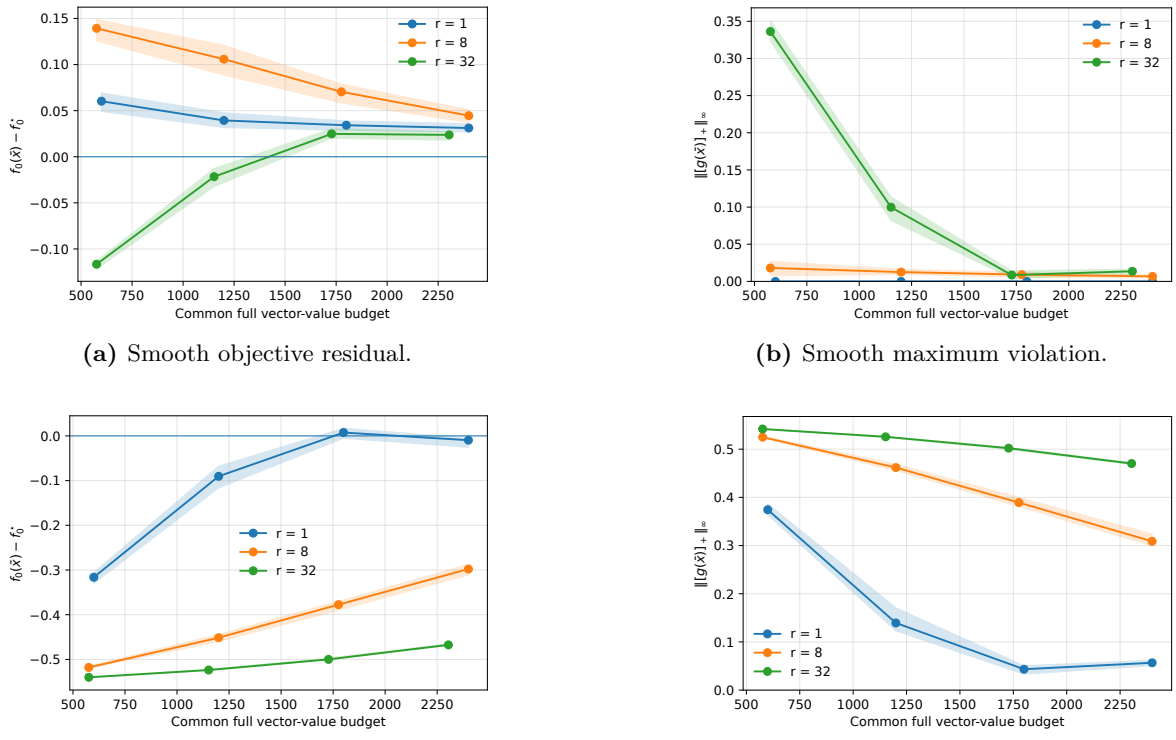
$$f_0(\bar{x}_k) - f_0^\star, \qquad \|[g(\bar{x}_k)]_+\|_\infty, \qquad J(\bar{x}_k) := \max\{[f_0(\bar{x}_k) - f_0^\star]_+, \|[g(\bar{x}_k)]_+\|_\infty\}.$$

The objective residual can be negative while $\bar{x}_k$ is infeasible; it must therefore be read jointly with the maximum-violation metric. The nonnegative joint metric J is used in every cross-method and scaling comparison below.

Fixed-horizon convergence panels and the complete terminal-metric table are reported in Appendix C. They show the expected depth–work tradeoff and also confirm that a negative objective residual must be read together with the constraint violation.

9.3 Equal-total-query comparison

Figure 1 reruns every batch size at common target budgets $Q \in \{600, 1200, 1800, 2400\}$, using exactly $N = \lfloor Q/(6r) \rfloor$ iterations and thirty seeds. Thus all curves occupy the common query range rather than terminating at batch-dependent horizons.



(a) Smooth objective residual.

(b) Smooth maximum violation.

(c) Nonsmooth objective residual.

(d) Nonsmooth maximum violation.

Figure 1. Equal-total-query comparison. Each point uses $N = \lfloor Q/(6r) \rfloor$ Mirror–Prox iterations, so batch size changes adaptive depth while the total vector-value budget is held fixed up to integer rounding.

At target $Q = 2400$, the nonsmooth violation is $5.67 \cdot 10^{-2}$ for $r = 1$, $3.09 \cdot 10^{-1}$ for $r = 8$, and $4.70 \cdot 10^{-1}$ for $r = 32$. Thus larger batches can improve a fixed sequential horizon without

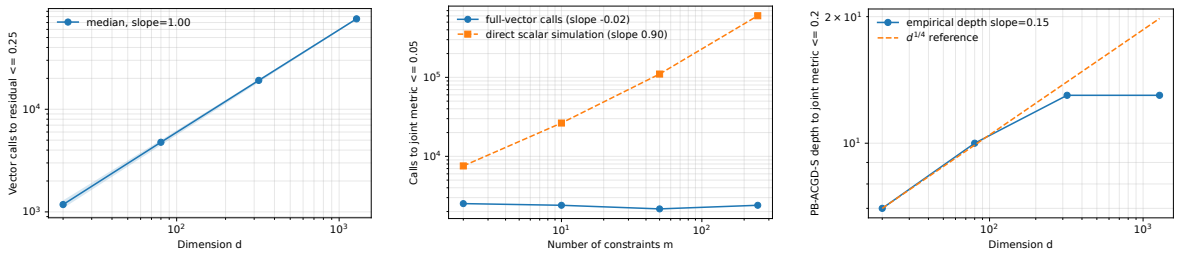
improving equal-budget query efficiency. The smooth results likewise show no uniformly dominant batch size.

9.4 Scaling in dimension and number of constraints

The two main oracle-accounting predictions are tested separately rather than inferred from the $d = 80, m = 10$ trajectories. For the dimension experiment we use the inactive-constraint linear hard subclass $f_0(x) = x_1$ on the Euclidean unit ball. Its common-randomness two-point estimator is du_1u, so this diagnostic isolates the variance-driven d factor while retaining exactly the paired vector-value semantics; each two-stage iteration is charged the same six vector values as Algorithm 1. For $d \in \{20, 80, 320, 1280\}$, sixteen seeds, and target residual 0.25, the median call counts are respectively $1179, 4770, 19104, 76044$. The fitted log–log slope is 1.002.

For the constraint-count experiment we return to the smooth constrained instance, vary $m \in \{2, 10, 50, 250\}$, and record first passage to $J \le 0.05$ for Mirror–Prox with $r = 8$. Across twelve seeds the median full-vector call counts are $2520, 2400, 2160, 2400$, yielding fitted slope -0.016. Charging the same trajectories by direct componentwise simulation gives $7560, 26400, 110160, 602400$ scalar values and slope 0.905. In this test the vector-round count is nearly constant in m, while direct scalar simulation grows almost linearly.

For PB-ACGD-S we also vary d in the nonsmooth regime and scale the supplied curvature input as $L_{A,\gamma} \propto \sqrt{d}$, as prescribed by the smoothing bound. Every one of the six runs reaches the coarse target $J \le 0.20$. For $d \in \{20, 80, 320, 1280\}$ the median adaptive depths are $7, 10, 13, 13$, with fitted finite-range slope 0.153. This is only a short-range sublinear diagnostic: the threshold is coarse and the data do not identify the theoretical exponent $1/4$.



(a) Vector-call scaling in d.

(b) Vector versus direct scalar scaling in m.

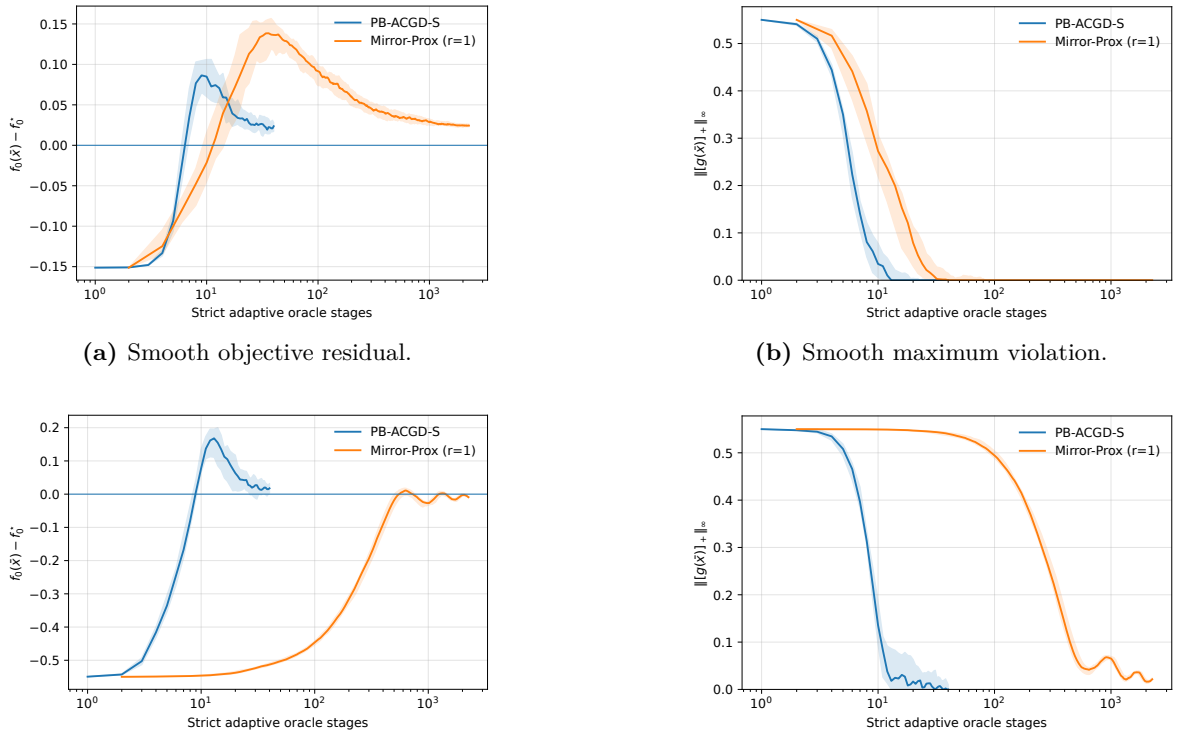
(c) PB–ACGD-S depth.

Figure 2. Scaling diagnostics. The d panel isolates the standard linear hard subclass, the m panel uses the constrained synthetic problem and the common joint metric J, and the PB panel reports first-passage phases under the Zhang–Lan center recursion.

9.5 Numerical check of PB-ACGD-S

The experiment runs Algorithm 2 on the same two synthetic problems and compares it with the low-work Mirror–Prox baseline $r = 1$. It uses the explicit bank recurrence of Definition 5 (the A/B split), the independent first-slot reserve, the Zhang–Lan center update (54) with stored $\underline{x}_{t-1}$, and actual vector-call accounting. The prescribed configuration has thirty seeds, $N = 40$ outer phases, batch $b = 8$, dual radius $R = 2$, sliding scale $r_{\mathrm{sl}} = 1$, Euclidean prox centers $x_0 = a$ and $y_0 = 0$, and $\lambda_N = 0.1$. The practical curvature inputs are $L_{A,\gamma} = 6$ in the smooth experiment and $L_{A,\gamma} = 12$ after nonsmooth smoothing, with $\bar{M}_\gamma = 1$. These constants are fixed before a run and do not use the true optimum or seed-specific trajectories. The implementation issues fixed-center samples sequentially for convenience. Since every query

location in a phase is known once $\underline{x}_t$ has been formed, the adaptive-depth count charges one bank stage per phase, as in Theorem 3.



(a) Smooth objective residual.

(b) Smooth maximum violation.

(c) Nonsmooth objective residual.

(d) Nonsmooth maximum violation.

Figure 3. PB-ACGD-S versus Mirror–Prox as a function of strict adaptive oracle stages. Curves are medians over thirty seeds with interquartile bands. One PB phase is one adaptive bank stage; a Mirror–Prox iteration has two sequential extragradient stages.

The results illustrate the work–depth distinction without showing query-efficiency dominance. At the smooth PB horizon of 6784 vector calls, PB–ACGD-S uses depth 40 and has median joint metric $J = 2.36 \cdot 10^{-2}$, whereas Mirror–Prox uses 6780 calls, depth 2260, and has $J = 2.44 \cdot 10^{-2}$. In the nonsmooth run, PB uses 4144 calls, depth 40, and reaches median $J = 2.33 \cdot 10^{-2}$; Mirror–Prox uses 4140 calls, depth 1380, and reaches $J = 2.22 \cdot 10^{-2}$. On this small instance, PB–ACGD-S therefore attains comparable coarse accuracy with far fewer adaptive stages and essentially the same full-vector budget.

A first-passage diagnostic gives the same qualitative picture. In the smooth experiment the median PB depths for $J \le 0.10, 0.08, 0.06, 0.05$ are $9, 11, 15.5, 16.5$ phases; all thirty seeds reach every threshold. In the nonsmooth experiment the medians for $J \le 0.12, 0.10, 0.08$ are $15, 18, 20$ phases, again with 30/30 successes. These ranges are too short for exponent estimation and are reported only as finite-horizon diagnostics.

Detailed fixed-horizon plots, high-noise tests, call-based PB–ACGD-S curves, parameter and radius sensitivity, the small-m Vaidya diagnostic, and the finite-difference noise ablation are collected in Appendix C. Keeping these secondary diagnostics outside the main narrative makes the work–depth comparison easier to read.

10 Conclusion

Under common-randomness vector feedback, shifted smoothing preserves feasibility and lets one random direction estimate the objective gradient and the full constraint Jacobian. Batched

Mirror–Prox has a paired zeroth-order contribution of order $\mathcal{O}(d/\varepsilon^2)$ in both smooth and Lipschitz nonsmooth problems. The general finite-moment analysis of the dual ℓ_∞ block adds the factor $1 + \log(2m)$ to its noise term, but no linear factor in m appears in the number of vector rounds.

PB–ACGD-S reduces smooth sequential depth to $\mathcal{O}(\varepsilon^{-1/2})$ while retaining leading vector work of order d/ε^2 up to logarithmic and fixed problem-scale factors. The Vaidya and affine reductions give alternatives for small dual dimension and known linear constraints. The lower bounds match the leading dependence on d and ε only in the paired-second-moment class; they do not establish optimal dependence on the Slater radius, a universal scalar-oracle dependence on m, or a parallel value-oracle lower bound. The experiments support the stated work and depth accounting but are not used as evidence of minimax optimality.

