Gradient-Free Methods for Stochastic Convex Optimization with Stochastic Functional Constraints:

Sequential Oracle Rounds, Oracle Calls and Matrix–Vector Products

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

[1]Innopolis University [2]Moscow Institute of Physics and Technology
[3]Kharkevich Institute for Information Transmission Problems RAS [4]HSE University
abronin.vd@phystech.edu, gasnikov@yandex.ru, dmdvinskikh@hse.ru

September 8, 2026

Abstract

We study convex stochastic optimization problems with m stochastic functional (expectation) constraints, $\min\{f(x) = \mathbb{E}F(x,\xi) : \ g_i(x) = \mathbb{E}G_i(x,\xi) \le 0, \ i = 1,\dots,m, \ x \in Q\}$, in the setting where only noisy values of the objective and of the constraints are available. The oracle returns the vector $(F(x,\xi), G_1(x,\xi), \dots, G_m(x,\xi))$ at two points for one common realization of ξ (two-point feedback). The objective and the constraints may be smooth or nonsmooth, and the objective may be strongly convex. We measure the complexity of a method by three separate quantities: the number of *sequential oracle rounds* N_{seq} (rounds of queries whose locations are known in advance and can be issued in parallel), the number of *oracle calls*, recorded separately for the objective and for the constraints, and the number of *inner operations* that require no new round (proximal steps or matrix–vector products with stored Jacobian samples).

Randomized smoothing and a shifted Lagrangian reformulation reduce the problem to a convex–concave saddle-point problem whose primal part is smooth. We give complete and self-contained analyses, with explicit constants, of (i) a batched stochastic Mirror–Prox method, which serves as a baseline with $N_{\text{seq}} = O(\varepsilon^{-1})$ rounds for smooth data, and (ii) a zeroth-order primal–dual *sliding* method whose outer loop follows the accelerated scheme of Zhang and Lan and whose inner steps use fresh mini-batches of Jacobian samples drawn at the same query point. The sliding method needs $N_{\text{seq}} = O(\sqrt{(L_0 + RL_g)D^2/\varepsilon})$ rounds for smooth data and $N_{\text{seq}} = O(d^{1/4}\sqrt{M_R M}\, D/\varepsilon)$ rounds for nonsmooth data, while the total number of oracle calls remains $O(dM_R^2 D^2\varepsilon^{-2} + R^2\sigma^2 \log(m+1)\varepsilon^{-2})$ up to logarithmic factors in m and a dimension-free term, of which only $O(dM_R^2 D^2\varepsilon^{-2})$ calls use the objective (separate primal and dual mini-batches make this accounting exact); here d is the dimension, R bounds the Lagrange multipliers, M and $M_R = M_0 + RM_g$ are Lipschitz scales of the data and of the Lagrangian, D is the diameter of Q, and σ^2 is the variance of the constraint values. These round complexities coincide, up to the multiplier scale, with the best known ones for *unconstrained* zeroth-order problems. For strongly convex objectives a restart scheme gives $N_{\text{seq}} = O(\sqrt{L/\mu}\log(1/\varepsilon))$ rounds, $O(dM_R^2/(\mu\varepsilon))$ calls in which the objective is used, and constraint calls with an additional $\tilde{O}(R^2\sigma^2\log(m+1)/\varepsilon^2)$ term for the constraint-level information, whose ε^{-2} order cannot be improved even under strong convexity. For exactly known affine constraints the number of matrix–vector products with the constraint matrix is separated from the number of objective calls in the spirit of the ACGD-S method. We complement the upper bounds with information-theoretic lower bounds: $\Omega(dM^2D^2/\varepsilon^2)$ two-point calls for the objective and $\Omega(dM_g^2D^2/\varepsilon^2)$ for the constraints, $\Omega(\sigma^2\log(1+m)/\varepsilon^2)$ vector calls for the constraint levels, and $\Omega(m\sigma^2/\varepsilon^2)$ scalar calls. The objective term is thus minimax optimal up to constants, the constraint-gradient term is matched in d, M_g, D and ε at unit multipliers, and one logarithm in m is necessary; the dependence on the multiplier scale, a second logarithmic factor and the nonsmooth round complexity (best known for the two-point model, not minimax) remain open. Numerical experiments illustrate the separation between rounds and calls.

Keywords. zeroth-order optimization; gradient-free methods; stochastic functional constraints; expectation constraints; two-point feedback; randomized smoothing; primal–dual sliding; accelerated methods; Mirror–Prox; restarts; oracle complexity; parallel complexity; lower bounds.

MSC 2020. 90C25, 90C15, 90C06, 65K05, 68Q25.

Contents

1 Introduction **3**
- 1.1 The problem . . . 3
- 1.2 Three measures of complexity . . . 4
- 1.3 Related work . . . 4
- 1.4 Contributions . . . 6
- 1.5 Organization . . . 7

2 Setting, oracle model and complexity measures **7**
- 2.1 Notation . . . 7
- 2.2 Assumptions . . . 8
- 2.3 Oracle model . . . 10
- 2.4 Target accuracy . . . 11

3 Randomized smoothing and two-point estimators **11**
- 3.1 Smoothing by averaging over a ball . . . 11
- 3.2 Estimators . . . 12
- 3.3 Mini-batches in the ℓ_∞-norm . . . 14
- 3.4 Suprema of martingale transforms over the dual set . . . 14

4 The smoothed Lagrangian and the certificate **15**
- 4.1 The smoothed and shifted problem . . . 15
- 4.2 Lagrange multipliers . . . 16
- 4.3 Lagrangian, restricted gap and certificate . . . 17
- 4.4 Proximal geometry . . . 17

5 Baseline: batched stochastic Mirror–Prox **18**

6 Zeroth-order primal–dual sliding **20**
- 6.1 Idea . . . 20
- 6.2 The method . . . 20
- 6.3 The deterministic skeleton . . . 21
- 6.4 Main result . . . 23
- 6.5 Choice of the parameters and complexity . . . 24
- 6.6 Discussion . . . 27

7 Strongly convex objectives: restarts **28**
- 7.1 Restarted sliding . . . 29
- 7.2 Restarted Mirror–Prox . . . 30

8 Exactly known affine constraints: matrix–vector products **30**

9 Lower bounds **32**
9.1 Two-point calls for the objective . . . 32
9.2 Two-point calls for the constraints . . . 33
9.3 Calls for the constraint levels . . . 34
9.4 Depth . . . 34

10 Summary of the rates **35**

11 Numerical experiments **36**
11.1 Test problems . . . 36
11.2 Rounds versus calls . . . 37
11.3 Restarts on the strongly convex instance . . . 38
11.4 Two diagnostics . . . 39

12 Conclusion and open problems **40**
12.1 Open problems . . . 40

A Auxiliary facts **46**

B Proofs for Section 3 **47**

C Proof of Theorem 5.1 **49**

D Proofs for Section 6 **51**
D.1 Proof of Lemma 6.1 . . . 51
D.2 Proof of Lemma 6.2 . . . 52
D.3 Proof of Lemma 6.3 . . . 52
D.4 Proof of Lemma 6.4 . . . 53
D.5 Proof details for Theorem 8.1 . . . 54

E Proofs for Section 9 **55**
E.1 Proof of Theorem 9.1 . . . 55
E.2 Proof of Theorem 9.2 . . . 56
E.3 Proof of Theorem 9.3 . . . 56
E.4 Proof of Remark 9.5 . . . 57
E.5 Proof of Theorem 9.4 . . . 57

F Experimental details **58**

1 Introduction

1.1 The problem

We consider the stochastic convex optimization problem with m stochastic functional constraints

$$\min_{x\in Q} f(x) := \mathbb{E}_\xi F(x,\xi) \qquad \text{s.t.} \qquad g_i(x) := \mathbb{E}_\xi G_i(x,\xi) \le 0, \quad i = 1,\dots,m, \tag{1}$$

where $Q \subset \mathbb{R}^d$ is a convex compact set with Euclidean diameter D, the random variable ξ has an unknown distribution P, and the functions $f, g_1, \dots, g_m$ are convex. Neither f nor g_i, nor their (sub)gradients, are available. The only access to the problem is a *stochastic zeroth-order oracle*: given two points x^+, x^-, it draws one fresh sample $\xi \sim P$ and returns the two vectors

$$\big(F(x^+,\xi), G_1(x^+,\xi), \dots, G_m(x^+,\xi)\big) \quad \text{and} \quad \big(F(x^-,\xi), G_1(x^-,\xi), \dots, G_m(x^-,\xi)\big). \tag{2}$$

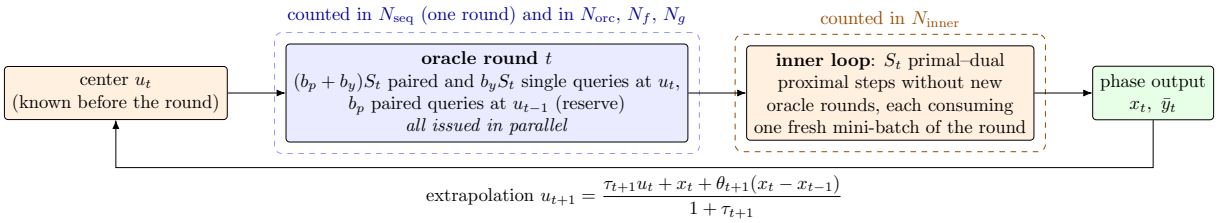


Figure 1: One phase of the zeroth-order primal–dual sliding method (Algorithm 2). All oracle queries of phase t are located at the current center u_t (plus one reserve mini-batch at u_{t-1}) and therefore form a single sequential round. The inner loop performs S_t proximal steps without further oracle rounds; each step uses a fresh mini-batch of the round, which keeps the estimation errors conditionally centered. The number of rounds is $N = O(\sqrt{LD^2/\varepsilon})$ for smooth data, while the total number of calls is $O\big((dM_R^2D^2 + R^2\sigma_g^2\log(m+1))/\varepsilon^2\big)$ up to logarithmic factors (gradient information with the factor d, level information with the factor $\log(m+1)$).

The common realization ξ at the two points is the classical *two-point feedback* model of Agarwal et al. [1], Duchi et al. [22], Shamir [56]. The constraints are *expectation constraints*: the oracle never reveals whether a point is feasible; feasibility can only be estimated from noisy values, and the noise in the constraint values, unlike the noise in the differences $F(x^+,\xi) - F(x^-,\xi)$, does not cancel. Problems of this type arise when the objective and the constraints are outputs of a simulator, of a physical experiment or of a non-differentiable model, and the constraints encode risk, budget or safety requirements that are themselves averages over uncertainty.