Acknowledgments

The research was supported by Russian Science Foundation (project No. 21-71-30005-), https://rscf.ru/en/project/21-71-30005/.

The authors acknowledge the use of AI-assisted writing tools, including OpenAI GPT models, Google Gemini, and Anthropic Claude (Sonnet), as well as coding assistants such as Codex and Claude Code. These tools were used solely to improve the clarity, grammar, organization, and presentation of the manuscript. All technical content, theoretical results, and conclusions are the original work of the authors. The authors reviewed and verified the entire manuscript and take full responsibility for its content.

A Proofs and technical details

A.1 Proof of Lemma 1

Proof. Convexity of f_i and $\mathbb{E}v = 0$ give Jensen's lower bound

$$f_i(x) = f_i(x + \gamma \mathbb{E}v) \le \mathbb{E} f_i(x + \gamma v) = f_{i,\gamma}(x).$$

Under Assumption 1,

$$f_{i,\gamma}(x) - f_i(x) \le \mathbb{E}\,|f_i(x + \gamma v) - f_i(x)| \le \gamma M_i,$$

which proves the nonsmooth bias bound. Standard ball-smoothing calculus gives the surface representation

$$\nabla f_{i,\gamma}(x) = \frac{d}{\gamma} \mathbb{E}_{u \sim \mathrm{Unif}(S_2^{d-1})}[f_i(x + \gamma u)u]. \tag{110}$$

For a unit vector a, subtracting the representations at x and z, using Lipschitz continuity, and then the estimate $\mathbb{E}\,|\langle u, a\rangle| \le d^{-1/2}$ up to a universal constant yields

$$\begin{aligned} |\langle \nabla f_{i,\gamma}(x) - \nabla f_{i,\gamma}(z), a\rangle| &\le \frac{d}{\gamma} \mathbb{E}\left[|f_i(x + \gamma u) - f_i(z + \gamma u)|\,|\langle u, a\rangle|\right] \\ &\le c_{\mathrm{sm}} \frac{\sqrt{d} M_i}{\gamma} \|x - z\|_2 . \end{aligned}$$

Taking the supremum over $\|a\|_2 = 1$ proves the displayed smoothness estimate. This is the standard estimate used in randomized smoothing; see also [Gasnikov et al., 2022, Nesterov, Spokoiny, 2017].

Under Assumption 3, the descent lemma gives

$$f_i(x + \gamma v) \le f_i(x) + \gamma \langle \nabla f_i(x), v\rangle + \frac{L_i \gamma^2}{2} \|v\|_2^2 .$$

Taking expectations and using $\mathbb{E}v = 0$ and $\|v\|_2 \le 1$ gives $f_{i,\gamma}(x) - f_i(x) \le L_i \gamma^2/2$. Averaging translations of an L_i-smooth function preserves the same gradient Lipschitz constant, so $L_{i,\gamma} \le L_i$.

Finally, if $f_i(x) \le 0$, then

$$h_{i,\gamma}(x) = f_{i,\gamma}(x) - b_i(\gamma) \le f_i(x) \le 0.$$

Applying this to $x = x^{\mathrm{s}}$ proves preservation of the Slater margin. □

A.2 Derivation of the multiplier radius

Proof. Let $x_\gamma^\star$ solve (12), and let $y_\gamma^\star$ be an optimal multiplier. Strong duality and saddle optimality give

$$f_{0,\gamma}(x_\gamma^\star) \le f_{0,\gamma}(x^{\mathrm{s}}) + \left\langle y_\gamma^\star, h_\gamma(x^{\mathrm{s}})\right\rangle \le f_{0,\gamma}(x^{\mathrm{s}}) - \rho \left\|y_\gamma^\star\right\|_1 .$$

Hence

$$\rho \left\|y_\gamma^\star\right\|_1 \le f_{0,\gamma}(x^{\mathrm{s}}) - f_{0,\gamma}(x_\gamma^\star).$$

The smoothed objective is M_0-Lipschitz on X, because it is an average of translations of an M_0-Lipschitz function. Therefore the right-hand side is at most $M_0 D_X$. □

Table 3. Frequently used quantities. Prox radii are measured from their specified initial centers; D_X is the pairwise Euclidean diameter.

Symbol	Meaning
d, m	primal dimension and number of functional constraints
D_X	Euclidean diameter of the primal domain X
R, Y_R	Slater-based ℓ_1 multiplier radius and localized dual domain
R_μ, Y_{R_μ}	regularized multiplier radius and dual domain for the Vaidya route
γ	randomized-smoothing and finite-difference radius
L_γ	Lipschitz constant of the smoothed saddle operator in the chosen product norm
$\kappa_m = 1 + \log(2m)$	logarithmic ℓ_∞-averaging factor for batching an ℓ_∞ block under finite second moments (Lemma 3)
$A_{t,s}, B_{t,s}, A^-_{t,1}$	sample bank of phase t: the primal and dual mini-batches and the cross-phase reserve (Definition 5)
slot (t, s)	one inner iteration of PB-ACGD-S; phase t has S_t slots
Σ_γ^2	batch-variance proxy: an average of r conditionally independent saddle-operator samples has conditional second moment at most Σ_γ^2/r
$D^2_{\omega,X}, D^2_{\omega,Y}, D^2_\omega$	primal, dual, and product center-based prox radii
$D^2_{\omega,X,\mathrm{ac}}, D^2_{\omega,Y,\mathrm{ac}}, \mathcal{D}_A$	PB primal, dual, and scaled-product center-based radii
N_{seq}	number of adaptive outer iterations; batches within an iteration are parallel
Q_{vec}	number of full vector-value calls $F(x, \xi) \in \mathbb{R}^{m+1}$
Q_{cmp}	number of scalar component calls in direct simulation of the vector algorithm
ε_D	target accuracy for the regularized dual problem in the Vaidya route
D_A^2	squared Euclidean product radius used in the affine sliding specialization

A.3 Proof of Theorem 1

Proof. Let $x^\star_\gamma$ and $y^\star_\gamma$ be a primal–dual solution of (12). Choosing $x = x^\star_\gamma$ and $y = 0$ in (17) gives

$$\begin{aligned}\mathrm{Gap}_{R,\gamma}(\widehat{z}) &\geq f_{0,\gamma}(\widehat{x}) - f_{0,\gamma}(x^\star_\gamma) - \langle \widehat{y}, h_\gamma(x^\star_\gamma)\rangle \\ &\geq f_{0,\gamma}(\widehat{x}) - f^\star_{0,\gamma},\end{aligned}$$

where the last inequality uses $\widehat{y} \geq 0$ and feasibility of $x^\star_\gamma$. Since $f_0(\widehat{x}) \leq f_{0,\gamma}(\widehat{x})$ and the original optimum $x^\star$ is feasible for the shifted problem,

$$f^\star_{0,\gamma} \leq f_{0,\gamma}(x^\star) \leq f^\star_0 + b_0(\gamma).$$

This proves the first inequality in (18).

For feasibility, let

$$v = \left\|[h_\gamma(\widehat{x})]_+\right\|_\infty.$$

If $v = 0$, then $f_i(\widehat{x}) \leq f_{i,\gamma}(\widehat{x}) = h_{i,\gamma}(\widehat{x}) + b_i(\gamma) \leq b_i(\gamma)$, and the second inequality in (18) follows. Suppose $v > 0$ and choose an index j with $h_{j,\gamma}(\widehat{x}) = v$ and $y = Re_j \in Y_R$. The saddle inequality for $y^\star_\gamma$ yields

$$f_{0,\gamma}(\widehat{x}) - f^\star_{0,\gamma} \geq -\langle y^\star_\gamma, h_\gamma(\widehat{x})\rangle \geq -\left\|y^\star_\gamma\right\|_1 v.$$

Consequently,

$$\begin{aligned}\mathrm{Gap}_{R,\gamma}(\widehat{z}) &\geq f_{0,\gamma}(\widehat{x}) - f^\star_{0,\gamma} + Rv - \langle \widehat{y}, h_\gamma(x^\star_\gamma)\rangle \\ &\geq (R - \left\|y^\star_\gamma\right\|_1)v \geq v,\end{aligned}$$

where the last inequality follows from $\left\|y_\gamma^\star\right\|_1 \le R-1$. Finally,

$$f_i(\widehat{x}) \le f_{i,\gamma}(\widehat{x}) = h_{i,\gamma}(\widehat{x}) + b_i(\gamma),$$

so $\|[g(\widehat{x})]_+\|_\infty \le v + b_g(\gamma)$. All inequalities are pointwise, hence remain valid after taking expectations. □

A.4 Proof of Lemma 2

Proof. For fixed i and ξ, the standard divergence-theorem identity for ball smoothing gives

$$\mathbb{E}_u\left[\frac{d}{2\gamma}\big(F_i(x+\gamma u,\xi) - F_i(x-\gamma u,\xi)\big)u\right] = \nabla_x \mathbb{E}_v[F_i(x+\gamma v,\xi)].$$

The two terms have equal expectation by symmetry of u. Taking expectation over ξ and interchanging expectation and differentiation, justified by the uniform Lipschitz bound, proves (23).

Fix $y \in Y_R$ and define the pathwise scalar function

$$\Psi_y(x,\xi) = F_0(x,\xi) + \sum_{i=1}^m y_i F_i(x,\xi).$$

It is convex and

$$|\Psi_y(x,\xi) - \Psi_y(z,\xi)| \le \left(M_0 + \sum_i y_i M_i\right)\|x-z\|_2 .$$

For later use in the accelerated proof, we record that convexity is *not* needed for the second-moment estimate. Let ψ be any M-Lipschitz real function on the required query neighborhood and, for u uniform on S^{d-1}, set

$$q(u) = \frac{\psi(x+\gamma u) - \psi(x-\gamma u)}{2\gamma}.$$

The function q is odd and M-Lipschitz on the sphere, so $\mathbb{E}q(u) = 0$. For $d \ge 2$, the spherical Poincaré inequality and Rademacher's theorem give

$$\mathbb{E}q(u)^2 = \mathrm{Var}(q(u)) \le \frac{1}{d-1}\mathbb{E}\|\nabla_S q(u)\|_2^2 \le \frac{M^2}{d-1}.$$

Consequently the central estimator $dq(u)u$ satisfies

$$\mathbb{E}\|dq(u)u\|_2^2 \le \frac{d^2}{d-1}M^2 \le 2dM^2, \qquad d \ge 2, \tag{111}$$

while for $d = 1$ the same conclusion holds directly with constant one. Thus the familiar $O(dM^2)$ bound for the two-point estimator follows from Lipschitzness alone; the convexity assumed in standard optimization statements such as [Gasnikov et al., 2022, Shamir, 2017] is not used in this variance calculation.

Applying (111) to Ψ_y yields

$$\mathbb{E}\left\|\widehat{\nabla}\Psi_{y,\gamma}(x)\right\|_2^2 \le c_{\mathrm{zo}} d\left(M_0 + \sum_i y_i M_i\right)^2 .$$

Since $y \ge 0$ and $\|y\|_1 \le R$, the last factor is at most M_R^2. More generally, for a signed coefficient vector a the same argument gives $c_{\mathrm{zo}} d(M_0 + \sum_i |a_i| M_i)^2$; this is the form used for the extrapolated multiplier in Step 2 of Theorem 3. The use of one common pair of vector values is crucial: estimating every row from independent oracle calls would lead to a different variance and query count. □

A.5 Proof of Lemma 3

Proof. All expectations in this proof are conditional on $\mathcal{H}$; we suppress this conditioning in the notation. Let $\xi'_1, \dots, \xi'_r$ be a conditionally independent copy of the batch. Conditional centering, Jensen's inequality, and the usual symmetrization identity give

$$\begin{aligned}\mathbb{E}\left\|\sum_{j=1}^{r}\xi_j\right\|_\infty^2 &\le \mathbb{E}\left\|\sum_{j=1}^{r}(\xi_j-\xi'_j)\right\|_\infty^2 \\ &= \mathbb{E}\mathbb{E}_\varepsilon\left\|\sum_{j=1}^{r}\varepsilon_j(\xi_j-\xi'_j)\right\|_\infty^2,\end{aligned}$$

where $\varepsilon_1, \dots, \varepsilon_r$ are independent Rademacher signs. Conditional on the two batches, each coordinate of the Rademacher sum is sub-Gaussian with variance proxy at most

$$\sum_{j=1}^{r}(\xi_{j,i}-\xi'_{j,i})^2 \le \sum_{j=1}^{r}\|\xi_j-\xi'_j\|_\infty^2.$$

The Rademacher maximal inequality over the m coordinates therefore yields

$$\mathbb{E}_\varepsilon\left\|\sum_{j=1}^{r}\varepsilon_j(\xi_j-\xi'_j)\right\|_\infty^2 \le C\log(2m)\sum_{j=1}^{r}\|\xi_j-\xi'_j\|_\infty^2.$$

Finally, $\|\xi_j-\xi'_j\|_\infty^2 \le 2\|\xi_j\|_\infty^2 + 2\|\xi'_j\|_\infty^2$. Taking expectation, absorbing numerical constants into C_{bt}, dividing by r^2, and using $\log(2m) \le \kappa_m$ proves the first inequality in (27). The second follows by bounding the sum of the r conditional moments by r times their maximum. □

A.6 Verification of the operator Lipschitz bound

Proof. Let $z = (x, y)$ and $z' = (x', y')$. For the primal component of (34), add and subtract $\sum_i y_i \nabla h_{i,\gamma}(x')$ to obtain

$$\begin{aligned}&\left\|\nabla f_{0,\gamma}(x) - \nabla f_{0,\gamma}(x') + Jh_\gamma(x)^\top y - Jh_\gamma(x')^\top y'\right\|_2 \\ &\quad\le (L_{0,\gamma} + RL_{g,\gamma})\left\|x-x'\right\|_2 + M_g\left\|y-y'\right\|_1.\end{aligned}$$

Here $\|\nabla h_{i,\gamma}(x')\|_2 \le M_i \le M_g$, because smoothing preserves Lipschitz continuity. For the dual component,

$$\left\|h_\gamma(x) - h_\gamma(x')\right\|_\infty \le M_g\left\|x-x'\right\|_2.$$

Combining the two inequalities with $(a+b)^2 \le 2a^2+2b^2$ proves (35) with a universal constant c_L. Substitution of the smoothing bounds in Lemma 1 gives (36). □

A.7 Verification of the batch-variance proxy

Condition on the history and on a fixed query point $z = (x, y)$. Write one centered operator error as $\delta_j = (\delta_j^x, \delta_j^h)$ in the Euclidean primal and ℓ_∞ dual blocks. Conditional independence and centering give the Hilbert-space identity

$$\mathbb{E}\left\|\frac{1}{r}\sum_{j=1}^{r}\delta_j^x\right\|_2^2 = \frac{1}{r^2}\sum_{j=1}^{r}\mathbb{E}\|\delta_j^x\|_2^2 \le \frac{c_{\mathrm{zo}}dM_R^2}{r}.$$

For the dual block, Assumption 5 and Lemma 3 instead give

$$\mathbb{E}\left\|\frac{1}{r}\sum_{j=1}^{r}\delta_j^h\right\|_\infty^2 \le \frac{C_{\mathrm{bt}}\kappa_m\sigma_h^2}{r}.$$

The squared product dual norm is the sum of these two squared block norms, which proves (37). This calculation pinpoints why the one-step ℓ_∞ moment in (26) cannot be replaced by σ_h^2/r without the logarithmic ℓ_∞-averaging factor. Under the stronger coordinatewise hypothesis of Lemma 4, applying the maximal inequality directly after averaging gives (29) and the stated sharper proxy.