We are interested in methods that produce, for a prescribed accuracy $\varepsilon > 0$, a point $\hat{x} \in Q$ with

$$\mathbb{E}\big[f(\hat{x})\big] - f^* \le \varepsilon \qquad \text{and} \qquad \mathbb{E}\Big[\max_{1\le i\le m}[g_i(\hat{x})]_+\Big] \le \varepsilon, \tag{3}$$

where f^* is the optimal value of (1), and in a precise accounting of their cost.

1.2 Three measures of complexity

Following Gasnikov et al. [26], we distinguish the number of oracle calls from the number of *sequential* rounds of calls. A *round* is a set of queries whose locations are determined by the information available before the round; all queries of a round can be issued simultaneously (in parallel, or as a single batched request to a simulator). The number of rounds N_{seq} is the quantity that governs the wall-clock time of a method on a parallel system; the number of oracle calls N_{orc} (point evaluations of the vector $(F, G_1, \dots, G_m)$) governs the total sampling cost. For constrained problems we record in addition how many of these calls are used for the objective, N_f, and how many for the constraint vector, N_g, because the two pieces of information have different prices and, as we show, different intrinsic complexities (a call may serve both channels, so $N_{\text{orc}} \le N_f + N_g$). Finally, some methods perform many cheap steps between two rounds (proximal steps in a low-dimensional dual space, or multiplications of vectors by a stored constraint Jacobian). Their number N_{inner} is the analogue of the number of matrix–vector products in the accounting of Zhang and Lan [64]. Figure 1 depicts one round of the main method of this paper.

1.3 Related work

Zeroth-order optimization. Gradient-free methods go back to Nemirovsky and Yudin [47], Polyak and Tsybakov [53], Spall [58]. The modern analysis is based on randomized smoothing: replacing f by $f_\gamma(x) = \mathbb{E}f(x+\gamma v)$ with v uniform in the unit ball, whose gradient admits the single-point representation $\nabla f_\gamma(x) = \frac{d}{\gamma}\mathbb{E}[f(x+\gamma u)u]$ with u uniform on the unit sphere [25, 48]. With two function evaluations per sample the estimator $\frac{d}{2\gamma}[F(x+\gamma u,\xi) - F(x-\gamma u,\xi)]u$

has second moment $O(dM^2)$ independently of γ [1, 22, 56]; the constant $2d$ with the correct dimensional behaviour can be obtained from the Poincaré inequality on the sphere [26, 56]. This leads to the complexity $O(dM^2D^2/\varepsilon^2)$ calls for nonsmooth stochastic convex problems, which is optimal [22, 56]. Smoothing also makes the function $\sqrt{d}M/\gamma$-smooth [26], which opens the way to accelerated methods. Gasnikov et al. [26] used this fact to separate the number of sequential iterations from the number of oracle calls: for nonsmooth M-Lipschitz objectives on a set of diameter D one can reach accuracy ε in $O(d^{1/4}MD/\varepsilon)$ iterations of an accelerated batched method with $O(dM^2D^2/\varepsilon^2)$ calls in total, and in $O(d^{1/4}\sqrt{M^2/(\mu\varepsilon)}\log(1/\varepsilon))$ iterations with $O(dM^2/(\mu\varepsilon))$ calls for μ-strongly convex objectives. The same paper treats adversarial (non-random) noise of bounded magnitude and one-point feedback. Surveys of the area are Conn et al. [17], Gasnikov et al. [27], Larson et al. [37]. Further developments include exploiting higher-order smoothness [2, 4, 29, 50], heavy-tailed noise [33], stochastic and adversarial noise [42], kernel approximations [43], gradient-free saddle-point methods [10], and gradient-free methods with inexact oracles [9, 28].

Parallel complexity and rounds. The number of rounds needed by any (randomized) algorithm that may issue polynomially many queries per round was studied by Nemirovski [44], Balkanski and Singer [7], Duchi et al. [20], Bubeck et al. [14] and Diakonikolas and Guzmán [19]. For nonsmooth Lipschitz convex minimization on the Euclidean ball the smoothing-based upper bound $O(d^{1/4}/\varepsilon)$ of Duchi et al. [21] is not known to be tight: the best lower bounds [14, 19] are of order $\tilde{\Omega}(\min\{\varepsilon^{-2}, d^{1/3}\varepsilon^{-2/3}\})$ in the regime of many queries per round, and in the regime $\varepsilon \lesssim d^{-1/4}$ the depth $\tilde{O}(d^{1/3}\varepsilon^{-2/3})$ is attained by the ball-acceleration methods of Bubeck et al. [14] and, with a polynomial total number of stochastic gradient queries, of Carmon et al. [15]. These methods use (stochastic) gradients; for the two-point value oracle considered here the smoothing bound $d^{1/4}/\varepsilon$ remains the best known depth. For smooth convex minimization the round complexity $\Theta(\sqrt{LD^2/\varepsilon})$ is attained by accelerated methods with mini-batches [24, 30, 34]. Very recently, Zhang et al. [63] obtained near-optimal lower bounds for randomized algorithms with an *exact* scalar value oracle; that model differs from the stochastic common-sample two-point oracle of this paper, in which differences of values are noise-free but the values themselves are not. Parameter-free stochastic zeroth-order methods with near-optimal oracle complexity for *unconstrained* convex problems were obtained by Ren and Luo [54], whose adaptive choice of the step and of the smoothing radius is complementary to the fixed schedules used here. Our results transfer the above regimes to problems with functional constraints.

Functional constraints with first-order information. Subgradient methods that alternate between objective and constraint steps go back to Polyak [52]; their complexity for nonsmooth convex problems is $O(M^2D^2/\varepsilon^2)$ subgradient calls [8, 47]. For stochastic constraints the cooperative stochastic approximation of Lan and Zhou [36] attains the rate $O(1/\varepsilon^2)$ for the optimality gap and the constraint violation without dual variables. Primal–dual methods solve the Lagrangian saddle-point problem: the ConEx method of Boob et al. [11] and the methods of Hamedani and Aybat [31], Xu [61] handle smooth and nonsmooth constraints in the stochastic first-order setting; level-set methods [3, 40, 41] and the level-constrained methods of Boob et al. [12] avoid the dependence on the size of the multipliers, and Deng et al. [18] obtain parameter-free first-order methods with optimal oracle complexity for function-constrained problems. For smooth convex constraints Zhang and Lan [64] showed that the number of gradient evaluations can be reduced to the unconstrained level $O(\sqrt{L/\varepsilon})$ by their accelerated constrained gradient descent (ACGD), where L is the smoothness constant of the Lagrangian, $L_f + \|y^*\| L_g$ (this is the origin of the constant $H = L_0 + RL_g$ below), and that in the "sliding" variant ACGD-S the $O(1/\varepsilon)$ cost of the bilinear coupling is paid only in matrix–vector multiplications with the Jacobian of the constraints, not in gradient evaluations; Zhang and Lan [64] also prove lower bounds showing that, in their first-order computation model, these complexities cannot be improved. The lower

bound of Ouyang and Xu [51] for bilinearly coupled saddle-point problems shows that the $O(1/\varepsilon)$ matrix–vector complexity cannot be improved by methods that access the coupling matrix only through matrix–vector products.

Functional constraints with zeroth-order information. Much less is known. Nguyen and Balasubramanian [49] proposed and analyzed a zeroth-order version of ConEx with Gaussian smoothing and per-function gradient estimates, reaching $O((m+1)d/\varepsilon^2)$ calls in the convex case with the violation measured in the Euclidean norm of the vector $[g(\hat{x})]_+$; the factor $m+1$ comes from estimating the $m+1$ gradients with independent samples, i.e. from an oracle that returns one function per call, whereas the oracle of this paper returns all $m+1$ values at once. Safe zeroth-order convex learning with unknown constraints was studied by Usmanova et al. [60]. Related zeroth-order works treat nonconvex constrained problems [6, 39], distributed gradient-free mirror descent with constraints [62], block methods for Lipschitz expectation-valued problems [57] and gradient-free saddle-point problems [10, 23]. To the best of our knowledge, none of the works above separates sequential rounds from calls, distinguishes objective from constraint information, or gives accelerated round complexities for constrained zeroth-order problems.

Saddle-point methods. The Mirror–Prox algorithm of Nemirovski [45] and its stochastic version [32, 46] solve convex–concave saddle-point problems with $O(L/\varepsilon)$ iterations plus $O(\sigma^2/\varepsilon^2)$ samples; the primal–dual methods of Chambolle and Pock [16] with dual extrapolation are the inner engine of the sliding methods of Lan [35], Zhang and Lan [64].

1.4 Contributions

We give a complete treatment of problem (1) under two-point zeroth-order stochastic feedback, in which every result stated in the main text is proved in full in the paper, with explicit constants. The results are the following.

(C1) Reduction to a smooth saddle-point problem (Sections 3 and 4). Randomized smoothing of the objective *and* of the constraints, with an explicit shift of the constraint levels, produces a problem whose feasible set contains the original one and whose Lagrangian is smooth in the primal variable. A certificate $\mathsf{C}_{R,\gamma}(x) = f_\gamma(x) - f_\gamma^* + R\|[h(x)]_+\|_\infty$ controls both the optimality gap and the violation of the original problem (Theorems 4.1 and 4.3). Two-point estimators of the Lagrangian gradient have second moment $2d(M_0 + RM_g)^2$ regardless of the smoothing radius; the constraint levels are estimated by single calls. A new explicit bound on the ℓ_∞-norm of mini-batches of Jacobian samples acting on a fixed vector (Theorem 3.4) removes the factor d from the dual noise at the price of a logarithmic factor in m.

(C2) Batched stochastic Mirror–Prox (Section 5). A baseline method with a complete proof (Theorem 5.1): $\mathbb{E}\mathsf{C} \le \sqrt{3}L_\alpha \mathcal{D}_\alpha^2/N + 5\sigma_* \mathcal{D}_\alpha/\sqrt{N}$, hence $N_{\mathrm{seq}} = O(\varepsilon^{-1})$ rounds for smooth data with $O(\varepsilon^{-2})$ calls.