A.8 Proof of Theorem 2

Proof. Write the two independent batched oracle errors in iteration k as

$$\widehat{G}_{\gamma,r}(z_k) - G_\gamma(z_k) = \Delta_k, \qquad \widehat{G}_{\gamma,r}(w_k) - G_\gamma(w_k) = \Delta_k'.$$

Let $\mathcal{F}_k$ be the history just before the first batch and let $\mathcal{F}_k^+$ include that batch and the resulting point w_k, but not the second batch. The errors are centered conditionally on $\mathcal{F}_k$ and $\mathcal{F}_k^+$, respectively. The Euclidean part of each batch averages by orthogonality, and the ℓ_∞ part averages by Lemma 3; hence (37) gives $\mathbb{E}[\|\Delta_k\|_{Z,*}^2 \mid \mathcal{F}_k] \le \Sigma_\gamma^2/r$ and $\mathbb{E}[\|\Delta_k'\|_{Z,*}^2 \mid \mathcal{F}_k^+] \le \Sigma_\gamma^2/r$.

We invoke the standard stochastic Mirror–Prox theorem of Juditsky, Nemirovski, and Tauvel [Juditsky et al., 2011]. In the notation used here, its proof combines the prox three-point inequality with an auxiliary mirror-descent sequence that controls the supremum of the martingale linear forms. This auxiliary sequence is essential: one cannot take a supremum over the comparison point and then simply declare the resulting martingale term to have zero expectation. With $V_Z(z_1, z) \le D_\omega^2$ and $\eta \le 1/(2L_\gamma)$, the theorem yields

$$\mathbb{E}\,\mathrm{Gap}_{R,\gamma}(\bar{z}_N) \le \frac{C_0 D_\omega^2}{\eta N} + C_1\eta\frac{\Sigma_\gamma^2}{r} + C_2\sqrt{\frac{\Sigma_\gamma^2 D_\omega^2}{Nr}}, \tag{112}$$

for universal constants C_0, C_1, C_2. The last term is precisely the expected supremum of the controlled martingale component.

If $L_\gamma = \Sigma_\gamma = 0$, (112) reduces to $\mathbb{E}\,\mathrm{Gap}_{R,\gamma}(\bar{z}_N) \le C_0 D_\omega^2/(\eta N)$ for every $\eta > 0$ because the stability restriction is vacuous. The step-size convention in Theorem 2 therefore gives the target gap with $N = r = 1$. Assume below that $L_\gamma + \Sigma_\gamma > 0$.

Choose

$$\eta = \min\left\{\frac{1}{2L_\gamma}, c_\eta\sqrt{\frac{D_\omega^2 r}{\Sigma_\gamma^2 N}}\right\}$$

with a sufficiently small universal c_η, omitting an entry with zero denominator. Then the first two terms in (112) are bounded, up to universal constants, by $L_\gamma D_\omega^2/N + \sqrt{\Sigma_\gamma^2 D_\omega^2/(Nr)}$. This proves (39). Conditions (40) make both contributions at most a constant fraction of ε_s. Finally, choose $N = \max\{1, \lceil CL_\gamma D_\omega^2/\varepsilon_s\rceil\}$ and

$$r = \max\left\{1, \left\lceil C\frac{\Sigma_\gamma^2 D_\omega^2}{N\varepsilon_s^2}\right\rceil\right\}.$$

Then $Nr = \mathcal{O}(L_\gamma D_\omega^2/\varepsilon_s + \Sigma_\gamma^2 D_\omega^2/\varepsilon_s^2 + 1)$, which gives (41) after absorbing the trivial one-call term. Fixed factors coming from the two Mirror–Prox oracle points and the constant number of vector values per operator sample are absorbed into C. □

A.9 Proofs of Corollaries 1 and 2

Proof. In the smooth regime, Assumption 4 supplies the dual radius, while Lemma 1 gives $L_{i,\gamma} \le L_i$ and bias $b_i(\gamma) \le L_i\gamma^2/2$. The choice $\gamma \le \bar{\gamma}$ makes all oracle points valid and makes both the objective and constraint biases at most $\varepsilon/3$. Substitute $L_\gamma = \mathcal{O}(L_0 + RL_g + M_g)$ and $\Sigma_\gamma^2 = \mathcal{O}(dM_R^2 + \kappa_m\sigma_h^2)$ into Theorem 2, set $\varepsilon_s = \varepsilon/3$, and use Theorem 1. Under Lemma 4, equation (29) gives the sharper dual specialization stated in the corollary.

In the nonsmooth regime, the exact choice (45) makes $b_0(\gamma), b_g(\gamma) \le \varepsilon/3$. Equation (36) gives

$$L_\gamma = \mathcal{O}\left(\sqrt{d}\, M_R \max\left\{\frac{1}{\bar{\gamma}}, \frac{3M_{\max}}{\varepsilon}\right\} + M_g\right).$$

Substituting this and (37) into (41) proves the general bounds (46)–(47). If $M_{\max} > 0$ and $\varepsilon \le 3M_{\max}\bar{\gamma}$, the second entry in the maximum is active and gives the stated small-accuracy specialization. The componentwise relation follows from (7). □

A.10 Proof of Theorem 3

Proof. We separate the argument into the deterministic ACGD-S telescoping, conditional moments of the phase banks, stochastic stabilization, and the arithmetic of the sliding weights.

Step 1: deterministic weighted energy and the exact ACGD-S mapping. Write $W_N = \sum_{t=1}^N \omega_t = N(N+1)/2$ and $a_{t,s} = \omega_t/S_t$. Let $u_t = \underline{x}_t$ and introduce the exact affine model

$$\ell_t(x,y) = f_{0,\gamma}(u_t) + \langle \nabla f_{0,\gamma}(u_t), x - u_t\rangle + \langle y, h_\gamma(u_t) + Jh_\gamma(u_t)(x-u_t)\rangle\,. \tag{113}$$

Here is the exact variable map to Algorithm 3 in Zhang–Lan v4 [Zhang, Lan, 2025]. Their query center $\underline{x}_t$, extrapolated point $\widetilde{x}_t$, and outer phase output x_t are the variables with the same names here; their inner primal point $y_s^{(t)}$ is our $p_{t,s}$, and their inner multiplier $\lambda_s^{(t)}$ is our $y_{t,s}$. Their phase carry $y_0^{(t+1)} = y_{S_t}^{(t)}$, $\lambda_0^{(t+1)} = \lambda_{S_t}^{(t)}$, and $\lambda_{-1}^{(t+1)} = \lambda_{S_t-1}^{(t)}$ is exactly our carry of $p_{t+1,0}, y_{t+1,0}, y_{t+1,-1}$. Most importantly, the v4 outer-center recursion is

$$\underline{x}_0 = x_0, \qquad \widetilde{x}_t = x_{t-1} + \theta_t(x_{t-1} - x_{t-2}), \qquad \underline{x}_t = \frac{\tau_t \underline{x}_{t-1} + \widetilde{x}_t}{1+\tau_t}, \tag{114}$$

which is (54); the first term in its numerator is the previous *query center*, not the previous phase output. Finally, set their smooth objective to $f_{0,\gamma}$, their constraint map to h_γ, their nonsmooth regularizer to zero, their phase data to $\pi_t = \nabla f_{0,\gamma}(\underline{x}_t)$ and $\nu_t = Jh_\gamma(\underline{x}_t)$, and their phase averages to our (x_t, y_t). Their Euclidean primal prox is precisely $V_X = \|\cdot\|_2^2/2$; replacing the Euclidean dual square by the one-strongly-convex V_Y used here invokes the identical three-point inequality in the matched ℓ_1/ℓ_∞ norms. Fenchel equality at a gradient point turns their nested-Lagrangian Q-term into the affine saddle difference generated by (113). Restricting the multiplier prox from $\mathbb{R}_+^m$ to the convex set Y_R leaves every prox inequality valid for the comparators $y \in Y_R$ used here.

Let $r_{t,s}$ denote the exact version of $\widehat{r}_{t,s}^A$ obtained by replacing every bank average by its conditional mean, and set $q_{t,s}(x) := h_\gamma(u_t) + Jh_\gamma(u_t)(x - u_t)$. We record the local inequalities because they are also where the stochastic perturbations enter. For every $x \in X$, the three-point inequality associated with (57) gives

$$\begin{aligned}\langle r_{t,s}, p_{t,s} - x\rangle \le{}& \eta_t\{V_X(x_{t-1},x) - V_X(x_{t-1},p_{t,s}) - V_X(p_{t,s},x)\}\\ &+ \beta_{t,s}\{V_X(p_{t,s-1},x) - V_X(p_{t,s-1},p_{t,s}) - V_X(p_{t,s},x)\}.\end{aligned} \tag{115}$$

Likewise, for every $y \in Y_R$, the exact dual prox step gives

$$\langle y - y_{t,s}, q_{t,s}(p_{t,s})\rangle \le \gamma_{t,s}\{V_Y(y_{t,s-1},y) - V_Y(y_{t,s-1},y_{t,s}) - V_Y(y_{t,s},y)\}. \tag{116}$$

The extrapolation cross terms are controlled in the matched norms by $\langle Jh_\gamma(u)d_x, d_y\rangle \le \bar{M}_\gamma\|d_x\|_2\|d_y\|_1$ and Young's inequality. The deterministic coefficients satisfy the exact coupling conditions

$$\gamma^{\mathrm{det}}_{t,s}\beta^{\mathrm{det}}_{t,s+1} \ge \rho_{t,s+1}\bar{M}^2_\gamma, \qquad \gamma^{\mathrm{det}}_{t-1,S_{t-1}}\beta^{\mathrm{det}}_{t,1} \ge \rho_{t,1}\bar{M}^2_\gamma, \tag{117}$$

whenever the indicated indices exist. Indeed, $\widetilde{M}_t \ge \bar{M}_\gamma$, while $\rho_{t,1} = \widetilde{M}_t/\widetilde{M}_{t-1}$; hence the second inequality reduces to $\widetilde{M}^2_{t-1} \ge \bar{M}^2_\gamma$. These are the norm-matched versions of the intra- and inter-phase conditions in Proposition 3 of Zhang–Lan. Equations (115)–(117) show explicitly that only one-strong convexity of the prox functions and the matched operator norm are used in the inner cancellation.

The outer smoothness step is also explicit. For every $p \in X$ and $y \in Y_R$,

$$\Phi_\gamma(p,y) - f^\star_{0,\gamma} \le \ell_t(p,y) - \ell_t(x^\star_\gamma, y_{t,s}) + \frac{L_{A,\gamma}}{2}\|p - u_t\|^2_2. \tag{118}$$

The first term is an upper model because $f_{0,\gamma} + \langle y, h_\gamma\rangle$ is $L_{A,\gamma}$-smooth for $y \in Y_R$. For the comparator term, convexity gives $\ell_t(x^\star_\gamma, y_{t,s}) \le f_{0,\gamma}(x^\star_\gamma) + \langle y_{t,s}, h_\gamma(x^\star_\gamma)\rangle \le f^\star_{0,\gamma}$ because $y_{t,s} \ge 0$ and $x^\star_\gamma$ is feasible.

Multiply the slot inequalities by $a_{t,s}$ and sum first over s and then over phases. The extrapolation terms cancel by (117). The primal and dual boundary divergences telescope because, for the conservative schedule,

$$a_{t,s}\beta^{\mathrm{det}}_{t,s} = L_{A,\gamma}, \qquad a_{t,s}r^2_{\mathrm{sl}}\gamma^{\mathrm{det}}_{t,s} = L_{A,\gamma} \quad \text{for every } (t,s). \tag{119}$$

The remaining outer smoothness terms in (118) cancel against the accelerated potential for $\omega_t = t$, $\tau_t = (t-1)/2$, and $\eta_t = L_{A,\gamma}/\tau_{t+1} = 2L_{A,\gamma}/t$, which is exactly the non-strongly-convex outer schedule in Theorem 5, equation (4.4), of Zhang–Lan. In particular, the first outer condition in their Proposition 3 is now literal: $\omega_t\eta_t = 2L_{A,\gamma} = \omega_{t-1}\eta_{t-1}$ for $t \ge 2$; moreover $\eta_{t-1}\tau_t = L_{A,\gamma} \ge (\omega_{t-1}/\omega_t)L_{A,\gamma}$ and $\eta_N(\tau_N + 1) = L_{A,\gamma}(N+1)/N \ge L_{A,\gamma}$. Hence there is no positive outer-prox leakage term. The initial primal and dual potentials are bounded by $CL_{A,\gamma}\mathcal{D}^2_A$ because (119) also holds at $(1,1)$. More precisely, the summed inequality holds for every fixed comparator $y \in Y_R$, and its right-hand side is uniform in that comparator because the initial dual divergence is bounded by the radius used in $\mathcal{D}^2_A$. Therefore convexity of $x \mapsto \Phi_\gamma(x,y)$, the definition $\bar{x}_N = W_N^{-1}\sum_t \omega_t x_t$, and only then taking $\sup_{y\in Y_R}$ imply

$$W_N\mathcal{C}_{R,\gamma}(\bar{x}_N) \le C_{\mathrm{det}}L_{A,\gamma}\mathcal{D}^2_A \tag{120}$$

in the exact-oracle case. Thus the deterministic backbone used below controls the primal certificate (19), not a stronger full saddle gap.

The stochastic stabilizer preserves all deterministic coupling inequalities. Indeed,

$$a_{t,s}\big(\beta_{t,s} - \beta^{\mathrm{det}}_{t,s}\big) = \lambda_N, \qquad a_{t,s}r^2_{\mathrm{sl}}\big(\gamma_{t,s} - \gamma^{\mathrm{det}}_{t,s}\big) = \lambda_N. \tag{121}$$

Hence (119) becomes the same constant $L_{A,\gamma} + \lambda_N$ at every slot. All coupling inequalities become weaker, while the additional boundary divergences telescope to at most $C\lambda_N\mathcal{D}^2_A$ and leave negative step-divergence terms that will absorb the iterate-dependent martingale part.

Step 2: conditional centering of the bank errors. Define $e^x_{t,s} := \widehat{r}^A_{t,s} - r_{t,s}$, $e^y_{t,s} := \widehat{q}^B_{t,s}(p_{t,s}) - q_{t,s}(p_{t,s})$, and $e_{t,s} := (e^x_{t,s}, e^y_{t,s})$ in the scaled product dual norm. Order inner slots lexicographically. Let $\mathcal{H}^-_{t,s}$ contain the complete outer history, all bank entries revealed before slot (t,s), and all iterates available before the $A_{t,s}$ batch is revealed. Let $\mathcal{H}^+_{t,s}$ additionally contain $A_{t,s}$ and the primal point generated from it, but not $B_{t,s}$. The multiplier combination used in the primal/extrapolation step is $\mathcal{H}^-_{t,s}$-measurable. The extrapolated multiplier has ℓ_1 norm

at most $5R$, as shown after (58). Applying the Lipschitz-only bound (111) from the proof of Lemma 2 to this possibly signed multiplier therefore gives

$$\mathbb{E}[e^x_{t,s} \mid \mathcal{H}^-_{t,s}] = 0, \qquad \mathbb{E}[\|e^x_{t,s}\|_2^2 \mid \mathcal{H}^-_{t,s}] \le \frac{cd(M_0 + 5RM_g)^2}{b} \le \frac{25cdM_R^2}{b}. \tag{122}$$

The fresh previous-center entry used in the first-slot extrapolation obeys the same conditional statement and changes only the universal constant.

The primal inner point is $\mathcal{H}^+_{t,s}$-measurable, while the $B_{t,s}$ bank is independent of $\mathcal{H}^+_{t,s}$. Hence the error in the dual linearized coefficient satisfies

$$\mathbb{E}[e^y_{t,s} \mid \mathcal{H}^+_{t,s}] = 0. \tag{123}$$

We first justify the query-domain bound using the exact recursion (114). For $t \ge 2$, $\theta_t = t/(t-1) \le 2$ and $x_{t-1}, x_{t-2} \in X$, so $\widetilde{x}_t \in X + 2D_X B_2^d$; this also holds at $t = 1$, when $\widetilde{x}_1 = x_0$. The Minkowski sum $X + 2D_X B_2^d$ is convex. Starting from $\underline{x}_0 = x_0$, induction in (114) therefore yields $\underline{x}_t \in X + 2D_X B_2^d$ for every t. Thus, for the actual slot point $p_{t,s} \in X$, writing $\underline{x}_t = x' + v$ with $x' \in X$ and $\|v\|_2 \le 2D_X$ gives $\|p_{t,s} - \underline{x}_t\|_2 \le D_X + 2D_X = 3D_X$.

Consider first one centered Jacobian-action sample. The common-direction two-point formula and pathwise Lipschitzness imply the one-sample estimate described below. Applying the conditional ℓ^m_∞ batching inequality (27) to the b independent samples then gives

$$\mathbb{E}\left[\|[\widehat{J}_\gamma(\underline{x}_t) - Jh_\gamma(\underline{x}_t)](p_{t,s} - \underline{x}_t)\|_\infty^2 \mid \mathcal{H}^+_{t,s}\right] \le \frac{c\kappa_m d M_g^2 D_X^2}{b}. \tag{124}$$

To see the dimension factor directly, for every row i the sampled gradient has the form $d\alpha_i(u)u$ with $|\alpha_i(u)| \le M_i$; therefore $\max_i |d\alpha_i(u)\langle u, p_{t,s} - \underline{x}_t\rangle|^2 \le d^2 M_g^2 |\langle u, p_{t,s} - \underline{x}_t\rangle|^2$. Spherical isotropy gives expectation at most $9dM_g^2 D_X^2$; the factor 9 is absorbed into the universal constant. Assumption 5 and the same conditional ℓ^m_∞ batching inequality give the dual-value contribution $c\kappa_m \sigma_h^2/b$. In contrast, for the primal A-bank the Hilbert-space variance identity gives

$$\mathbb{E}[\|e^x_{t,s}\|_2^2 \mid \mathcal{H}^-_{t,s}] \le \frac{c_x\, dM_R^2}{b} \tag{125}$$

after absorbing the signed-extrapolation and reserve constants, with no factor κ_m. Combining the Euclidean primal part and the scaled dual part proves

$$\mathbb{E}[\|e_{t,s}\|_{A,*}^2 \mid \mathcal{H}^-_{t,s}] \le \frac{\Sigma_{A,\gamma}^2}{b} \tag{126}$$

for the combined perturbation, after enlarging the universal constants in (59).

The fact that the whole bank was physically queried before the inner loop does not affect (122)–(126). The reason is independence: the joint law of a family of independent variables does not depend on the order in which they are generated. This licenses an equivalent *deferred-revelation* construction: we may pretend that each stored variable comes into existence only at the moment of its first use in its slot. With respect to the filtration ordered in this way, the error of each batch is conditionally centered before its slot, and the martingale estimate below applies unchanged. Formally, refine each slot into two half-slots. On the A half-slot use the product state $(p_{t,s}, y_{t,s-1})$ and perturbation $(e^x_{t,s}, 0)$; only the primal coordinate changes. On the subsequent B half-slot use $(p_{t,s}, y_{t,s})$ and perturbation $(0, e^y_{t,s})$; only the dual coordinate changes. Thus the stabilizer-generated primal and dual negative divergences are exactly the squared increments of this refined product-state sequence, up to the one-strong-convexity factor. The A-error is centered relative to $\mathcal{H}^-_{t,s}$ and the B-error relative to $\mathcal{H}^+_{t,s}$. Applying the martingale estimate below to the refined filtration therefore merely doubles the number of increments and changes only a universal constant.

Step 3: stochastic perturbation of the telescoping inequality. If $\Sigma_{A,\gamma} = 0$, equation (126) implies that every bank error is zero almost surely. With $\lambda_N = 0$, equation (120) directly gives the theorem, so no quotient by λ_N occurs. In the remainder of this step assume $\Sigma_{A,\gamma} > 0$, and hence $\lambda_N > 0$. Insert the noisy linear coefficients into the same primal and dual three-point inequalities used for (120). All mean terms reproduce the deterministic ACGD-S expression. The remainder is a sum of weighted martingale perturbations. With k denoting the lexicographic slot index, it has the generic form

$$\sum_k a_k \langle e_k, z_k - z \rangle, \qquad z \in X \times Y_R, \tag{127}$$

plus the negative step-divergence produced by the additional curvature in (121).

We use the following standard auxiliary-prox estimate, stated here in the exact form needed. If e_k is conditionally centered before slot k, $\mathbb{E}[\|e_k\|_{A,*}^2 \mid \mathcal{H}_{k-1}] \le v_k^2$, and V_A is one-strongly convex in the scaled product norm, then for any adapted z_k and any $\lambda > 0$,

$$\mathbb{E} \sup_{z \in X \times Y_R} \left\{ \sum_k a_k \langle e_k, z_k - z \rangle - \frac{\lambda}{4} \sum_k \|z_k - z_{k-1}\|_A^2 \right\} \le C\lambda \mathcal{D}_A^2 + \frac{C}{\lambda} \sum_k a_k^2 v_k^2. \tag{128}$$

For clarity, this does not assume that z_k is independent of e_k. First split $z_k - z = (z_{k-1} - z) + (z_k - z_{k-1})$. Young's inequality gives

$$a_k \langle e_k, z_k - z_{k-1} \rangle \le \frac{\lambda}{8} \|z_k - z_{k-1}\|_A^2 + \frac{2a_k^2}{\lambda} \|e_k\|_{A,*}^2,$$

so this part is absorbed by the displayed negative step term. For the predictable part introduce the auxiliary mirror-descent sequence $w_k^{\mathrm{aux}} = \arg\min_w \{a_k \langle e_k, w \rangle + \lambda V_A(w_{k-1}^{\mathrm{aux}}, w)\}$ and write

$$a_k \langle e_k, z_{k-1} - z \rangle = a_k \langle e_k, w_{k-1}^{\mathrm{aux}} - z \rangle + a_k \langle e_k, z_{k-1} - w_{k-1}^{\mathrm{aux}} \rangle.$$

The second summand has conditional expectation zero because both z_{k-1} and w_{k-1}^{aux} are predictable. For the first summand, split $w_{k-1}^{\mathrm{aux}} - z = (w_{k-1}^{\mathrm{aux}} - w_k^{\mathrm{aux}}) + (w_k^{\mathrm{aux}} - z)$, use Young's inequality on the first difference, and use the prox three-point inequality

$$a_k \langle e_k, w_k^{\mathrm{aux}} - z \rangle \le \lambda \{ V_A(w_{k-1}^{\mathrm{aux}}, z) - V_A(w_k^{\mathrm{aux}}, z) - V_A(w_{k-1}^{\mathrm{aux}}, w_k^{\mathrm{aux}}) \}.$$

One-strong convexity absorbs the Young term into $V_A(w_{k-1}^{\mathrm{aux}}, w_k^{\mathrm{aux}})$; summation leaves at most one initial radius $\lambda \sup_z V_A(w_0^{\mathrm{aux}}, z)$ plus $C \sum_k a_k^2 \|e_k\|_{A,*}^2 / \lambda$. Taking the supremum and expectation therefore proves (128).

Applying (128) with $v_k^2 \le \Sigma_{A,\gamma}^2 / b$ and using the stabilizer-generated negative divergence gives

$$\mathbb{E}\big[W_N \mathcal{C}_{R,\gamma}(\bar{x}_N)\big] \le C_{\mathrm{det}} L_{A,\gamma} \mathcal{D}_A^2 + C\lambda_N \mathcal{D}_A^2 + C \frac{A_N \Sigma_{A,\gamma}^2}{\lambda_N b}, \tag{129}$$

where $A_N = \sum_{t=1}^N \sum_{s=1}^{S_t} a_{t,s}^2$. The choice of λ_N in Algorithm 2 balances the last two terms and yields

$$\mathbb{E}\mathcal{C}_{R,\gamma}(\bar{x}_N) \le C \frac{L_{A,\gamma} \mathcal{D}_A^2}{W_N} + C \frac{\Sigma_{A,\gamma} \mathcal{D}_A \sqrt{A_N / b}}{W_N}. \tag{130}$$

Step 4: weight arithmetic. Set $c = \bar{M}_\gamma r_{\mathrm{sl}} / L_{A,\gamma}$, so $S_t = \max\{1, \lceil ct \rceil\}$. The ceiling gives

$$K_N = \sum_{t=1}^N S_t \le 2N + cN(N+1), \tag{131}$$

which is (61) up to a universal constant. Furthermore,

$$A_N = \sum_{t=1}^{N} \frac{t^2}{S_t} \le C \frac{W_N^2}{K_N}. \tag{132}$$

One elementary verification is to split into two regimes. If $cN \le 1$, then $K_N = \Theta(N)$, $A_N \le \sum_{t \le N} t^2 = O(N^3)$, and $W_N^2/K_N = \Theta(N^3)$. If $cN > 1$, split the sum at $t_0 = \lceil 1/c \rceil$; then $A_N = O(c^{-3}) + O(\sum_{t>t_0} t/c) = O(N^2/c)$, while $K_N = \Theta(cN^2)$ and again $W_N^2/K_N = \Theta(N^2/c)$.

Substituting (132) and $W_N = \Theta(N^2)$ into (130) proves

$$\mathbb{E}\mathcal{C}_{R,\gamma}(\bar{x}_N) \le C_0 \frac{L_{A,\gamma}\mathcal{D}_A^2}{N(N+1)} + C_1 \frac{\Sigma_{A,\gamma}\mathcal{D}_A}{\sqrt{bK_N}},$$

which is (60).

Step 5: query count and adaptive depth. Each A-bank entry uses a constant number of full vector values for the common-randomness pair; each independent B-entry uses a second pair and one constraint-value vector. The first-slot previous-center reserve adds only a constant-factor overhead. Therefore $Q_{\mathrm{vec}} \le c_Q b K_N$. All points queried in phase t are functions only of the already known outer center(s), sampled directions, and fresh seeds; none depends on an inner iterate because the stored full Jacobian rows are combined with the adaptive multiplier only after the query. Thus the entire bank is issued in a constant number of parallel stages per phase and the strict adaptive depth is $O(N)$.

Finally, impose the target floor (62), choose N so that the deterministic term in (60) is at most $\varepsilon_s/2$, and choose b as in (63). Since $bK_N \le K_N + C\Sigma_{A,\gamma}^2 \mathcal{D}_A^2/\varepsilon_s^2$, and (131) with $L_{A,\gamma}\mathcal{D}_A^2/\varepsilon_s \ge 1$ gives $N^2 = O(L_{A,\gamma}\mathcal{D}_A^2/\varepsilon_s)$ and hence

$$K_N = O\left(N + \frac{\bar{M}_\gamma r_{\mathrm{sl}} \mathcal{D}_A^2}{\varepsilon_s}\right),$$

we obtain (64)–(65). □

A.11 Proof of Corollaries 3 and 4

Proof. In the smooth regime, Lemma 1 permits the natural majorant $L_0 + RL_g$, and the smoothing radii can be chosen so that the objective and shifted-constraint biases are each at most a fixed fraction of ε. Use $L_{A,\gamma} = \max\{L_0 + RL_g, \varepsilon_s/\mathcal{D}_A^2\}$ as prescribed by (62). The floor contributes only the explicit constant in the depth bound. Apply Theorem 3 with $\varepsilon_s = \Theta(\varepsilon)$ and use

$$\Sigma_{A,\gamma}^2 = O\left(dM_R^2 + \kappa_m r_{\mathrm{sl}}^2 (dD_X^2 M_g^2 + \sigma_h^2)\right).$$

The certificate-to-primal inequalities (20) control the smoothed objective and shifted-constraint violation. The smoothing inequalities of Lemma 1 then transfer both quantities to the original problem. This proves (67).