(C3) Zeroth-order primal–dual sliding (Section 6). The main method (Algorithm 2) combines the outer recursion of ACGD-S [64] with inner proximal steps driven by fresh mini-batches drawn at the same center. We prove (Theorem 6.5)

$$\mathbb{E}\,\mathsf{C}_{R,\gamma}(\bar{x}_N) \;\le\; \frac{2LD_A^2}{N(N+1)} + 4D_A\sqrt{\frac{17dM_R^2}{b_p K_N} + \frac{\varrho^2\sigma_y^2}{b_y K_N}},$$

where N is the number of rounds, $K_N = \sum_t S_t$ the number of inner steps and b_p, b_y the sizes of the primal and dual mini-batches. Consequently $N_{\mathrm{seq}} = O(\sqrt{(L_0 + RL_g)D^2/\varepsilon})$ for smooth data and $N_{\mathrm{seq}} = O(d^{1/4}\sqrt{M_R M D/\varepsilon})$ for nonsmooth data, with $N_f = O(N_{\mathrm{inner}} + dM_R^2 D^2/\varepsilon^2)$ calls for the objective and $N_g = N_{\mathrm{orc}} = O\big(N_{\mathrm{inner}} + [dM_R^2 D^2 + \Lambda_m R^2(\sigma_g^2 +$

$\nu_m M_g^2 D^2)]/\varepsilon^2\big)$ calls for the constraints, where $\Lambda_m = O((1+\log m)\log(m+1))$ and $\nu_m = O(\min\{d, 1+\log m\})$ (Theorems 6.9 and 6.10). These round complexities coincide, up to the multiplier scale, with the best known upper bounds for the unconstrained two-point problem. The analysis is organized around two inequalities that are of independent interest: a *sliding inequality* for arbitrary sequences (Theorem 6.3) and an *energy inequality* for the inner primal–dual loop with an explicit accounting of the stochastic errors (Theorem 6.4).

(C4) Strongly convex objectives (Section 7). A restart scheme based on the certificate yields $N_{\mathrm{seq}} = O(\sqrt{L/\mu}\log(\mu D^2/\varepsilon))$ rounds and, thanks to the separate primal and dual batch sizes, $N_f = O(N_{\mathrm{inner}} + dM_R^2/(\mu\varepsilon))$ calls in which the objective is used, while the constraint calls carry in addition an $\tilde{O}(R^2\sigma_g^2\log(m+1)/\varepsilon^2)$ term for the constraint-level information (Theorems 7.2 and 7.3); the latter term cannot be removed, even for strongly convex objectives (Theorem 9.3 and Theorem 9.5).

(C5) Exactly known affine constraints (Section 8). When $g(x) = Ax - c$ is known, one mini-batch per round suffices and the coupling cost appears only in matrix–vector products: $N_f = O(\sqrt{L_{0,\gamma}D^2/\varepsilon} + dM_0^2D^2/\varepsilon^2)$, $N_g = 0$ and $N_{\mathrm{inner}} = O(\sqrt{L_{0,\gamma}D^2/\varepsilon} + \|A\|_{2\to\infty} RD\sqrt{\log(m+1)}/\varepsilon)$ (Theorem 8.1), in exact analogy with ACGD-S.

(C6) Lower bounds (Section 9). Any algorithm with T two-point calls suffers an error $\Omega(MD\sqrt{d/T})$ on a class of problems satisfying our assumptions (Theorem 9.1); through a scaled epigraph reduction the bound $\Omega(M_gD\sqrt{d/T})$ holds for the constraint calls, with universal constants (Theorem 9.2); detecting the active constraint requires $\Omega(\sigma_g^2\log(1+m)/\varepsilon^2)$ vector calls and $\Omega(m\sigma_g^2/\varepsilon^2)$ scalar calls (Theorems 9.3 and 9.4), which explains the $\log(m+1)$ factors of the entropy dual geometry. The objective term is matched up to constants, the constraint-gradient term in d, M_g, D and ε but not in the multiplier factor R^2, and the level term up to a factor $\kappa_m R^2$; the nonsmooth depth is the best known one for the two-point model but not a minimax statement (Table 5).

(C7) Experiments (Section 11). On a smooth and a nonsmooth instance with $d = 24$ and $m = 8$, run with the schedule of Theorem 6.5, we observe the predicted separation between rounds and calls, and we illustrate the restart scheme (also at equal budgets) and the two information-theoretic effects (common-sample variance and logarithmic dependence on m).

Open questions suggested by these results are collected in Section 12.1. Table 1 summarizes the rates; Table 2 compares them with the literature.

1.5 Organization

Section 2 fixes the notation, the assumptions, the oracle model and the complexity measures. Section 3 collects the properties of randomized smoothing and of the two-point estimators, the ℓ_∞ mini-batch lemma and the martingale-supremum lemma. Section 4 introduces the smoothed shifted Lagrangian, the multiplier bound and the certificate. Section 5 analyzes the batched Mirror–Prox baseline. Section 6 is the core of the paper: the sliding method, its deterministic skeleton and the stochastic theorem. Section 7 treats strongly convex objectives, Section 8 exactly known affine constraints, Section 9 lower bounds. Section 10 collects all rates in tables, Section 11 reports the experiments and Section 12 discusses open problems. Long proofs are in the appendices, which are self-contained.

2 Setting, oracle model and complexity measures

2.1 Notation

$\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^d$ and $\langle\cdot,\cdot\rangle$ the Euclidean inner product; $\|\cdot\|_1$ and $\|\cdot\|_\infty$ are the ℓ_1- and ℓ_∞-norms on $\mathbb{R}^m$. For a matrix $J \in \mathbb{R}^{m\times d}$ with rows $J_1, \dots, J_m$ we write

Table 1: Overview of the upper bounds proved in this paper (constants and log factors in m omitted; see Section 10 for the exact statements). Here $M_R = M_0 + RM_g$, $H = L_0 + RL_g$, σ_g^2 is the variance proxy of the constraint values, $\Lambda_m = (1 + \log 2m)\log(m+1)$, $\nu_m = \min\{4d, 92\kappa_m\}$, N_{seq} counts sequential oracle rounds, N_f and N_g count the vector calls (point evaluations of the whole vector (F, G)) in which the objective, resp. the constraint component is used, and N_{inner} counts inner steps that require no new round. In all methods every call carries the constraint vector, so $N_{\mathrm{orc}} = N_g$ (and $N_g = 0$, $N_{\mathrm{orc}} = N_f$ for known affine constraints). The terms of the call counts of the order of N_{inner} are omitted.

Regime	N_{seq}	N_f	$N_g = N_{\mathrm{orc}}$	N_{inner}
Mirror–Prox baseline (Theorem 5.1)				
smooth f, smooth g	$\dfrac{HD^2 + M_g RD\sqrt{\log m}}{\varepsilon}$	$\dfrac{dM_R^2 D^2 + \Lambda_m R^2 \sigma_g^2}{\varepsilon^2}$	$= N_f$	$= N_{\mathrm{seq}}$
nonsmooth	$\dfrac{\sqrt{d} M_R M D^2}{\varepsilon^2}$	$\dfrac{dM_R^2 D^2 + \Lambda_m R^2 \sigma_g^2}{\varepsilon^2}$	$= N_f$	$= N_{\mathrm{seq}}$
Zeroth-order sliding (Theorem 6.5)				
smooth f, smooth g	$\sqrt{\dfrac{HD^2}{\varepsilon}}$	$\dfrac{dM_R^2 D^2}{\varepsilon^2}$	$\dfrac{dM_R^2 D^2 + \Lambda_m R^2(\sigma_g^2 + \nu_m M_g^2 D^2)}{\varepsilon^2}$	$N_{\mathrm{seq}} + \dfrac{M_g RD\sqrt{\log m}}{\varepsilon}$
nonsmooth	$\dfrac{d^{1/4}\sqrt{M_R M} D}{\varepsilon}$	$\dfrac{dM_R^2 D^2}{\varepsilon^2}$	$\dfrac{dM_R^2 D^2 + \Lambda_m R^2(\sigma_g^2 + \nu_m M_g^2 D^2)}{\varepsilon^2}$	$N_{\mathrm{seq}} + \dfrac{M_g RD\sqrt{\log m}}{\varepsilon}$
μ-s.c., smooth	$\sqrt{\dfrac{H}{\mu}} \log \dfrac{\mu D^2}{\varepsilon}$	$\dfrac{dM_R^2}{\mu\varepsilon}$	$\dfrac{dM_R^2}{\mu\varepsilon} + \dfrac{\Lambda_m R^2(\sigma_g^2 + \nu_m M_g^2 D^2)}{\varepsilon^2}$	$N_{\mathrm{seq}} + M_g R\sqrt{\dfrac{\log m}{\mu\varepsilon}}$
μ-s.c., nonsmooth	$d^{1/4}\sqrt{\dfrac{M_R M}{\mu\varepsilon}} \log \dfrac{\mu D^2}{\varepsilon}$	$\dfrac{dM_R^2}{\mu\varepsilon}$	$\dfrac{dM_R^2}{\mu\varepsilon} + \dfrac{\Lambda_m R^2(\sigma_g^2 + \nu_m M_g^2 D^2)}{\varepsilon^2}$	$N_{\mathrm{seq}} + M_g R\sqrt{\dfrac{\log m}{\mu\varepsilon}}$
Exactly known affine constraints (Theorem 8.1), $N_g = 0$				
smooth f	$\sqrt{\dfrac{L_0 D^2}{\varepsilon}}$	$\dfrac{dM_0^2 D^2}{\varepsilon^2}$	0	$N_{\mathrm{seq}} + \dfrac{\|A\|_{2\to\infty} RD\sqrt{\log m}}{\varepsilon}$
nonsmooth f	$\dfrac{d^{1/4} M_0 D}{\varepsilon}$	$\dfrac{dM_0^2 D^2}{\varepsilon^2}$	0	$N_{\mathrm{seq}} + \dfrac{\|A\|_{2\to\infty} RD\sqrt{\log m}}{\varepsilon}$

$\|J\|_{2\to\infty} := \max_i \|J_i\|$; then

$$\|Jw\|_\infty \le \|J\|_{2\to\infty} \|w\| \quad (w \in \mathbb{R}^d), \qquad \left\|J^\top y\right\| \le \|J\|_{2\to\infty} \|y\|_1 \quad (y \in \mathbb{R}^m). \tag{4}$$

$\mathbb{B} = \{v \in \mathbb{R}^d : \|v\| \le 1\}$ is the unit ball, $\mathbb{S} = \{u : \|u\| = 1\}$ the unit sphere, $U(\mathbb{B})$ and $U(\mathbb{S})$ the uniform distributions on them. $[z]_+ = \max\{z, 0\}$ componentwise, and $\mathsf{v}(x) := \|[g(x)]_+\|_\infty = \max_i [g_i(x)]_+$ is the maximal constraint violation. $Q \subset \mathbb{R}^d$ is convex and compact, $D := \max_{x,x' \in Q} \|x - x'\|$ its diameter and $Q_{\bar\gamma} := Q + \bar\gamma\mathbb{B}$ its $\bar\gamma$-neighbourhood, where $\bar\gamma > 0$ is a fixed upper bound for the smoothing radii used below. Throughout,

$$\kappa_m := 1 + \log(2m), \qquad \Lambda_m := \kappa_m \log(m+1), \tag{5}$$

and log is the natural logarithm. Table 3 lists the recurring symbols.