In the Lipschitz nonsmooth regime, choose the largest radius allowed by (45), namely $\gamma = \gamma_\varepsilon$ from (68). When $M_A > 0$, Lemma 1 gives the natural majorant $O(\sqrt{d}\, M_R/\gamma_\varepsilon)$ up to fixed problem scales; when $M_A = 0$, that natural majorant vanishes. Impose the target floor (62) in either case. Its contribution is absorbed by the explicit leading constant below. Thus

$$N_{\mathrm{seq}} = O\left(1 + \sqrt{\frac{L_{A,\gamma}\mathcal{D}_A^2}{\varepsilon}}\right) = O\left(1 + d^{1/4}\sqrt{\frac{M_R \mathcal{D}_A^2}{\gamma_\varepsilon \varepsilon}}\right)$$

up to those scales. If $\bar{\gamma}$ does not bind, then $\gamma_\varepsilon = \Theta(\varepsilon)$ up to the displayed Lipschitz scale and this specializes to $O(d^{1/4}/\varepsilon)$. If it binds, the depth is instead $O(d^{1/4}/\sqrt{\bar{\gamma}\varepsilon})$ up to problem scales. Substitution of (59) into (65) gives the remaining terms displayed in (69); in particular, the dual Jacobian-action and value contributions carry κ_m, while the primal Euclidean contribution does not. Shifted smoothing again transfers the result to the original constraints. □

A.12 Proof of Theorem 4

Proof. Fix $z \in Y$. Since $d_\mu(z)$ is the minimum of $\mathcal{L}_\mu(\cdot, z)$,

$$\begin{aligned} d_\mu(z) &\le \mathcal{L}_\mu(\widetilde{x}(y), z) \\ &= \mathcal{L}_\mu(\widetilde{x}(y), y) + \langle g(\widetilde{x}(y)), z - y\rangle \\ &\le d_\mu(y) + \delta_{\text{in}} + \langle g(\widetilde{x}(y)), z - y\rangle . \end{aligned}$$

Multiplying by -1 gives

$$\varphi_\mu(z) \ge \varphi_\mu(y) + \langle -g(\widetilde{x}(y)), z - y\rangle - \delta_{\text{in}},$$

which is (75).

If $\widehat{g}$ is used, then

$$\begin{aligned} \langle -g(\widetilde{x}(y)), z - y\rangle &= \langle -\widehat{g}, z - y\rangle + \langle \widehat{g} - g(\widetilde{x}(y)), z - y\rangle \\ &\ge \langle -\widehat{g}, z - y\rangle - \tau D_Y. \end{aligned}$$

Substitution proves (76). □

A.13 Justification of Theorem 5

The deterministic δ-subgradient estimate in Theorem 1 of [Gladin et al., 2022] states, in dimension m, that after N cuts the optimization error of the best queried point is the sum of an exponentially decaying localization term and the uniform additive subgradient error. With fixed volumetric-center parameters, the localization term is at most ε_D once $N = \mathcal{O}\big(m \log\big(e + m^{3/2} R_Y B_\varphi/(\rho_Y \varepsilon_D)\big)\big)$. Substituting the error from Theorem 4 therefore adds $\delta_{\text{in}} + D_Y\tau$, not N times this quantity.

The cited result uses the best queried point in true objective value and contains no noisy-incumbent claim. Let $y^{\text{best}} \in \arg\min_{j \le N} \varphi_\mu(y_j)$ and let $\widehat{y} \in \arg\min_{j \le N} \widehat{\varphi}(y_j)$. On the event $\max_{j \le N} |\widehat{\varphi}(y_j) - \varphi_\mu(y_j)| \le \eta_{\text{val}}$,

$$\varphi_\mu(\widehat{y}) \le \widehat{\varphi}(\widehat{y}) + \eta_{\text{val}} \le \widehat{\varphi}(y^{\text{best}}) + \eta_{\text{val}} \le \varphi_\mu(y^{\text{best}}) + 2\eta_{\text{val}},$$

which proves (79). Thus the $2\eta_{\text{val}}$ loss is a separate elementary selection bound, not part of Theorem 1 of [Gladin et al., 2022].

Equations (82)–(84), under Assumption 7, give a concrete realization of the cut and incumbent events in Theorem 5; the deterministic inner bias is at most δ_{in} and the fresh validation batch controls the stochastic remainder. If all cut and incumbent events hold, the deterministic theorem plus the preceding selection bound gives the claimed $\mathcal{O}(\varepsilon_D)$ dual error. Assigning conditional failure probability at most β/N_V to each outer query and applying a union bound yields total failure probability at most β. This is why the inner target is $\Theta(\varepsilon_D)$ rather than $\Theta(\varepsilon_D/N_V)$.

A.14 Proof of the regularization bias

Proof. For every $x \in X$,

$$0 \le \frac{\mu}{2}\|x - x_c\|_2^2 \le \frac{\mu}{2} D_X^2$$

when $x_c \in X$. Evaluating the regularized problem at an optimizer of the unregularized problem and comparing optimal values proves that the regularized objective optimum differs by at most $\mu D_X^2/2$. □

A.15 Dual smoothness and proof of Corollary 5

Proof. Let $x(y)$ and $x(z)$ minimize the two μ-strongly convex inner objectives. Strong convexity gives

$$\mathcal{L}_\mu(x(z), y) \ge \mathcal{L}_\mu(x(y), y) + \frac{\mu}{2}\|x(z) - x(y)\|^2,$$
$$\mathcal{L}_\mu(x(y), z) \ge \mathcal{L}_\mu(x(z), z) + \frac{\mu}{2}\|x(z) - x(y)\|^2.$$

Adding these inequalities yields

$$\mu\|x(z) - x(y)\|^2 \le \langle y - z, g(x(z)) - g(x(y))\rangle.$$

If $\|g(x) - g(z)\|_* \le \overline{M}_g\|x - z\|$ in paired norms, then

$$\|x(y) - x(z)\| \le \frac{\overline{M}_g}{\mu}\|y - z\|, \qquad \|g(x(y)) - g(x(z))\|_* \le \frac{\overline{M}_g^2}{\mu}\|y - z\|.$$

Uniqueness and Danskin's theorem give $\nabla\varphi_\mu(y) = -g(x(y))$, so $L_D \le \overline{M}_g^2/\mu$. For $g(x) = Ax - b$ in Euclidean norms, $\overline{M}_g = \|A\|$ and hence $L_D \le \|A\|^2/\mu$.

For a convex function whose gradient is L_D-Lipschitz, choose the strictly positive majorant $\widehat{L}_D \ge L_D$ from (88). The projected-gradient step $y^+ = \Pi_Y(y - \nabla\varphi_\mu(y)/\widehat{L}_D)$ satisfies

$$\varphi_\mu(y^+) \le \varphi_\mu(y) - \frac{1}{2\widehat{L}_D}\left\|\mathcal{G}_{\widehat{L}_D}(y)\right\|^2.$$

Since $\varphi_\mu(y^+) \ge \varphi_\mu^\star$, inequality (89) follows. Combining it with (90) gives

$$\mathcal{E}_{\text{primal}}(x_\mu(y)) \le C_{\text{rec}}\sqrt{2\widehat{L}_D(\varphi_\mu(y) - \varphi_\mu^\star)}.$$

When $C_{\text{rec}} > 0$, the right-hand side is at most ε under (91). When $C_{\text{rec}} = 0$ and the primal error is nonnegative, (90) already gives zero error. The positive-majorant convention makes the same proof valid when the least gradient-Lipschitz constant is $L_D = 0$. The statement under the linear recovery inequality follows directly. □

A.16 Affine sliding correspondence

A.17 Proof of Theorem 6 and Corollaries 6 and 7

Proof. We spell out the degenerate bilinear specialization instead of using only a notation map. Fix $0 < \gamma \le \bar{\gamma}$ and define $\Phi_\gamma^{\text{aff}}(x, y) := f_{0,\gamma}(x) + \langle y, Ax - b\rangle$ and the restricted weak gap

$$\text{Gap}_\gamma^{\text{aff}}(\bar{z}) := \sup_{y \in Y_R} \Phi_\gamma^{\text{aff}}(\bar{x}, y) - \inf_{x \in X} \Phi_\gamma^{\text{aff}}(x, \bar{y}).$$

Table 4. Correspondence between the affine specialization and the notation of Nguyen–Gasnikov.

Nguyen–Gasnikov notation or assumption	Present specialization
bounded product domain, radius Ω	$X \times Y_R$, $\Omega \leftrightarrow D_A^2$
L_x and stochastic variance σ_x^2	L_0 and $\sigma_{\mathrm{zo}}^2 = \mathcal{O}(dM_0^2)$
L_{xy}	$\|A\|$
$g(y)$ block	$\langle b, y\rangle$, linear and exact ($L_y = \sigma_y = 0$)
exact bilinear products	exact A, $A^\top$, and b

For (93), the corresponding smoothed monotone saddle operator is

$$G(z) = G_f(z)+G_A(z)+G_b, \quad G_f(x,y) = (\nabla f_{0,\gamma}(x), 0), \quad G_A(x,y) = (A^\top y, -Ax), \quad G_b = (0,b).$$

The first block is the only stochastic block and has smoothness $L_{0,\gamma}$ and conditional variance σ_f^2; the second block is exact, skew-symmetric, and $\|A\|$-Lipschitz; the last block is exact and constant. The common-randomness estimator used for the first block satisfies

$$\mathbb{E}_{u,\xi}\left[\frac{d}{2\gamma}\big(F_0(x+\gamma u,\xi) - F_0(x-\gamma u,\xi)\big)u\right] = \nabla f_{0,\gamma}(x)$$

by Lemma 2; in general it is not unbiased for $\nabla f_0(x)$. The three-point prox inequality for the recursive sliding updates, summed over the outer objective steps and the inner bilinear steps, gives the expected smoothed weak-gap estimate

$$\mathbb{E}\,\mathrm{Gap}_\gamma^{\mathrm{aff}}(\bar z) \le C\left[\frac{L_{0,\gamma}D_A^2}{Q_f^2} + \sqrt{\frac{\sigma_f^2 D_A^2}{Q_f}} + \frac{\|A\|\,D_A^2}{Q_A}\right]. \tag{133}$$

For completeness, the first term is the usual accelerated telescope of the $L_{0,\gamma}$-smooth block. The bilinear cross terms cancel because $\langle G_A(z) - G_A(z'), z - z'\rangle = 0$; the remaining inner boundary terms telescope and leave $C\,\|A\|\,D_A^2/Q_A$. If e_k denotes the centered stochastic error of the objective block, the only random remainder is a predictable martingale sum. Introducing the standard auxiliary mirror sequence and applying its three-point inequality yields $\mathbb{E}\sup_z \sum_{k=1}^{Q_f}\langle e_k, z_k - z\rangle \le C\sqrt{Q_f\sigma_f^2 D_A^2}$; division by the accumulated objective weight gives the middle term of (133). Thus no stochastic term is charged to A, $A^\top$, or b. Choosing each of the three terms in (133) below a fixed fraction of the internal smoothed-gap target gives

$$Q_f = \mathcal{O}\left(\sqrt{\frac{L_{0,\gamma}D_A^2}{\varepsilon}} + \frac{\sigma_f^2 D_A^2}{\varepsilon^2}\right), \qquad Q_A = \mathcal{O}\left(\frac{\|A\|\,D_A^2}{\varepsilon}\right). \tag{134}$$

This proves Theorem 6 after identifying σ_f^2 with the objective estimator variance σ_{zo}^2 and absorbing the constant number of value calls per two-point sample. This is precisely the degenerate stochastic bilinear specialization of Corollary K.9 and Section K.2 of Nguyen and Gasnikov [Nguyen, Gasnikov, 2026]; the argument above records the specialization needed here without requiring the reader to reconstruct it from the source notation. Under common-randomness two-point feedback, the objective-only specialization of Lemma 2 gives $\sigma_f^2 = \mathcal{O}(dM_0^2)$. In the smooth case, Lemma 1 gives $L_{0,\gamma} \le L_0$ and

$$0 \le f_{0,\gamma}(x) - f_0(x) \le \frac{L_0\gamma^2}{2} \quad \text{uniformly on } X.$$

Consequently the original weak gap is at most the smoothed weak gap plus $L_0\gamma^2/2$. Running the smoothed method with target $\varepsilon/2$ and choosing (95) makes the transfer loss at most $\varepsilon/12$. Substitution in (134) proves (96). If $L_0 = 0$, then ∇f_0 is constant on the convex query neighborhood, hence f_0 is affine and ball smoothing leaves it unchanged. Any $0 < \gamma \le \bar{\gamma}$ is then admissible, $L_{0,\gamma} = 0$, and the same argument applies with no curvature or bias term.

For a nonsmooth objective, choose $\gamma = \gamma_{f,\varepsilon}$ from (97). Lemma 1 gives $L_{0,\gamma} = \mathcal{O}(\sqrt{d}M_0/\gamma_{f,\varepsilon})$. Therefore

$$\sqrt{\frac{L_{0,\gamma}D_A^2}{\varepsilon}} = \mathcal{O}\left(d^{1/4}\sqrt{\frac{M_0 D_A^2}{\gamma_{f,\varepsilon}\varepsilon}}\right).$$

The variance term remains $\mathcal{O}(dM_0^2D_A^2/\varepsilon^2)$, and the exact coupling count is unchanged. The smoothing bias is absorbed by a constant-factor split of the target accuracy. If $\varepsilon \le 3M_0\bar{\gamma}$, substitution of $\gamma_{f,\varepsilon} = \varepsilon/(3M_0)$ gives (99); otherwise the general expression is (98). The case $M_0 = 0$ is degenerate and has no objective-oracle contribution. □

A.18 Proofs of Propositions 1 and 2

Proof. For Proposition 1, set every constraint equal to the known constant $-\rho$. Feasibility is then automatic, the Slater condition holds everywhere, and the output requirement in Definition 1 contains the unconstrained objective requirement. Any constrained algorithm therefore induces an unconstrained paired-value algorithm with the same number of objective observations. Proposition 1 of [Duchi et al., 2015], specialized to the Euclidean ball and its second-moment oracle class, gives expected error at least a universal multiple of $M_0D_X\sqrt{d/T}$ in the nontrivial range. Its hard losses are linear. Solving this inequality for T proves (103).