2.2 Assumptions

The functions $F(\cdot, \xi)$ and $G_i(\cdot, \xi)$ are defined on $\mathbb{R}^d$ for every ξ, and all functions of x below are assumed measurable in ξ and integrable, so that the expectations exist. We say that a differentiable function φ is L-smooth if $\nabla\varphi$ is L-Lipschitz, and μ-strongly convex if $\varphi - \frac{\mu}{2}\|\cdot\|^2$ is convex.

Assumption 2.1 (convexity)**.** $f, g_1, \dots, g_m : \mathbb{R}^d \to \mathbb{R}$ are convex.

Assumption 2.2 (Lipschitz samples)**.** For every ξ the function $F(\cdot, \xi)$ is $M_0(\xi)$-Lipschitz and $G_i(\cdot, \xi)$ is $M_i(\xi)$-Lipschitz on the neighbourhood $Q_{\bar\gamma}$ of Q, $i = 1, \dots, m$, where $M_g(\xi) := \max_i M_i(\xi)$ and

$$\mathbb{E}\, M_0(\xi)^2 \le M_0^2, \qquad \mathbb{E}\, M_g(\xi)^2 \le M_g^2,$$

with constants $M_0, M_g > 0$. We put $M := \max\{M_0, M_g\}$.

Table 2: Comparison with related results (constants omitted). "FO" means stochastic first-order information, "ZO" zeroth-order information, "pairs" two-point calls, d is the dimension. Complexities are for the convex nonsmooth case unless stated otherwise; N_{seq} is listed only for papers that account for it. The information models differ: in Nguyen and Balasubramanian [49] a call returns one function, the $m+1$ gradients are estimated from independent samples and the violation is measured in ℓ_2; in this paper one call returns all $m+1$ values with a common sample and the violation is measured in ℓ_∞, so part of the factor $m+1$ reflects the oracle model rather than the algorithm.

Method	Setting	Calls	Rounds / remarks
CSA [36]	FO, stochastic g	$1/\varepsilon^2$	no dual variables
ConEx [11]	FO, stochastic g	$1/\varepsilon^2$	$\sqrt{1/\varepsilon}$ gradients for smooth deterministic data
ACGD-S [64]	deterministic FO, smooth f, g	$\sqrt{L/\varepsilon}$ gradients	$\|\nabla g\|_{2\to\infty}\, RD/\varepsilon$ matrix–vector products
SZO-ConEx [49]	ZO, stochastic g	$(m+1)d/\varepsilon^2$	violation in ℓ_2; independent estimates per function
Gasnikov et al. [26]	ZO pairs, unconstrained	dM^2D^2/ε^2	$N_{\text{seq}} = d^{1/4}MD/\varepsilon$
Theorem 5.1	ZO pairs, stochastic g	$(dM_R^2D^2 + \Lambda_m R^2\sigma_g^2)/\varepsilon^2$	$N_{\text{seq}} = \sqrt{d}M_RMD^2/\varepsilon^2$ (nonsmooth), HD^2/ε (smooth)
Theorem 6.5	ZO pairs, stochastic g	$N_f = dM_R^2D^2/\varepsilon^2$; $N_g = dM_R^2D^2/\varepsilon^2 + \Lambda_m R^2(\sigma_g^2 + \nu_m M_g^2 D^2)/\varepsilon^2$	$N_{\text{seq}} = d^{1/4}\sqrt{M_R M}D/\varepsilon$ (nonsmooth), $\sqrt{HD^2/\varepsilon}$ (smooth)
Theorem 8.1	ZO pairs, known affine g	$N_f = dM_0^2D^2/\varepsilon^2$, $N_g = 0$	$N_{\text{inner}} = \|A\|_{2\to\infty}\, RD\sqrt{\log m}/\varepsilon$
Theorems 9.1 and 9.2	ZO pairs	$\Omega(dM_0^2D^2/\varepsilon^2)$ for f, $\Omega(dM_g^2D^2/\varepsilon^2)$ for g	lower bounds (unit multiplier)
Theorems 9.3 and 9.4	ZO, constraint levels	$\Omega(\sigma_g^2 \log(1+m)/\varepsilon^2)$ vector calls, $\Omega(m\sigma_g^2/\varepsilon^2)$ scalar calls	lower bounds

By Jensen's inequality, Theorem 2.2 implies that f is M_0-Lipschitz and every g_i is M_g-Lipschitz on $Q_{\bar\gamma}$. Note that no bound on the absolute values $F(x,\xi)$ or on their variance is required: the methods below never use the value of the objective at a single point, only differences at pairs of points evaluated with the same ξ. The Lipschitz property is required only on $Q_{\bar\gamma}$, which contains every point the algorithms query; Theorem 3.1(e) below extends the expected functions from $Q_{\bar\gamma}$ to $\mathbb{R}^d$ whenever a global property is convenient.

Assumption 2.3 (noise of the constraint values)**.** $\mathbb{E}\,\|G(x,\xi) - g(x)\|_\infty^2 \le \sigma_g^2$ for all $x \in Q_{\bar\gamma}$, where $G = (G_1, \dots, G_m)$.

Assumption 2.4 (smoothness, used when stated)**.** f is L_0-smooth and each g_i is L_g-smooth on $\mathbb{R}^d$.

Assumption 2.5 (Slater condition)**.** There exist $x^{\mathrm{s}} \in Q$ and $\rho > 0$ with $g_i(x^{\mathrm{s}}) \le -\rho$ for all i.

Assumption 2.6 (strong convexity, used when stated)**.** f is μ-strongly convex on $Q_{\bar\gamma}$, i.e. $f - \frac{\mu}{2}\|\cdot\|^2$ is convex on $Q_{\bar\gamma}$.

Remark 2.7 (on the standing assumptions). (i) *Local versus global properties.* The Lipschitz property (Theorem 2.2) and the strong convexity (Theorem 2.6) are required only on the neighbourhood $Q_{\bar\gamma}$ of the compact set Q. This is not a mere convenience: a function cannot be both M-Lipschitz and μ-strongly convex on all of $\mathbb{R}^d$ (for a unit vector v, strong convexity gives $\varphi(x+tv) + \varphi(x-tv) - 2\varphi(x) \ge \mu t^2$, the Lipschitz property gives $\le 2Mt$, and the two are incompatible for $t > 2M/\mu$), so global versions of Theorems 2.2 and 2.6 would define an empty class. On $Q_{\bar\gamma}$ the two assumptions are compatible (a strongly convex quadratic is Lipschitz on every bounded set), and $Q_{\bar\gamma}$ is all the algorithms ever see: the Lipschitz property of the *samples* enters only through the moment bounds of Theorems 3.3 to 3.5 and the bias bounds

Table 3: Recurring notation.

Symbol	Meaning	Defined in
$d,\ m$	dimension, number of constraints	(1)
$Q,\ D,\ Q_{\bar\gamma}$	feasible set, its diameter, its $\bar\gamma$-neighbourhood	Section 2
$M_0, M_g,\ M := \max\{M_0, M_g\}$	Lipschitz moduli (mean square) of the samples	Theorem 2.2
$L_0,\ L_g$	Lipschitz constants of ∇f, ∇g_i (when smooth)	Theorem 2.4
$\sigma_g^2,\ \sigma_h^2$	variance proxies of the constraint values	Theorem 2.3, (10)
$\rho,\ x^{\mathrm{s}}$	Slater margin and Slater point	Theorem 2.5
μ	strong convexity parameter of f	Theorem 2.6
$\gamma,\ f_\gamma,\ g_{i,\gamma}$	smoothing radius and smoothed functions	(6)
$L_{0,\gamma},\ L_{g,\gamma},\ H,\ L$	smoothness constants after smoothing	Theorem 3.1, (13)
$b_0,\ b_i,\ h_i = g_{i,\gamma} - b_i$	bias bounds and shifted smoothed constraints	(11)
$R,\ Y_R,\ \tilde Y_R,\ M_R$	multiplier bound, dual sets, $M_R := M_0 + RM_g$	(14), (15)
$\Phi_\gamma,\ \mathsf{C}_{R,\gamma}$	smoothed Lagrangian and certificate	(16), (17)
$x_\gamma^*,\ f_\gamma^*,\ y_\gamma^*$	solution, value and multipliers of the smoothed problem	Theorem 4.2
$\mathcal{A},\ \mathcal{B}$	primal and dual prox radii: $\mathcal{A} = \frac{1}{2}\left\|x_0 - x_\gamma^*\right\|^2$ (unknown; replaced by $D^2/2$ in parameter choices), $\mathcal{B} = R^2\log(m+1)$	(21)
$V_{\mathcal{X}},\ V_{\mathcal{Y}}$	Euclidean and entropic Bregman distances	(20)
$\hat g,\ \hat J,\ \hat h$	two-point gradient, Jacobian and one-point level estimators	(9)
ν_m	constant in the Jacobian-action bound: $\nu_m = \min\{4d, 92\kappa_m\}$	Theorem 3.4
$N_{\mathrm{orc}},\ N_f,\ N_g,\ N_{\mathrm{seq}},\ N_{\mathrm{inner}}$	vector calls, calls used for f and for g, rounds, inner steps	Theorem 2.8
$b,\ \lambda,\ \varrho$	mini-batch size, stabilizer, dual scale	Algorithm 2
$S_t,\ a_t,\ \eta_t,\ \beta_t,\ K_N,\ A_N,\ W_N$	inner schedule of the sliding method	(28)
$D_A,\ \Sigma_b,\ \sigma_y$	radius and noise constants of the sliding method	(37), (39)
$b_p,\ b_y$	sizes of the primal and dual mini-batches of Algorithm 2	Algorithm 2

of Theorem 3.1(a), all evaluated at query points in $Q_{\bar\gamma}$; the Lipschitz property of the *expected* functions is used on Q in Theorem 4.2; and strong convexity is used on Q in Theorem 4.3. The only place where a global property is needed is the sliding inequality (Theorem 6.3), which requires the smoothed Lagrangian $\Phi_\gamma(\cdot, y)$ to be convex and L-smooth on $\mathbb{R}^d$. In the nonsmooth regime this is obtained by the extension device of Theorem 3.1(e): a convex function that is M-Lipschitz on $Q_{\bar\gamma}$ has a convex M-Lipschitz extension to $\mathbb{R}^d$ that coincides with it on $Q_{\bar\gamma}$, hence has the same smoothing on Q and produces the same oracle answers; the extension is used only for the smoothness of the Lagrangian, never for its strong convexity (Theorem 3.2). In the smooth regime Theorem 2.4 is stated on $\mathbb{R}^d$; it is the only global requirement of the paper, it is compatible with Theorems 2.2 and 2.6 (quadratics satisfy all three), and it may be read as the assumption that the data admit convex L_0-, L_g-smooth extensions, since a function that is convex and L-smooth only on a neighbourhood of Q need not have such an extension with the same constant. The test instances of Section 11 satisfy Theorems 2.2, 2.4 and 2.6 in exactly this form (Section 11.1). (ii) Theorem 2.3 is a genuine additional assumption: Theorem 2.2 bounds the differences $G_i(x,\xi) - G_i(x',\xi)$ but says nothing about $G_i(x,\xi) - g_i(x)$, which may have infinite variance (a random offset independent of x). The constraint levels enter the dual updates and their noise does not cancel. (iii) Theorem 2.5 is used only to bound the Lagrange multipliers; any known upper bound on $\left\|y_\gamma^*\right\|_1$ can replace it (see (14)). (iv) The constants M_0, M_g are taken positive and, in the smooth regime, $\max\{L_0, L_g\} > 0$ (if $L_0 = L_g = 0$ all data are affine and Section 8 applies); this only excludes degenerate cases in which the corresponding terms of the bounds are absent, and avoids divisions by zero in the parameter rules.

2.3 Oracle model

We fix the units in which oracle access is measured. A *vector call* (one point evaluation) at a point $x \in Q_{\bar\gamma}$ returns $(F(x,\xi), G(x,\xi)) \in \mathbb{R}^{m+1}$ for one realization ξ; it is the unit of the physical oracle cost. A *paired query* (or two-point call) at (x^+, x^-) consists of *two* vector calls with one

common fresh $\xi \sim P$ independent of everything else, as in (2). A *single query* is one vector call with a fresh ξ. A method may use only the objective component or only the constraint component of a call; since the two pieces of information have different prices and, as we show, different intrinsic complexities, we count the calls in which each of them is used in addition to the total number of calls:

Definition 2.8 (complexity measures)**.** For a run of an algorithm we denote by

- N_{orc} the total number of vector calls (point evaluations) issued by the algorithm; N_f the number of those calls whose objective component $F(x,\xi)$ is used, and N_g the number of those calls whose constraint component $G(x,\xi)$ is used (one call returns all m components). A call may be counted in both channels, so that $\max\{N_f, N_g\} \le N_{\mathrm{orc}} \le N_f + N_g$; in all methods of this paper every call is used in at least one channel, and N_{orc} is the number that is compared across methods and in the experiments;
- N_{seq} the number of *sequential oracle rounds*: the run is split into rounds, and the locations of all queries of a round are measurable with respect to the information (oracle answers and internal randomness) available before the round;
- N_{inner} the number of inner operations that require no new oracle round: proximal steps on Q and on the dual set, or matrix–vector products with a constraint Jacobian. In Algorithm 2 these steps consume oracle samples that were drawn in the current round (see Section 6.6); in Algorithm 4 they are oracle-free.

The lower bounds of Section 9 are stated in the number T of paired queries or of single vector calls, as indicated there; a bound of T paired queries is a bound of $2T$ vector calls, which only affects constants. In the *scalar* oracle model of Theorem 9.4 a call returns a single component $G_i(x,\xi)$ chosen by the algorithm; in the *vector* model all m components are returned at once. All algorithms of this paper use the vector model.

2.4 Target accuracy

For $\varepsilon > 0$ a random point $\hat{x} \in Q$ is an *ε-solution in expectation* if (3) holds. All upper bounds are stated for this criterion. The lower bounds of Section 9 bound from below the expectation of the *sum* of the optimality gap and of the violation (or the gap alone when the constraints are inactive); since (3) implies $\mathbb{E}[f(\hat{x}) - f^*] + \mathbb{E}\mathsf{v}(\hat{x}) \le 2\varepsilon$, a lower bound on the sum is a lower bound for the criterion (3) up to a factor 2 in ε, so that upper and lower bounds are comparable. (Note that $f(\hat{x}) - f^*$ may be negative at infeasible points, which is why the sum rather than the maximum is used.) The passage from bounds in expectation to bounds in probability by independent repetitions is standard and is discussed in Section 12.1.

3 Randomized smoothing and two-point estimators

3.1 Smoothing by averaging over a ball

For a function $\varphi : \mathbb{R}^d \to \mathbb{R}$ and a radius $\gamma > 0$ define

$$\varphi_\gamma(x) := \mathbb{E}_{v \sim U(\mathbb{B})} \varphi(x + \gamma v), \qquad x \in \mathbb{R}^d. \tag{6}$$

We apply (6) to f and to every g_i and write f_γ, $g_{i,\gamma}$, $g_\gamma = (g_{1,\gamma}, \dots, g_{m,\gamma})$ and $J_\gamma(x) \in \mathbb{R}^{m \times d}$ for the matrix with rows $\nabla g_{i,\gamma}(x)^\top$.

Lemma 3.1 (properties of the smoothing)**.** *Let $\varphi : \mathbb{R}^d \to \mathbb{R}$ be convex and $\gamma > 0$; for a convex set $C \subseteq \mathbb{R}^d$ write $C_\gamma := C + \gamma\mathbb{B}$.*

(a) *φ_γ is convex. If φ is M-Lipschitz on C_γ, then φ_γ is M-Lipschitz on C and $\varphi(x) \le \varphi_\gamma(x) \le \varphi(x)+\gamma M$ for all $x \in C$. If φ is L-smooth on $\mathbb{R}^d$, then $\varphi_\gamma(x) \le \varphi(x)+\frac{L\gamma^2}{2}\frac{d}{d+2} \le \varphi(x)+\frac{L\gamma^2}{2}$ for all x. If φ is affine, $\varphi_\gamma = \varphi$.*

(b) *If φ is M-Lipschitz on $\mathbb{R}^d$, or L-smooth on $\mathbb{R}^d$, then φ_γ is continuously differentiable on $\mathbb{R}^d$ and, for $u \sim U(\mathbb{S})$,*

$$\nabla\varphi_\gamma(x) = \frac{d}{\gamma}\,\mathbb{E}\big[\varphi(x+\gamma u)\,u\big] = \frac{d}{2\gamma}\,\mathbb{E}\big[\big(\varphi(x+\gamma u)-\varphi(x-\gamma u)\big)u\big]. \tag{7}$$

(c) *If φ is M-Lipschitz on $\mathbb{R}^d$, then $\nabla\varphi_\gamma$ is L_γ-Lipschitz on $\mathbb{R}^d$ with $L_\gamma := \sqrt{d}\,M/\gamma$. If φ is L-smooth on $\mathbb{R}^d$, then $\nabla\varphi_\gamma(x) = \mathbb{E}\nabla\varphi(x+\gamma v)$ and $\nabla\varphi_\gamma$ is L-Lipschitz on $\mathbb{R}^d$.*

(d) *If φ is μ-strongly convex on C_γ, then φ_γ is μ-strongly convex on C.*

(e) *(extension) If φ is M-Lipschitz on $C_{\bar\gamma}$ for some $\bar\gamma \ge \gamma$, then*

$$\tilde\varphi(x) := \inf_{z\in C_{\bar\gamma}} \big\{\varphi(z) + M\,\|x-z\|\big\}, \qquad x \in \mathbb{R}^d, \tag{8}$$

is convex and M-Lipschitz on $\mathbb{R}^d$, coincides with φ on $C_{\bar\gamma}$, and $\tilde\varphi_\gamma = \varphi_\gamma$ on C.

The proof is given in Section B. The representation (7) is classical [25, 48]; we include a proof because the functions involved are merely Lipschitz, and the constant $\sqrt{d}M/\gamma$ in (c) is the one obtained in Gasnikov et al. [26]. Figure 2 illustrates (a) and the level shift used in Section 4.

Convention 3.2 (which functions are smoothed). Under Theorem 2.2 the expected functions f, g_i are Lipschitz on $Q_{\bar\gamma}$ only. For each of them we fix once and for all the function to which the smoothing (6) is applied: if the function is L-smooth on $\mathbb{R}^d$ (Theorem 2.4, or an affine constraint), it is smoothed as it is; otherwise it is replaced by its extension (8) from $Q_{\bar\gamma}$ (with $C = Q$ and $M = M_0$, resp. M_g), which is convex and Lipschitz on $\mathbb{R}^d$. Throughout, f_γ and $g_{i,\gamma}$ denote the smoothings of the functions so fixed. Nothing that the algorithms or the guarantees can see depends on this choice: the oracle is only queried in $Q_{\bar\gamma}$, where the extension coincides with the original function (Theorem 3.1(e)), so the estimators of Section 3.2 are unchanged; by Theorem 3.1(e) the smoothed functions on Q, hence the smoothed problem (12), its value, its solutions and its multipliers, are unchanged; and the transfer to the original problem (Theorem 4.1) only involves points of Q. What the convention buys is that each f_γ, $g_{i,\gamma}$ is convex and smooth on all of $\mathbb{R}^d$ (Theorem 3.1(c)), with constant L_0, resp. L_g, in the smooth case and $\sqrt{d}M_0/\gamma$, resp. $\sqrt{d}M_g/\gamma$, in the Lipschitz case; this is the property needed in Theorem 6.3. The extension is never used for strong convexity: Theorem 2.6 concerns f on $Q_{\bar\gamma}$, where the extension equals f, and Theorem 3.1(d) with $C = Q$ gives the strong convexity of f_γ on Q, which is where Theorem 4.3 uses it.

3.2 Estimators

Let $u \sim U(\mathbb{S})$ and $v \sim U(\mathbb{B})$ be drawn independently of $\xi \sim P$. For $x \in Q$, $y \in \mathbb{R}^m$ and a vector of shifts $b = (b_1, \dots, b_m)$ (fixed in (11) below) we use the estimators

$$\begin{aligned} \hat g_0(x) &:= \frac{d}{2\gamma}\big[F(x+\gamma u,\xi) - F(x-\gamma u,\xi)\big]u, & \hat J(x) &:= \frac{d}{2\gamma}\big[G(x+\gamma u,\xi) - G(x-\gamma u,\xi)\big]u^\top, \\ \hat g_y(x) &:= \hat g_0(x) + \hat J(x)^\top y, & \hat h(x) &:= G(x+\gamma v,\xi) - b \in \mathbb{R}^m, \end{aligned} \tag{9}$$

with $\hat J(x) \in \mathbb{R}^{m\times d}$. $\hat g_0$ and $\hat J$ are computed from one paired query at $(x+\gamma u, x-\gamma u)$; $\hat h$ from one single query. The following lemma is the reason why the smoothing radius may be taken arbitrarily small when the data are smooth: the second moment of the two-point estimator does not depend on γ.