For Proposition 2, Lipschitz continuity and $\phi(x^{\mathrm{ref}}) = 0$ imply $|\phi(x)| \le M_\phi D_X$ on X, so the epigraph optimum $(x^\star, \phi^\star)$ belongs to Z. The stated reference point is strictly feasible because

$$g(x^{\mathrm{ref}}, M_\phi D_X + \rho) = -(M_\phi D_X + \rho) \le -\rho.$$

For every realization,

$$\phi(\widehat{x}) - \phi^\star = \phi(\widehat{x}) - \widehat{t} + \widehat{t} - \phi^\star \le [\phi(\widehat{x}) - \widehat{t}]_+ + (\widehat{t} - \phi^\star).$$

Taking expectations and applying the two accuracy inequalities gives (105). A value of the unknown constraint at (x, t) is exactly a value of $\phi(x)$ followed by subtraction of the known scalar t, so an algorithm for (104) would solve the original hard two-point family with the same order of observations. □

A.19 Bounded pathwise hard-family transfer (proof sketch)

This paragraph records the missing step needed to transfer the Gaussian linear family to bounded sampled slopes. It is not used as a proved same-class minimax result. Fix $d \ge 2$ and the nontrivial accuracy range

$$0 < \varepsilon \le c_0 M_0 D_X \tag{135}$$

for a sufficiently small universal c_0. In the usual normalization, take fresh Gaussian coefficients $X_v \sim N(\mu_v, s^2 I_d)$ with $s \le M_0/\sqrt{d}$, equal mean norms $\|\mu_v\|_2 \le M_0/4$, and let P_v denote the law of X_v. Condition on $E = \{\|X_v\|_2 \le 4M_0\}$. Gaussian concentration gives the common probability $p = P_v(E) \ge 1/2$: the event is radial and every hypothesis has the same mean norm. The conditioned linear losses have slopes bounded by $4M_0$, and a rescaling by four restores pathwise modulus M_0. Radial symmetry gives $\mathbb{E}[X_v \mid E] = \kappa\mu_v$ with the same κ for every hypothesis. In the displayed mean range, Gaussian tail bounds keep κ between

two positive universal constants. The objective separation and the admissible range (135) therefore change only by universal factors.

At the coefficient-law level, if $P_v^E = P_v(\cdot \mid E)$, equal conditioning probabilities give

$$\chi^2(P_v^E \| P_w^E) \le p^{-1}\chi^2(P_v \| P_w) \le 2\chi^2(P_v \| P_w). \tag{136}$$

This one-shot comparison is not enough for an adaptive paired-value algorithm. Condition on the algorithm's internal randomness and let $K_{v,t}^E(\cdot \mid H_{t-1})$ denote the reply kernel at the query pair selected from the previous history. The required uniform step is

$$\mathrm{KL}\big(K_{v,t}^E(\cdot \mid h) \big\| K_{w,t}^E(\cdot \mid h)\big) \le C_{\mathrm{tr}}\,\mathrm{KL}(K_{v,t}(\cdot \mid h) \| K_{w,t}(\cdot \mid h)) \quad \text{for every reachable } h, \tag{137}$$

with a universal C_{tr}. If (137) holds, the KL chain rule for the complete transcript gives the same testing bound up to C_{tr}. Inequality (136) and data processing motivate this transfer, but the latent radial conditioning event still requires a uniform likelihood-ratio calculation for every paired marginal. That calculation is not supplied here. The argument is therefore a proof sketch, not a theorem for the pathwise-Lipschitz class.

A.20 Proof of Proposition 3

Proof. Let $X = [-1, 1]$, $f_0(x) = -x$, and consider two one-constraint instances, indexed by $b \in \{-8\varepsilon, +8\varepsilon\}$, with $g_b(x) = x - b$ and $\varepsilon \le 1/16$. The point $x^{\mathrm{s}} = -1$ is Slater with margin at least $1/2$ for both instances, and their optima are $x_b^\star = b$. Suppose the constraint oracle returns $g_b(x) + Z$, where $Z \sim N(0, \sigma_h^2)$, while the objective is exact. For any output $\widehat{x}$, the two accuracy inequalities give

$$\mathbb{E}_b(b - \widehat{x}) \le \varepsilon, \qquad \mathbb{E}_b[\widehat{x} - b]_+ \le \varepsilon.$$

Hence $\mathbb{E}_b|\widehat{x} - b| \le 3\varepsilon$. The test that decides the sign of b from the sign of $\widehat{x}$ has error at most $3/8$ under either hypothesis by Markov's inequality. At every adaptively selected query point, the two Gaussian observation laws have means separated by 16ε. The KL chain rule therefore gives transcript divergence at most $128 Q_{\mathrm{vec}} \varepsilon^2/\sigma_h^2$. Le Cam's two-point inequality requires a positive universal lower bound on this divergence for testing error below $1/2$, which proves (106). □

A.21 Verification of the synthetic test problem

The random affine constraints $2, \dots, m-1$ depend only on coordinates $2{:}d$ and have right-hand side 0.3; the quadratic constraint also depends only on these coordinates. At $x^\star = 0.2e_1$, all of them are strictly inactive, while $g_1(x^\star) = 0$. Both objectives in (107) are radially increasing in the distance from $a = 0.75e_1$. Therefore the closest feasible point to a is $0.2e_1$, proving (108). The KKT multiplier of g_1 equals 0.55 for the smooth objective and 1 for the nonsmooth objective. At $x^{\mathrm{s}} = 0$, the affine slacks are 0.2 and 0.3, and the quadratic slack is 0.05, so $\rho = 0.05$.

The experimental code uses the same estimator structure as Algorithm 1. For each operator sample it evaluates two vectors at $x \pm \gamma u$ and one vector at $x + \gamma v$. Two independent operator batches are used in each Mirror–Prox iteration. Thus the exact number of full vector values through iteration k is $6rk$, and direct componentwise simulation produces $(m+1)6rk$ scalar component values. The source archive contains the deterministic instance arrays, complete configuration, thirty main evaluation seeds, the thirty-seed radius ablation, raw per-seed trajectories, equal-budget summaries, and the aggregation script.

A.22 Independent additive value noise

Corollary 9. *Suppose that every scalar component observation is corrupted by an additive measurement noise ζ_i that is conditionally centered and satisfies $\mathbb{E}[\zeta_i^2 \mid \mathcal{F}] \le \sigma_{\mathrm{val}}^2$ uniformly, with independent noises at the two perturbed points. Define*

$$\sigma_{\mathrm{val,R}}^2 := \sup_{y \in Y_R} \mathbb{E}\left[\left(\zeta_0 + \sum_{i=1}^{m} y_i \zeta_i\right)^2 \middle| \mathcal{F}\right]. \tag{138}$$

Without assuming independence across coordinates at one point,

$$\sigma_{\mathrm{val,R}}^2 \le \sigma_{\mathrm{val}}^2 (1+R)^2. \tag{139}$$

If the coordinate noises are conditionally independent, the sharper bound $\sigma_{\mathrm{val,R}}^2 \le \sigma_{\mathrm{val}}^2(1+R^2)$ holds. Independent noises at the two perturbed points therefore add $d^2\sigma_{\mathrm{val,R}}^2/(2\gamma^2)$ to the Euclidean primal part of Σ_γ^2.

The absolute constraint-value sample in (25) is corrupted as well. Write

$$\sigma_{h,\zeta}^2 := \sup \mathbb{E}[\|\zeta_{1:m}\|_\infty^2 \mid \mathcal{F}]. \tag{140}$$

Finite coordinatewise variances give $\sigma_{h,\zeta}^2 \le m\sigma_{\mathrm{val}}^2$. By Lemma 3, the generic batch proxy receives the additional contribution

$$C\left(\frac{d^2\sigma_{\mathrm{val,R}}^2}{\gamma^2} + \kappa_m \sigma_{h,\zeta}^2\right). \tag{141}$$

Under conditional coordinatewise sub-Gaussian noise, applying the maximal inequality directly to the averaged coordinates replaces the second term by $C\sigma_{\mathrm{val}}^2 \log(2m)$. In the Lipschitz nonsmooth regime with (45), the additional total-query term is

$$\mathcal{O}\left(\frac{d^2\sigma_{\mathrm{val,R}}^2 D_\omega^2}{\varepsilon^2} \max\left\{\frac{1}{\bar{\gamma}^2}, \frac{9M_{\max}^2}{\varepsilon^2}\right\} + \frac{\kappa_m \sigma_{h,\zeta}^2 D_\omega^2}{\varepsilon^2}\right). \tag{142}$$

When $M_{\max} > 0$ and $\varepsilon \le 3M_{\max}\bar{\gamma}$, the first summand is $\mathcal{O}(d^2\sigma_{\mathrm{val,R}}^2 M_{\max}^2 D_\omega^2/\varepsilon^4)$. This exponent is an upper-bound penalty of the central-difference construction, not a stated minimax law.

A.23 Central-difference calculation

Proof. Let the two additive noises be independent, centered, and have variance at most σ_{val}^2. Their difference has variance at most $2\sigma_{\mathrm{val}}^2$. The noise part of the estimator is

$$\frac{d}{2\gamma}(\zeta^+ - \zeta^-)u.$$

Since $\|u\|_2 = 1$,

$$\mathbb{E}\left\|\frac{d}{2\gamma}(\zeta^+ - \zeta^-)u\right\|_2^2 \le \frac{d^2}{4\gamma^2}\, 2\sigma_{\mathrm{val}}^2 = \frac{d^2\sigma_{\mathrm{val}}^2}{2\gamma^2}.$$

This term is absent when the stochastic sample enters through a pathwise Lipschitz function and the same sample is used at both perturbed points, as in Definition 2. In the ablation of Section 9, the empirical second moment is computed directly at the smooth saddle point; no optimization trajectory is used to infer the γ^{-2} law. □

A.24 Exact scalar accounting

If a full vector value at a point is simulated by evaluating every component $F_i(x, \xi)$, then each vector call costs exactly $m + 1$ scalar component calls, which proves (7). A common sample identifier reproduces the vector oracle's joint law; without it, the count remains valid but the simulated joint noise law may differ. The identity does not apply to a sampled-constraint oracle, because that oracle does not reconstruct the full vector and its convergence proof uses different information and assumptions.

B Open questions and scope boundaries

The sharpness statements in the main text cover only parameters for which a matching lower bound is available. The following questions remain open.

1. The upper bounds contain $M_R^2 = (M_0 + RM_g)^2$, with $R = 1 + M_0 D_X/\rho$. Proposition 2 already gives d/ε^2 hardness with a fixed Slater margin. Multiplier norms can diverge as the margin shrinks, but the necessity of the full ρ^{-2} factor is not established.
2. Proposition 3 closes the unweighted σ_h^2/ε^2 lower bound. A matching lower bound for its R-weighted accelerated counterpart is unknown.
3. The accelerated bank variance contains $\kappa_m d r_{\mathrm{sl}}^2 D_X^2 M_g^2/\varepsilon^2$ from the ℓ_∞ Jacobian linearization. A phase-bank control variate or an additional predictable directional query may reduce this term; the present theorem does not claim it is intrinsic.
4. For a scalar component oracle, an additive law of order $\widetilde{\Theta}((d+m)/\varepsilon^2)$ is a natural conjecture up to problem scales. Naive uniform constraint sampling multiplies the dual variance by m; whether adaptive importance sampling can remove this loss remains open.
5. The adaptive-depth bounds for the paired vector-value oracle have no matching parallel lower bound. Existing nonsmooth lower bounds use different local-oracle semantics and therefore do not settle the present model.
6. Under independent additive measurement noise, the ε^{-4} term in Corollary 9 is an upper bound for central differences. No matching lower bound is asserted for the constrained vector model.

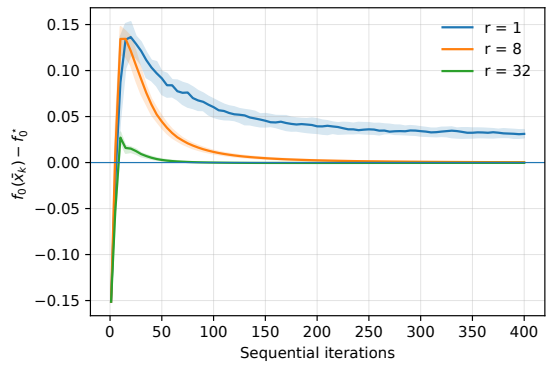
C Additional numerical diagnostics

This section contains the secondary diagnostics omitted from the main narrative. They use the same test problems, oracle accounting, and joint metric J defined in Section 9.

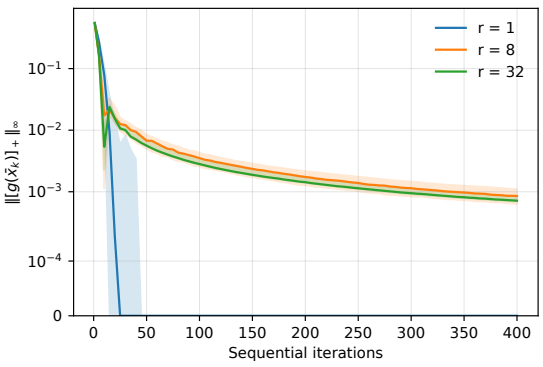
C.1 Fixed-horizon Mirror–Prox runs

C.2 High-noise stress test

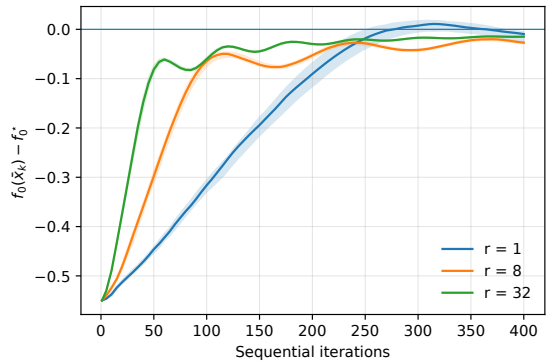
We repeat both regimes with $\sigma_0 = 0.20$, $\sigma_i = 0.08$, sixteen seeds, and 500 iterations. At the common iteration horizon, the smooth median J decreases from $2.67 \cdot 10^{-2}$ for $r = 1$ to $1.25 \cdot 10^{-3}$ for $r = 8$ and $6.60 \cdot 10^{-4}$ for $r = 32$. In the nonsmooth problem, the corresponding values are $6.05 \cdot 10^{-2}$, $1.84 \cdot 10^{-2}$, and $1.11 \cdot 10^{-2}$. The call-based panels charge the proportional batch cost.



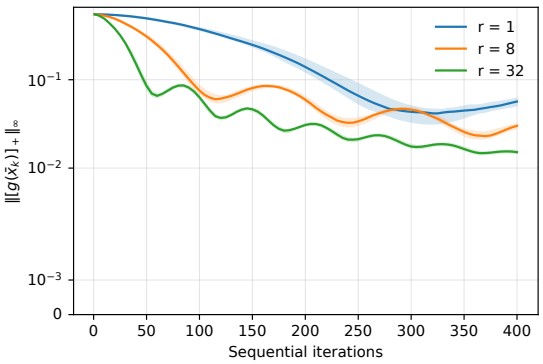
(a) Smooth objective residual.



(b) Smooth maximum violation.



(c) Nonsmooth objective residual.

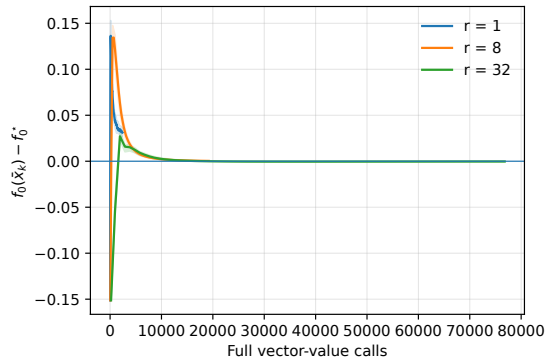


(d) Nonsmooth maximum violation.

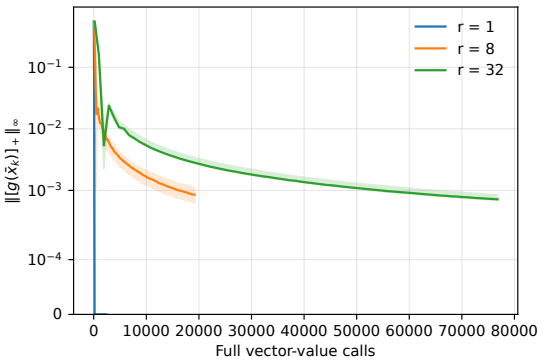
Figure 4. Convergence versus sequential Mirror–Prox iterations. Curves show medians over thirty seeds; shaded bands are interquartile ranges.

Table 5. Final median metrics after 400 iterations; brackets contain the interquartile range. The last column is the nonnegative joint metric J.

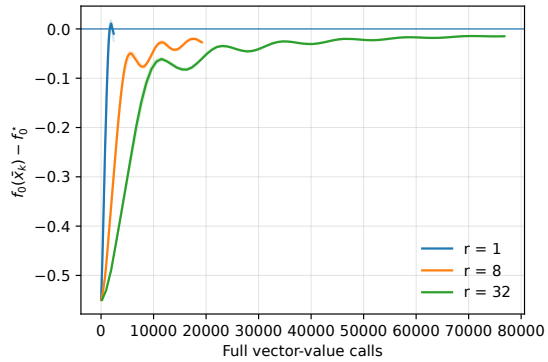
Regime	Batch	Objective residual	Maximum violation	Joint metric
Smooth	1	$3.12 \cdot 10^{-2}$ $[2.69, 3.58] \cdot 10^{-2}$	0 $[0, 0]$	$3.12 \cdot 10^{-2}$ $[2.69, 3.58] \cdot 10^{-2}$
Smooth	8	$3.54 \cdot 10^{-4}$ $[2.12, 5.61] \cdot 10^{-4}$	$8.53 \cdot 10^{-4}$ $[6.25, 11.01] \cdot 10^{-4}$	$1.01 \cdot 10^{-3}$ $[0.72, 1.16] \cdot 10^{-3}$
Smooth	32	$-3.00 \cdot 10^{-4}$ $[-3.72, -2.69] \cdot 10^{-4}$	$7.16 \cdot 10^{-4}$ $[6.92, 8.50] \cdot 10^{-4}$	$7.16 \cdot 10^{-4}$ $[6.92, 8.50] \cdot 10^{-4}$
Nonsmooth	1	$-9.61 \cdot 10^{-3}$ $[-2.53, -0.52] \cdot 10^{-2}$	$5.67 \cdot 10^{-2}$ $[5.10, 6.16] \cdot 10^{-2}$	$5.67 \cdot 10^{-2}$ $[5.10, 6.16] \cdot 10^{-2}$
Nonsmooth	8	$-2.71 \cdot 10^{-2}$ $[-2.86, -2.54] \cdot 10^{-2}$	$3.01 \cdot 10^{-2}$ $[2.82, 3.14] \cdot 10^{-2}$	$3.01 \cdot 10^{-2}$ $[2.82, 3.14] \cdot 10^{-2}$
Nonsmooth	32	$-1.48 \cdot 10^{-2}$ $[-1.54, -1.44] \cdot 10^{-2}$	$1.51 \cdot 10^{-2}$ $[1.46, 1.57] \cdot 10^{-2}$	$1.51 \cdot 10^{-2}$ $[1.46, 1.57] \cdot 10^{-2}$



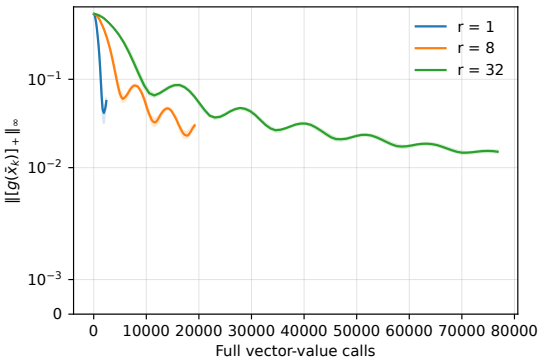
(a) Smooth objective residual.



(b) Smooth maximum violation.

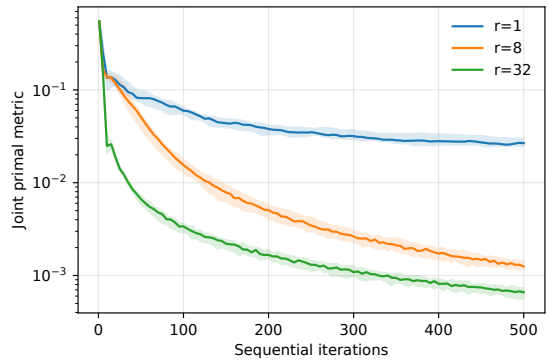


(c) Nonsmooth objective residual.

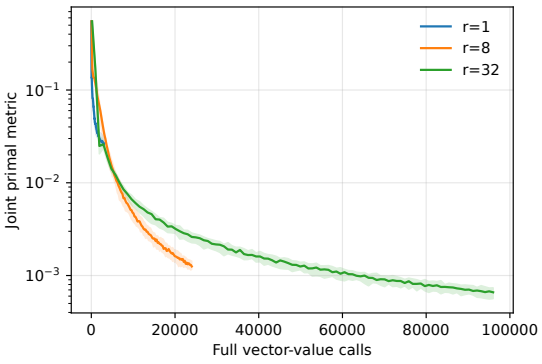


(d) Nonsmooth maximum violation.

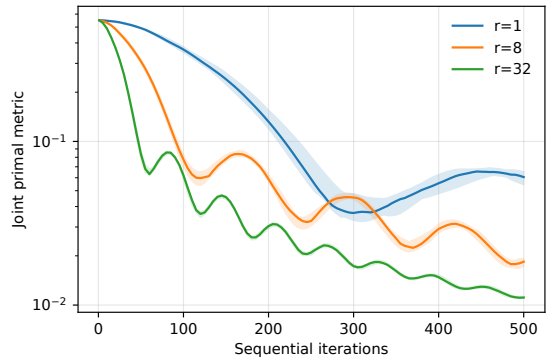
Figure 5. The same trajectories versus full vector-value calls. Larger batches trade adaptive depth for proportionally more total oracle work.



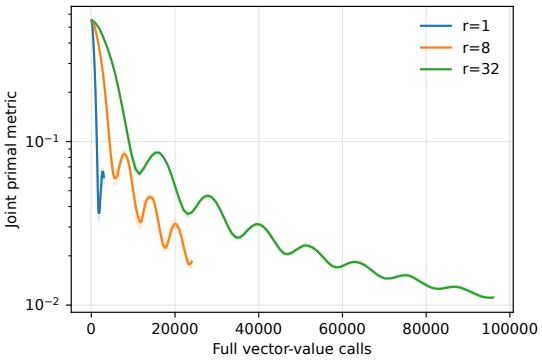
(a) Smooth: iterations.



(b) Smooth: vector calls.

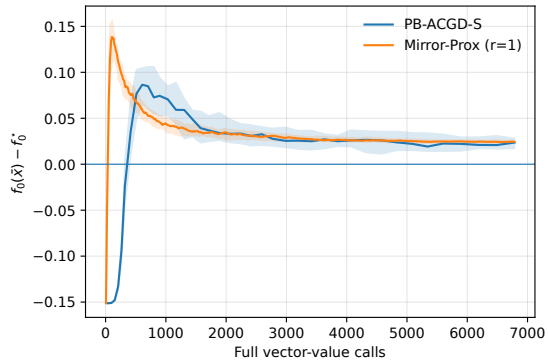


(c) Nonsmooth: iterations.

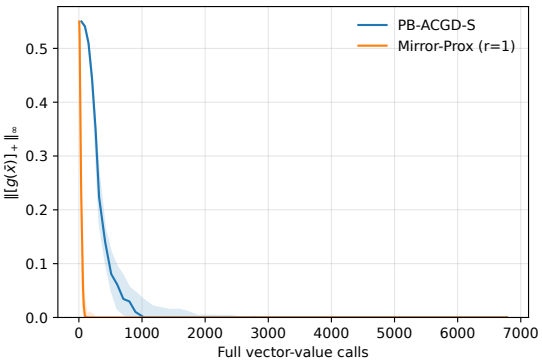


(d) Nonsmooth: vector calls.

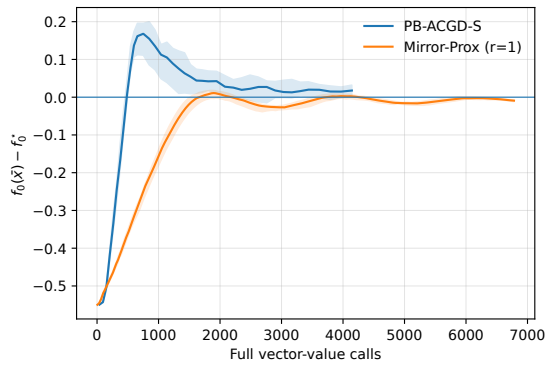
Figure 6. High-noise stress test using the common nonnegative metric J.



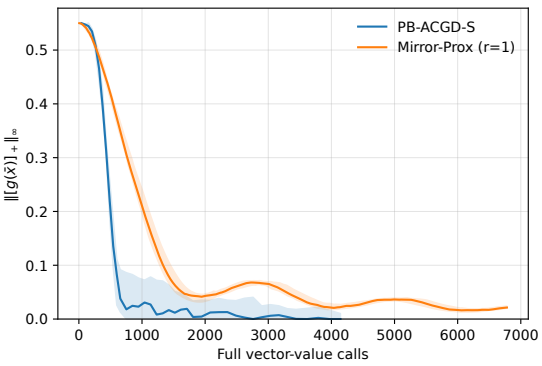
(a) Smooth objective residual.



(b) Smooth maximum violation.



(c) Nonsmooth objective residual.



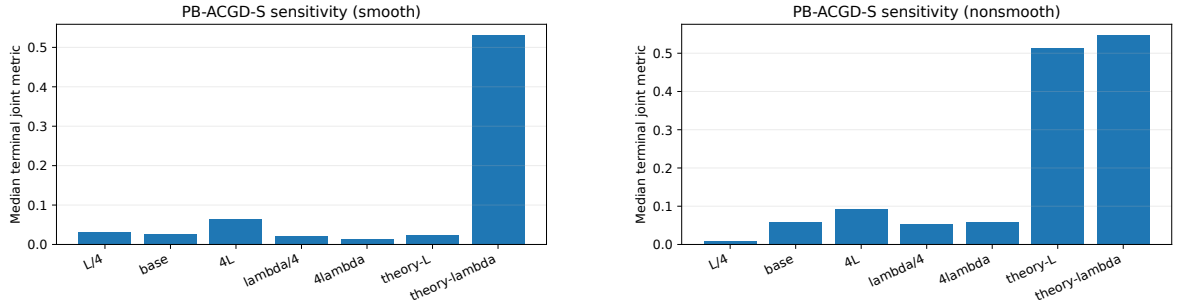
(d) Nonsmooth maximum violation.

Figure 7. PB–ACGD-S and Mirror–Prox charged by the actual number of full vector-value calls, including the A/B banks and cross-phase reserve.

C.3 PB–ACGD-S call accounting

C.4 PB–ACGD-S parameter sensitivity

The sensitivity script perturbs $L_{A,\gamma}$ and λ_N by factors of four using ten seeds and thirty phases. The smooth base configuration has median terminal $J = 2.59 \cdot 10^{-2}$; changing λ_N to $4\lambda_N$ and $\lambda_N/4$ gives $1.43 \cdot 10^{-2}$ and $2.22 \cdot 10^{-2}$. The nonsmooth base gives $5.86 \cdot 10^{-2}$, and the same perturbations give $5.87 \cdot 10^{-2}$ and $5.23 \cdot 10^{-2}$. Formula-based worst-case stabilizers are much more conservative, as shown below.