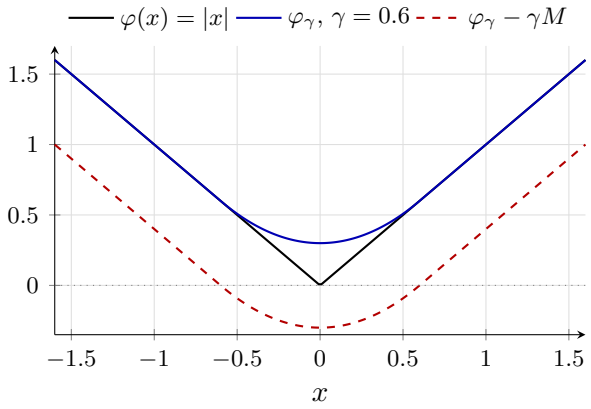


Figure 2: One-dimensional illustration of Theorem 3.1(a) for $\varphi(x) = |x|$ ($M = 1$, $d = 1$): $\varphi \le \varphi_\gamma \le \varphi + \gamma M$, and the shifted function $h = \varphi_\gamma - \gamma M$ satisfies $h \le \varphi$, so that the set $\{h \le 0\}$ contains $\{\varphi \le 0\}$. In the constrained problem this inclusion guarantees that the optimal value of the smoothed problem does not exceed the original one (Theorem 4.1).

Lemma 3.3 (two-point estimator)**.** *Let $x \in Q$ and $\gamma \le \bar\gamma$. Let $\Phi(\cdot,\xi)$ be $\Lambda(\xi)$-Lipschitz on $Q_{\bar\gamma}$ for every ξ, with $\mathbb{E}\Lambda(\xi)^2 \le \Lambda^2$, let $\phi := \mathbb{E}\Phi(\cdot,\xi)$, smoothed according to Theorem 3.2 (so that ϕ_γ is differentiable on $\mathbb{R}^d$), and $\hat\phi(x) := \frac{d}{2\gamma}[\Phi(x+\gamma u,\xi) - \Phi(x-\gamma u,\xi)]u$ with $u \sim U(\mathbb{S})$ independent of ξ. Then*

$$\mathbb{E}\hat\phi(x) = \nabla\phi_\gamma(x), \qquad \mathbb{E}\left\|\hat\phi(x)\right\|^2 \le \frac{d^2}{d-1}\Lambda^2 \le 2d\Lambda^2 \quad (d \ge 2), \qquad \mathbb{E}\left\|\hat\phi(x)\right\|^2 \le \Lambda^2 \quad (d = 1).$$

In particular $\mathbb{E}\left\|\hat\phi(x) - \nabla\phi_\gamma(x)\right\|^2 \le 2d\Lambda^2$ for every $d \ge 1$.

Proof. Unbiasedness: by Fubini, $\mathbb{E}\hat\phi(x) = \frac{d}{2\gamma}\mathbb{E}_u[(\phi(x+\gamma u) - \phi(x-\gamma u))u]$, where the integrability follows from $|\Phi(x+\gamma u,\xi) - \Phi(x-\gamma u,\xi)| \le 2\gamma\Lambda(\xi)$ (the points $x \pm \gamma u$ lie in $Q_{\bar\gamma}$). The points $x \pm \gamma u$ also lie in the set on which ϕ coincides with the function smoothed according to Theorem 3.2, so the right-hand side equals $\nabla\phi_\gamma(x)$ by Theorem 3.1(b), applied to the smooth, resp. globally Lipschitz, function that is actually smoothed. For the second moment fix ξ and put $q_\xi(u) := \frac{1}{2\gamma}[\Phi(x+\gamma u,\xi) - \Phi(x-\gamma u,\xi)]$. Then q_ξ is odd, so $\mathbb{E}_u q_\xi(u) = 0$, and it is $\Lambda(\xi)$-Lipschitz on $\mathbb{S}$ with respect to the Euclidean distance. For $d \ge 2$ the Poincaré inequality on the sphere (Theorem A.1) gives $\mathbb{E}_u q_\xi(u)^2 \le \Lambda(\xi)^2/(d-1)$, whence $\mathbb{E}\left\|\hat\phi(x)\right\|^2 = d^2\,\mathbb{E}_\xi\mathbb{E}_u q_\xi(u)^2 \le d^2\Lambda^2/(d-1) \le 2d\Lambda^2$. For $d = 1$, $|q_\xi(u)| \le \Lambda(\xi)$ and $\left\|\hat\phi\right\| = |q_\xi|$. Finally $\mathbb{E}\left\|\hat\phi - \mathbb{E}\hat\phi\right\|^2 \le \mathbb{E}\left\|\hat\phi\right\|^2$. □

Lemma 3.4 (Jacobian estimator and its action)**.** *Under Theorem 2.2, for fixed $x \in Q$, $y \in \mathbb{R}^m$ and $w \in \mathbb{R}^d$:*

(a) $\mathbb{E}\hat J(x) = J_\gamma(x)$, $\mathbb{E}\hat g_y(x) = \nabla f_\gamma(x) + J_\gamma(x)^\top y$ *and* $\mathbb{E}\left\|\hat g_y(x) - \nabla f_\gamma(x) - J_\gamma(x)^\top y\right\|^2 \le 2d\,(M_0 + \|y\|_1 M_g)^2$.

(b) $\|J_\gamma(x)\|_{2\to\infty} \le M_g$, *hence* $\|J_\gamma(x)w\|_\infty \le M_g\|w\|$.

(c) $\mathbb{E}\left\|\hat J(x)w\right\|_\infty^2 \le d\,M_g^2\|w\|^2$ *and* $\mathbb{E}\left\|(\hat J(x) - J_\gamma(x))w\right\|_\infty^2 \le 4d\,M_g^2\|w\|^2$.

(d) $\mathbb{E}\left\|\hat J(x)w\right\|_\infty^2 \le 45\,\kappa_m M_g^2\|w\|^2$ *and* $\mathbb{E}\left\|(\hat J(x) - J_\gamma(x))w\right\|_\infty^2 \le 92\,\kappa_m M_g^2\|w\|^2$.

We write $\nu_m := \min\{4d, 92\kappa_m\}$, so that $\mathbb{E}\left\|(\hat J(x) - J_\gamma(x))w\right\|_\infty^2 \le \nu_m M_g^2\|w\|^2$.

Proof. (a) The i-th row of $\hat J(x)$ is the two-point estimator of $G_i(\cdot,\xi)$, so $\mathbb{E}\hat J(x) = J_\gamma(x)$ by Theorem 3.3. The vector $\hat g_y(x) = \frac{d}{2\gamma}[\Phi(x+\gamma u,\xi) - \Phi(x-\gamma u,\xi)]u$ is the two-point estimator of

$\Phi(\cdot,\xi) := F(\cdot,\xi)+\langle y, G(\cdot,\xi)\rangle$, which is $(M_0(\xi)+\|y\|_1 M_g(\xi))$-Lipschitz; by Minkowski's inequality $\mathbb{E}(M_0(\xi)+\|y\|_1 M_g(\xi))^2 \le (M_0+\|y\|_1 M_g)^2$, and Theorem 3.3 applies. (b) $g_{i,\gamma}$ is M_g-Lipschitz by Theorem 3.1(a), so $\|\nabla g_{i,\gamma}(x)\| \le M_g$; use (4). (c) With $q_i(u) := \frac{1}{2\gamma}[G_i(x+\gamma u,\xi) - G_i(x-\gamma u,\xi)]$ we have $|q_i(u)| \le M_i(\xi) \le M_g(\xi)$ and $(\hat{J}(x)w)_i = d\, q_i(u)\,\langle u, w\rangle$, hence $\mathbb{E}\max_i(\hat{J}(x)w)_i^2 \le d^2\,\mathbb{E}[M_g(\xi)^2]\,\mathbb{E}\,\langle u,w\rangle^2 = dM_g^2\,\|w\|^2$ because u and ξ are independent and $\mathbb{E}\,\langle u,w\rangle^2 = \|w\|^2/d$. Then $\mathbb{E}\left\|(\hat{J}-J_\gamma)w\right\|_\infty^2 \le 2\mathbb{E}\left\|\hat{J}w\right\|_\infty^2 + 2\,\|J_\gamma w\|_\infty^2 \le (2d+2)M_g^2\,\|w\|^2 \le 4dM_g^2\,\|w\|^2$ by (b). Part (d) is proved in Section B. □

Lemma 3.5 (level estimator). *Under Theorems 2.2 and 2.3, for $x \in Q$ and $\gamma \le \bar{\gamma}$, $\mathbb{E}\hat{h}(x) = g_\gamma(x) - b =: h(x)$ and*

$$\mathbb{E}\left\|\hat{h}(x) - h(x)\right\|_\infty^2 \le \sigma_h^2 := 2\sigma_g^2 + 8\gamma^2 M_g^2. \tag{10}$$

Proof. $\mathbb{E}_\xi G(x+\gamma v,\xi) = g(x+\gamma v)$ and $\mathbb{E}_v g(x+\gamma v) = g_\gamma(x)$ give unbiasedness. Moreover $\hat{h}(x) - h(x) = [G(x+\gamma v,\xi) - g(x+\gamma v)] + [g(x+\gamma v) - g_\gamma(x)]$; the first bracket has $\mathbb{E}\,\|\cdot\|_\infty^2 \le \sigma_g^2$ because $x+\gamma v \in Q_{\bar{\gamma}}$, and $|g_i(x+\gamma v) - g_{i,\gamma}(x)| \le |g_i(x+\gamma v) - g_i(x)| + |g_i(x) - g_{i,\gamma}(x)| \le 2\gamma M_g$ by Theorem 3.1(a). Use $\|a+c\|_\infty^2 \le 2\,\|a\|_\infty^2 + 2\,\|c\|_\infty^2$. □

3.3 Mini-batches in the ℓ_∞-norm

Averaging b independent copies of a centered random vector divides its *Euclidean* second moment by b. In the ℓ_∞-norm, which is the norm relevant for the dual (entropy) geometry, the same is true up to a logarithmic factor in m, but only after the batch size exceeds that factor.