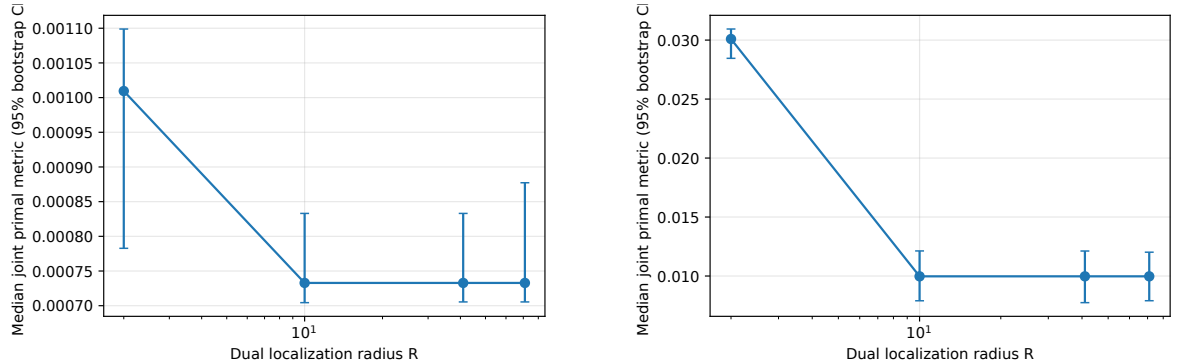
(a) Smooth sensitivity.

(b) Nonsmooth sensitivity.

Figure 8. Sensitivity of PB–ACGD-S to curvature and stabilizer inputs. Bars show median terminal J over ten seeds.

C.5 Dual-radius ablation

The main experiment uses the engineered localization $R = 2$. We rerun $r = 8$ for thirty seeds at $R \in \{2, 10, 41, 71\}$. The nonsmooth Slater formula gives $R = 41$ for the displayed problem data; $R = 71$ is a representative Slater-scale radius for the smooth quadratic. Error bars are nonparametric 95% bootstrap intervals for the median joint metric.



(a) Smooth radius ablation.

(b) Nonsmooth radius ablation.

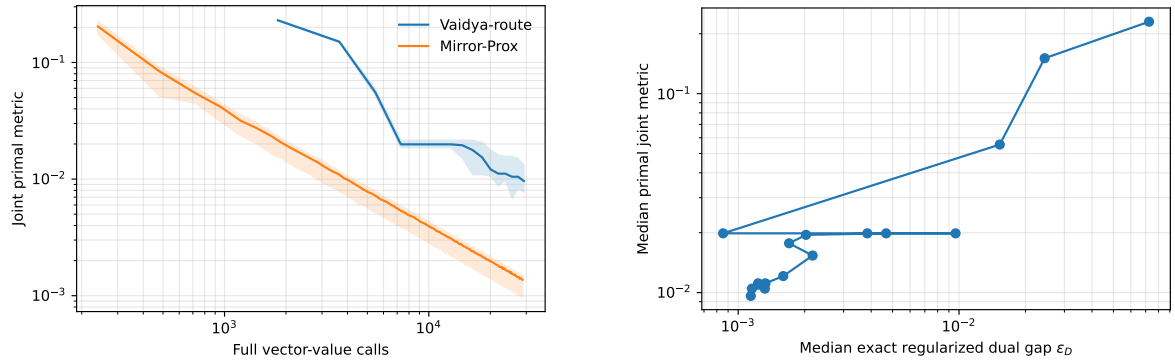
Figure 9. Sensitivity to the dual localization radius for batch $r = 8$ and thirty seeds.

For the nonsmooth problem, the median joint metric is $3.01 \cdot 10^{-2}$ at $R = 2$ and $9.97 \cdot 10^{-3}$ for $R \geq 10$. In the smooth problem it is $1.01 \cdot 10^{-3}$ at $R = 2$ and $7.33 \cdot 10^{-4}$ for $R \geq 10$, with strongly overlapping confidence intervals. This benign instance does not exhibit the worst-case radius dependence.

C.6 Small-constraint dual-route diagnostic

We use a separate $d = 20$, $m = 3$ smooth quadratic problem with three affine constraints, Slater margin 0.2, radius $R = 2$, and regularization $\mu = 0.05$. A minimal volumetric-center implementation uses sixteen cuts and omits inactive-cut deletion, so this is a route diagnostic rather than a test of asymptotic cutting-plane complexity.

Across twelve seeds, the dual route reaches $J \le 0.02, 0.015, 0.01$ in 12/12, 10/12, and 8/12 runs. The respective first-passage medians are 7296, 17328, and 23712 vector calls, compared with 2040, 2760, and 4080 for Mirror–Prox. The two stricter dual-route medians are conditional on successful runs; the remaining runs are right-censored at 29184 calls.



(a) Primal accuracy versus full-vector work. **(b)** Regularized dual gap versus primal metric.

Figure 10. Minimal $m = 3$ volumetric-center diagnostic of the Vaidya dual route. The stricter first-passage summaries are conditional on success and right-censored at sixteen cuts.

C.7 Finite-difference noise ablation

We estimate the second moment of the two-point Lagrangian gradient at the exact smooth saddle point from 30000 independent samples per radius. It is nearly independent of γ under common pathwise randomness and grows as γ^{-2} under fresh additive measurement noise.

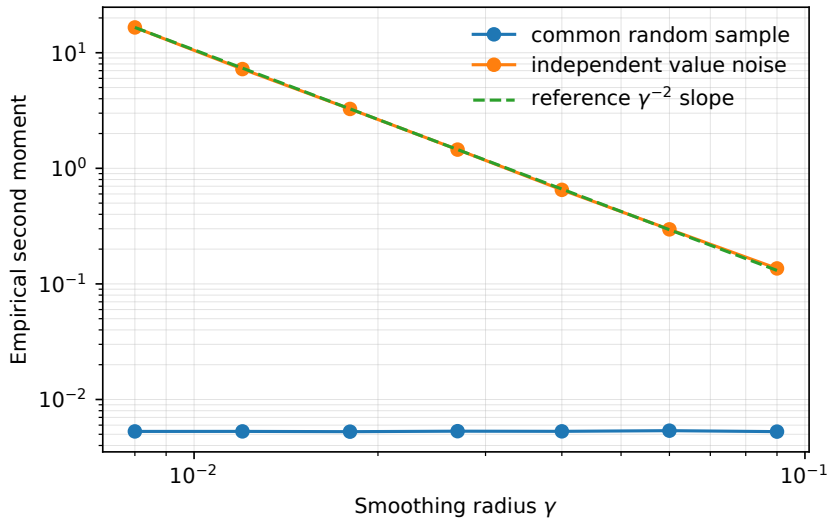


Figure 11. Empirical second moment of the two-point estimator versus smoothing radius. Independent measurement noise exhibits the predicted γ^{-2} amplification.

References

[Akhavan et al., 2022] Akhavan A., Chzhen E., Pontil M., Tsybakov A. B. A gradient estimator via ℓ_1-randomization for online zero-order optimization with two point feedback. // arXiv:2205.13910. — 2022.

[Akhavan et al., 2020] Akhavan A., Pontil M., Tsybakov A. B. Exploiting higher order smoothness in derivative-free optimization and continuous bandits. // arXiv:2006.07862. — 2020.

[Bach, Perchet, 2016] Bach F., Perchet V. Highly-smooth zero-th order online optimization. // Proceedings of the 29th Conference on Learning Theory. — PMLR, 2016. — Vol. 49. — P. 257–283.

[Flaxman et al., 2005] Flaxman A. D., Kalai A. T., McMahan H. B. Online convex optimization in the bandit setting: gradient descent without a gradient. // Proceedings of the 16th Annual ACM–SIAM Symposium on Discrete Algorithms. — 2005. — P. 385–394.

[Larson et al., 2019] Larson J., Menickelly M., Wild S. M. Derivative-free optimization methods. // Acta Numerica. — 2019. — Vol. 28. — P. 287–404. — DOI:10.1017/S0962492919000060.

[Usmanova et al., 2019] Usmanova I., Krause A., Kamgarpour M. Safe convex learning under uncertain constraints. // Proceedings of AISTATS. — PMLR, 2019. — Vol. 89. — P. 2106–2114.

[Yu, Neely, 2016] Yu H., Neely M. J. A low complexity algorithm with $O(\sqrt{T})$ regret and $O(1)$ constraint violations for online convex optimization with long term constraints. // arXiv:1604.02218. — 2016.

[Bayandina et al., 2018] Bayandina A., Dvurechensky P., Gasnikov A., Stonyakin F., Titov A. Mirror descent and convex optimization problems with non-smooth inequality constraints. // Large-Scale and Distributed Optimization / Eds. P. Giselsson, A. Rantzer. — Cham: Springer, 2018. — Lecture Notes in Mathematics, Vol. 2227. — P. 181–215. — DOI:10.1007/978-3-319-97478-1_8.

[Beznosikov et al., 2020] Beznosikov A., Sadiev A., Gasnikov A. Gradient-free methods with inexact oracle for convex–concave stochastic saddle-point problem. // Mathematical Optimization Theory and Operations Research. — Cham: Springer, 2020. — Communications in Computer and Information Science, Vol. 1275. — P. 105–119. — DOI:10.1007/978-3-030-58657-7_11.

[Boob et al., 2023] Boob D., Deng Q., Lan G. Stochastic first-order methods for convex and nonconvex functional constrained optimization. // Mathematical Programming. — 2023. — Vol. 197. — P. 215–279. — DOI:10.1007/s10107-021-01742-y.

[Bubeck et al., 2019] Bubeck S., Jiang Q., Lee Y.-T., Li Y., Sidford A. Complexity of highly parallel non-smooth convex optimization. // Advances in Neural Information Processing Systems. — 2019. — Vol. 32. — P. 13900–13909.

[Deng et al., 2024] Deng Q., Lan G., Lin Z. Uniformly optimal and parameter-free first-order methods for convex and function-constrained optimization. // arXiv:2412.06319v3. — first submitted 2024; v3 revised 2026.

[Diakonikolas, Guzmán, 2020] Diakonikolas J., Guzmán C. Lower bounds for parallel and randomized convex optimization. // Journal of Machine Learning Research. — 2020. — Vol. 21, no. 5. — P. 1–31.

[Duchi et al., 2015] Duchi J. C., Jordan M. I., Wainwright M. J., Wibisono A. Optimal rates for zero-order convex optimization: the power of two function evaluations. // IEEE Transactions on Information Theory. — 2015. — Vol. 61, no. 5. — P. 2788–2806. — DOI:10.1109/TIT.2015.2409256.

[Gasnikov et al., 2022] Gasnikov A., Novitskii A., Novitskii V., Abdukhakimov F., Kamzolov D., Beznosikov A., Takáč M., Dvurechensky P., Gu B. The power of first-order smooth optimization for black-box non-smooth problems. // Proceedings of the 39th International Conference on Machine Learning. — PMLR, 2022. — Vol. 162. — P. 7241–7265.

[Gladin et al., 2021] Gladin E., Sadiev A., Gasnikov A., Dvurechensky P., Beznosikov A., Alkousa M. Solving smooth min-min and min-max problems by mixed oracle algorithms. // Mathematical Optimization Theory and Operations Research: Recent Trends. — 2021. — CCIS, Vol. 1476. — P. 19–40. — DOI:10.1007/978-3-030-86433-0_2.

[Gladin et al., 2022] Gladin E., Gasnikov A., Ermakova E. Vaidya's method for convex stochastic optimization problems in small dimension. // Matematicheskie Zametki. — 2022. — Vol. 112, no. 2. — P. 179–187. — DOI:10.4213/mzm13430.

[Juditsky et al., 2011] Juditsky A., Nemirovski A., Tauvel C. Solving variational inequalities with stochastic Mirror–Prox algorithm. // Stochastic Systems. — 2011. — Vol. 1, no. 1. — P. 17–58. — DOI:10.1214/10-SSY011.

[Lan, Zhou, 2020] Lan G., Zhou Z. Algorithms for stochastic optimization with function or expectation constraints. // Computational Optimization and Applications. — 2020. — Vol. 76, no. 2. — P. 461–498. — DOI:10.1007/s10589-020-00179-x.

[Li et al., 2022] Li Z., Chen P.-Y., Liu S., Lu S., Xu Y. Zeroth-order optimization for composite problems with functional constraints. // Proceedings of the AAAI Conference on Artificial Intelligence. — 2022. — Vol. 36, no. 7. — P. 7453–7461. — DOI:10.1609/aaai.v36i7.20709.

[Nemirovski, 1994] Nemirovski A. On parallel complexity of nonsmooth convex optimization. // Journal of Complexity. — 1994. — Vol. 10, no. 4. — P. 451–463. — DOI:10.1006/jcom.1994.1025.

[Nemirovski, Yudin, 1983] Nemirovski A. S., Yudin D. B. Problem complexity and method efficiency in optimization. — New York: Wiley-Interscience, 1983.

[Nesterov, Spokoiny, 2017] Nesterov Y., Spokoiny V. Random gradient-free minimization of convex functions. // Foundations of Computational Mathematics. — 2017. — Vol. 17. — P. 527–566. — DOI:10.1007/s10208-015-9296-2.

[Nguyen, Balasubramanian, 2023] Nguyen A., Balasubramanian K. Stochastic zeroth-order functional constrained optimization: oracle complexity and applications. // INFORMS Journal on Optimization. — 2023. — Vol. 5, no. 3. — P. 256–272. — DOI:10.1287/ijoo.2022.0085.

[Nguyen, Gasnikov, 2026] Nguyen N. T., Gasnikov A. Sliding methods for Hölder-smooth convex–concave minimax optimization with bilinear coupling. // arXiv:2608.03846v1. — 4 Aug. 2026.

[Ouyang, Xu, 2021] Ouyang Y., Xu Y. Lower complexity bounds of first-order methods for convex–concave bilinear saddle-point problems. // Mathematical Programming. — 2021. — Vol. 185, no. 1–2. — P. 1–35. — DOI:10.1007/s10107-019-01420-0.

[Rajoriya et al., 2025] Rajoriya V., Pradhan P. P., Rajawat K. Hinge-proximal stochastic gradient methods for convex optimization with functional constraints. // arXiv:2512.13017. — 2025.

[Sanyal et al., 2025] Sanyal P., Thomdapu S. T., Rajawat K. Stochastic sequential quadratic programming for optimization with functional constraints. // arXiv:2511.20178v3. — first submitted 2025; v3 revised 2026.

[Shamir, 2017] Shamir O. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. // Journal of Machine Learning Research. — 2017. — Vol. 18, no. 52. — P. 1–11.

[Vaidya, 1996] Vaidya P. M. A new algorithm for minimizing convex functions over convex sets. // Mathematical Programming. — 1996. — Vol. 73, no. 3. — P. 291–341. — DOI:10.1007/BF02592216.

[Zhang, Lan, 2025] Zhang Z., Lan G. Solving convex smooth function constrained optimization is almost as easy as unconstrained optimization. // arXiv:2210.05807v4. — 30 Nov. 2025.