Lemma 3.6 (ℓ_∞ mini-batch lemma). *Let $Z_1,\dots,Z_b \in \mathbb{R}^m$ be independent with $\mathbb{E}Z_j = 0$ and $\mathbb{E}\,\|Z_j\|_\infty^2 \le \mathsf{H}$. Then*

$$\mathbb{E}\Big\|\frac{1}{b}\sum_{j=1}^b Z_j\Big\|_\infty^2 \le \min\Big\{1, \frac{8\kappa_m}{b}\Big\}\mathsf{H}, \qquad \kappa_m = 1+\log(2m).$$

The same holds conditionally on a σ-field $\mathcal{G}$ if the Z_j are independent, centered and bounded as above conditionally on $\mathcal{G}$.

The proof (Section B) uses symmetrization and the sub-Gaussian maximal inequality; the factor $\min\{1,\cdot\}$ is Jensen's inequality. For $b \le 8\kappa_m$ the bound of the lemma is the trivial one, H: the worst-case analysis does not certify any reduction of the ℓ_∞ noise before the batch size reaches the logarithmic scale (for particular distributions the actual variance may of course decrease earlier). This is why the dual noise appears with the factor κ_m in all bounds below and why the parameter choices never rely on dual mini-batches smaller than $8\kappa_m$.

3.4 Suprema of martingale transforms over the dual set

The certificate of Section 4 requires a bound that holds uniformly over the dual variable y. Stochastic errors enter the analysis in the form $\sum_k a_k\,\langle e_k, y - \zeta_k\rangle$ with predictable ζ_k; the supremum over y of such a sum is controlled by the following classical device [32, 46], which we state for a general Bregman setup.

Lemma 3.7 (martingale supremum). *Let $Z \subseteq Z'$ be convex sets in a finite-dimensional normed space $(\mathbb{E}, \|\cdot\|_Z)$ with dual norm $\|\cdot\|_{Z,*}$, Z compact, let ω be a distance-generating function on Z' (differentiable on Z', or on its relative interior, and 1-strongly convex with respect to $\|\cdot\|_Z$ on Z'), $V(z,z') := \omega(z') - \omega(z) - \langle\nabla\omega(z), z'-z\rangle$ the associated Bregman distance, $z_0 \in Z'$ a point at which ω is differentiable (not necessarily in Z) and $\Theta \ge \sup_{z\in Z} V(z_0,z)$. Let $(\mathcal{F}_k)_{k\ge0}$ be a filtration, e_k an $\mathcal{F}_k$-measurable random vector with $\mathbb{E}[e_k \mid \mathcal{F}_{k-1}] = 0$ and $\mathbb{E}\,\|e_k\|_{Z,*}^2 \le v_k^2$, ζ_k an*

$\mathcal{F}_{k-1}$-measurable bounded random vector in $\mathbb{E}$ (not necessarily in Z) and $a_k > 0$ deterministic, $k = 1, \dots, K$. Then for every $\lambda > 0$

$$\mathbb{E} \sup_{z \in Z} \sum_{k=1}^{K} a_k \langle e_k, z - \zeta_k \rangle \le \lambda \Theta + \frac{1}{2\lambda} \sum_{k=1}^{K} a_k^2 v_k^2,$$

and consequently, optimizing over λ,

$$\mathbb{E} \sup_{z \in Z} \sum_{k=1}^{K} a_k \langle e_k, z - \zeta_k \rangle \le \left(2\Theta \sum_{k=1}^{K} a_k^2 v_k^2 \right)^{1/2}.$$

Proof. Define the auxiliary sequence $w_0 := z_0$, $w_k := \arg\min_{w \in Z} \{ a_k \langle -e_k, w \rangle + \lambda V(w_{k-1}, w) \}$; the minimizers exist since Z is compact and lie in $Z \subseteq Z'$, so that all Bregman distances below are defined ($w_0 = z_0$ may lie outside Z, which is why the lemma is stated with the ambient set Z'; Theorem A.6 covers this case). By the three-point inequality (Theorem A.6), for every $z \in Z$: $a_k \langle -e_k, w_k - z \rangle \le \lambda [V(w_{k-1}, z) - V(w_k, z) - V(w_{k-1}, w_k)]$. Hence

$$\begin{aligned} a_k \langle e_k, z - w_{k-1} \rangle &= a_k \langle e_k, z - w_k \rangle + a_k \langle e_k, w_k - w_{k-1} \rangle \\ &\le \lambda [V(w_{k-1}, z) - V(w_k, z)] - \lambda V(w_{k-1}, w_k) + a_k \|e_k\|_{Z,*} \|w_k - w_{k-1}\|_Z . \end{aligned}$$

By Young's inequality and $V(w_{k-1}, w_k) \ge \frac{1}{2} \|w_k - w_{k-1}\|_Z^2$, $a_k \|e_k\|_{Z,*} \|w_k - w_{k-1}\|_Z \le \frac{a_k^2}{2\lambda} \|e_k\|_{Z,*}^2 + \lambda V(w_{k-1}, w_k)$. Summing over k and telescoping,

$$\sum_k a_k \langle e_k, z - w_{k-1} \rangle \le \lambda V(w_0, z) + \frac{1}{2\lambda} \sum_k a_k^2 \|e_k\|_{Z,*}^2 \le \lambda \Theta + \frac{1}{2\lambda} \sum_k a_k^2 \|e_k\|_{Z,*}^2$$

for all $z \in Z$ simultaneously. Finally

$$\sum_k a_k \langle e_k, z - \zeta_k \rangle = \sum_k a_k \langle e_k, z - w_{k-1} \rangle + \sum_k a_k \langle e_k, w_{k-1} - \zeta_k \rangle ,$$

where w_{k-1} and ζ_k are $\mathcal{F}_{k-1}$-measurable and bounded, so the last sum has zero expectation. Taking the supremum over z in the first sum and then expectations proves the first bound; the second follows with $\lambda = \left(\sum_k a_k^2 v_k^2 / (2\Theta) \right)^{1/2}$. □

4 The smoothed Lagrangian and the certificate

4.1 The smoothed and shifted problem

Fix $\gamma \in (0, \bar{\gamma}]$, and let f_γ, $g_{i,\gamma}$ be the smoothed functions of Theorem 3.2. By Theorem 3.1(a) (with $C = Q$) the smoothed constraints satisfy $g_i \le g_{i,\gamma} \le g_i + b_i$ on Q with the *known* bias bounds

$$b_i := \begin{cases} \gamma M_g, & \text{in general,} \\ L_g \gamma^2 / 2, & \text{if } g_i \text{ is } L_g\text{-smooth,} \\ 0, & \text{if } g_i \text{ is affine,} \end{cases} \qquad b_0 := \begin{cases} \gamma M_0, & \text{in general,} \\ L_0 \gamma^2 / 2, & \text{if } f \text{ is } L_0\text{-smooth.} \end{cases} \tag{11}$$

We consider the *smoothed shifted problem*

$$\min_{x \in Q} \ f_\gamma(x) \qquad \text{s.t.} \qquad h_i(x) := g_{i,\gamma}(x) - b_i \le 0, \quad i = 1, \dots, m, \tag{12}$$

whose data are smooth on all of $\mathbb{R}^d$: writing $L_{0,\gamma}$ and $L_{g,\gamma}$ for the smoothness constants of f_γ and of the $g_{i,\gamma}$ given by Theorem 3.1(c) and Theorem 3.2 ($L_{0,\gamma} = L_0$ if f is L_0-smooth on $\mathbb{R}^d$

and $L_{0,\gamma} = \sqrt{d}M_0/\gamma$ otherwise; $L_{g,\gamma}$ is the largest of the corresponding constants of the $g_{i,\gamma}$, i.e. L_g if all constraints are L_g-smooth, 0 if they are affine, and $\sqrt{d}M_g/\gamma$ if some of them are merely Lipschitz), we put

$$H := L_{0,\gamma} + RL_{g,\gamma}, \qquad L \geq H \quad \text{(any upper bound used by an algorithm)}, \tag{13}$$

where R is the multiplier bound (14) below. We denote by f_γ^* the optimal value of (12), by x_γ^* an optimal solution and by $b_g := \max_i b_i$.

Lemma 4.1 (transfer between the original and the smoothed problem)**.** *Under Theorems 2.1 and 2.2:*

(a) every feasible point of (1) is feasible for (12), and $f_\gamma^* \leq f^* + b_0$;

(b) Theorem 2.5 for (1) implies $h_i(x^{\mathrm{s}}) \leq -\rho$ for all i, i.e. the Slater condition for (12) with the same point and margin;

(c) if $\hat{x} \in Q$ satisfies $f_\gamma(\hat{x}) - f_\gamma^* \leq \varepsilon'$ and $\|[h(\hat{x})]_+\|_\infty \leq \varepsilon'$, then $f(\hat{x}) - f^* \leq \varepsilon' + b_0$ and $\mathsf{v}(\hat{x}) = \|[g(\hat{x})]_+\|_\infty \leq \varepsilon' + b_g$.

Proof. (a) If $g(x) \leq 0$ then $h_i(x) = g_{i,\gamma}(x) - b_i \leq g_i(x) \leq 0$. Hence the feasible set of (12) contains that of (1) and $f_\gamma^* \leq \min\{f_\gamma(x) : g(x) \leq 0, x \in Q\} \leq \min\{f(x) + b_0 : g(x) \leq 0, x \in Q\} = f^* + b_0$ by Theorem 3.1(a). (b) $h_i(x^{\mathrm{s}}) \leq g_i(x^{\mathrm{s}}) \leq -\rho$. (c) $f(\hat{x}) - f^* \leq f_\gamma(\hat{x}) - f^* \leq (f_\gamma(\hat{x}) - f_\gamma^*) + (f_\gamma^* - f^*) \leq \varepsilon' + b_0$ by (a). Moreover $g_i(\hat{x}) \leq g_{i,\gamma}(\hat{x}) = h_i(\hat{x}) + b_i \leq [h_i(\hat{x})]_+ + b_g$. □

4.2 Lagrange multipliers

Lemma 4.2 (multipliers)**.** *Under Theorems 2.1, 2.2 and 2.5, problem (12) has an optimal solution $x_\gamma^* \in Q$ and a vector $y_\gamma^* \in \mathbb{R}_+^m$ such that*

$$x_\gamma^* \in \arg\min_{x \in Q} \big\{ f_\gamma(x) + \langle y_\gamma^*, h(x) \rangle \big\}, \qquad \langle y_\gamma^*, h(x_\gamma^*) \rangle = 0, \qquad \|y_\gamma^*\|_1 \leq \frac{f_\gamma(x^{\mathrm{s}}) - f_\gamma^*}{\rho} \leq \frac{M_0 D}{\rho}.$$

Proof. f_γ and h_i are finite convex functions on $\mathbb{R}^d$ (Theorem 3.1(a) and Theorem 3.2), Q is compact and the Slater condition holds by Theorem 4.1(b); the existence of x_γ^* follows from continuity and compactness, the existence of Kuhn–Tucker multipliers with the saddle-point property and complementary slackness is the classical Kuhn–Tucker theorem for convex programs under Slater's condition (e.g. 55, Thm. 28.2 and Cor. 28.3.1). For the bound, the saddle-point property gives $f_\gamma^* = \min_{x \in Q}\{f_\gamma(x) + \langle y_\gamma^*, h(x) \rangle\} \leq f_\gamma(x^{\mathrm{s}}) + \langle y_\gamma^*, h(x^{\mathrm{s}}) \rangle \leq f_\gamma(x^{\mathrm{s}}) - \rho \|y_\gamma^*\|_1$, and $f_\gamma(x^{\mathrm{s}}) - f_\gamma^* = f_\gamma(x^{\mathrm{s}}) - f_\gamma(x_\gamma^*) \leq M_0 \|x^{\mathrm{s}} - x_\gamma^*\| \leq M_0 D$ since f_γ is M_0-Lipschitz on Q (Theorem 3.1(a) with $C = Q$). □

Throughout we fix a constant

$$R \;\geq\; \|y_\gamma^*\|_1 + 1, \qquad \text{for instance} \qquad R := 1 + \frac{M_0 D}{\rho}, \tag{14}$$

and we write $M_R := M_0 + RM_g$. The dual variable ranges over the truncated orthant and its lifting to a simplex,

$$Y_R := \{y \in \mathbb{R}_+^m : \|y\|_1 \leq R\}, \qquad \tilde{Y}_R := \Big\{\tilde{y} = (\tilde{y}_0, y) \in \mathbb{R}_+^{m+1} : \; \tilde{y}_0 + \textstyle\sum_{i=1}^m y_i = R\Big\}, \tag{15}$$

where the map $\tilde{y} \mapsto y$ (dropping the slack coordinate $\tilde{y}_0$) is a bijection from $\tilde{Y}_R$ onto Y_R.

4.3 Lagrangian, restricted gap and certificate

The smoothed Lagrangian and the certificate are

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

$$\mathsf{C}_{R,\gamma}(x) := \sup_{y\in Y_R} \Phi_\gamma(x,y) - f_\gamma^* = f_\gamma(x) - f_\gamma^* + R\,\|[h(x)]_+\|_\infty\,, \qquad x\in Q. \tag{17}$$

The identity in (17) holds because $\sup_{y\in Y_R}\langle y,h\rangle = R\max\{0,\max_i h_i\} = R\,\|[h]_+\|_\infty$. For $y\in Y_R$ the function $\Phi_\gamma(\cdot,y)$ is convex and L-smooth on $\mathbb{R}^d$ with L from (13): indeed $\nabla_x\Phi_\gamma(x,y) = \nabla f_\gamma(x) + J_\gamma(x)^\top y$ and $\|(J_\gamma(x)-J_\gamma(x'))^\top y\| \le \sum_i y_i\,\|\nabla g_{i,\gamma}(x) - \nabla g_{i,\gamma}(x')\| \le RL_{g,\gamma}\,\|x-x'\|$.

Lemma 4.3 (certificate)**.** *Under Theorems 2.1, 2.2 and 2.5 and (14), for every $x\in Q$,*

$$\mathsf{C}_{R,\gamma}(x) \;\ge\; \max\bigl\{f_\gamma(x) - f_\gamma^*,\; \|[h(x)]_+\|_\infty\bigr\} \;\ge\; 0, \tag{18}$$

and for every $\hat y\in Y_R$

$$\Phi_\gamma(x_\gamma^*,\hat y) \le f_\gamma^*, \qquad \textit{hence} \qquad \mathsf{C}_{R,\gamma}(x) \le \sup_{y\in Y_R}\Phi_\gamma(x,y) - \Phi_\gamma(x_\gamma^*,\hat y). \tag{19}$$

If in addition Theorem 2.6 holds, then $\frac{\mu}{2}\|x-x_\gamma^*\|^2 \le \mathsf{C}_{R,\gamma}(x)$ for all $x\in Q$.

Proof. By the saddle-point property of Theorem 4.2, for all $x\in Q$, $f_\gamma(x) + \langle y_\gamma^*, h(x)\rangle \ge f_\gamma(x_\gamma^*) + \langle y_\gamma^*, h(x_\gamma^*)\rangle = f_\gamma^*$, hence $f_\gamma(x) - f_\gamma^* \ge -\langle y_\gamma^*, h(x)\rangle \ge -\langle y_\gamma^*, [h(x)]_+\rangle \ge -\|y_\gamma^*\|_1\,\|[h(x)]_+\|_\infty \ge -(R-1)\,\|[h(x)]_+\|_\infty$. Adding $R\,\|[h(x)]_+\|_\infty$ yields $\mathsf{C}_{R,\gamma}(x) \ge \|[h(x)]_+\|_\infty \ge 0$; and $\mathsf{C}_{R,\gamma}(x) \ge f_\gamma(x) - f_\gamma^*$ is obvious. For (19), $h(x_\gamma^*)\le 0$ and $\hat y\ge 0$ give $\Phi_\gamma(x_\gamma^*,\hat y) \le f_\gamma(x_\gamma^*) = f_\gamma^*$. Under Theorem 2.6, f_γ is μ-strongly convex on Q (Theorem 3.1(d) with $C=Q$, $C_\gamma\subseteq Q_{\bar\gamma}$; the function smoothed under Theorem 3.2 coincides with f on $Q_{\bar\gamma}$), so $\Phi_\gamma(\cdot,y_\gamma^*)$ is μ-strongly convex on Q and minimized over Q at x_γ^*; thus $\Phi_\gamma(x,y_\gamma^*) - f_\gamma^* \ge \frac{\mu}{2}\|x-x_\gamma^*\|^2$ (strong convexity plus the first-order optimality condition on Q), and $\Phi_\gamma(x,y_\gamma^*) \le \sup_{y\in Y_R}\Phi_\gamma(x,y)$ because $y_\gamma^*\in Y_R$. □

Theorems 4.1 and 4.3 reduce the task to bounding $\mathbb{E}\sup_{y\in Y_R}\Phi_\gamma(\hat x,y) - \Phi_\gamma(x_\gamma^*,\hat y)$ for the output $\hat x$ and an arbitrary dual output $\hat y$ of a method: an expected bound ε' on this *restricted duality gap* gives an $(\varepsilon'+b_0,\varepsilon'+b_g)$-solution of the original problem in expectation. Note that the supremum is over y only; the primal comparator is the fixed point x_γ^*. This asymmetry is exploited in the stochastic analysis: primal errors need only be centered, dual errors must be controlled uniformly (Theorem 3.7).

4.4 Proximal geometry

The primal space carries the Euclidean distance and the dual space the entropy on the lifted simplex $\tilde Y_R$:

$$V_{\mathcal{X}}(x,x') := \tfrac12\|x-x'\|^2\,, \qquad V_{\mathcal{Y}}(\tilde y,\tilde y') := R\sum_{i=0}^{m}\tilde y_i'\log\frac{\tilde y_i'}{\tilde y_i} \quad (\tilde y\in \operatorname{ri}\tilde Y_R,\ \tilde y'\in\tilde Y_R). \tag{20}$$

Both are Bregman distances of distance-generating functions ($\frac12\|x\|^2$ and $R\sum_i \tilde y_i\log\tilde y_i$). We use the uniform dual center $\tilde y^{\mathrm u} := \frac{R}{m+1}(1,\dots,1)$ and the radii

$$\mathcal{A} := V_{\mathcal{X}}(x_0,x_\gamma^*) = \tfrac12\|x_0-x_\gamma^*\|^2 \le \tfrac12 D^2, \qquad \mathcal{B} := \max_{\tilde y\in\tilde Y_R} V_{\mathcal{Y}}(\tilde y^{\mathrm u},\tilde y) = R^2\log(m+1). \tag{21}$$

Lemma 4.4 (facts about the proximal steps)**.** *(a) (Pinsker) For all $\tilde y\in\operatorname{ri}\tilde Y_R$, $\tilde y'\in\tilde Y_R$ with last m coordinates y,y':*

$$V_{\mathcal{Y}}(\tilde y,\tilde y') \ge \tfrac12\|\tilde y'-\tilde y\|_1^2 \ge \tfrac12\|y'-y\|_1^2\,.$$

(b) (radius) $V_{\mathcal{Y}}(\tilde{y}^{\mathrm{u}}, \tilde{y}) \le R^2 \log(m+1) = \mathcal{B}$ *for all* $\tilde{y} \in \tilde{Y}_R$.

(c) (dual step) For $\tilde{y} \in \operatorname{ri} \tilde{Y}_R$, $q \in \mathbb{R}^m$, $\tau > 0$ *and* $\tilde{q} := (0, q) \in \mathbb{R}^{m+1}$, *the point* $\tilde{y}^+ := \arg\max_{\tilde{y}' \in \tilde{Y}_R} \{\langle \tilde{q}, \tilde{y}' \rangle - \tau V_{\mathcal{Y}}(\tilde{y}, \tilde{y}')\}$ *is given by* $\tilde{y}_i^+ = R\, \tilde{y}_i e^{\tilde{q}_i/(\tau R)} / \sum_{l=0}^{m} \tilde{y}_l e^{\tilde{q}_l/(\tau R)}$, *lies in* $\operatorname{ri} \tilde{Y}_R$, *and satisfies for all* $\tilde{y}' \in \tilde{Y}_R$

$$\langle q, y' - y^+ \rangle = \langle \tilde{q}, \tilde{y}' - \tilde{y}^+ \rangle \le \tau \big[V_{\mathcal{Y}}(\tilde{y}, \tilde{y}') - V_{\mathcal{Y}}(\tilde{y}^+, \tilde{y}') - V_{\mathcal{Y}}(\tilde{y}, \tilde{y}^+) \big].$$

(d) (primal step) For $x_{\mathrm{c}}, p \in Q$, $r \in \mathbb{R}^d$ *and* $\eta, \beta \ge 0$ *with* $\eta + \beta > 0$, *the point* $p^+ := \arg\min_{x \in Q} \{\langle r, x \rangle + \eta V_{\mathcal{X}}(x_{\mathrm{c}}, x) + \beta V_{\mathcal{X}}(p, x)\}$ *is the Euclidean projection onto* Q *of* $(\eta x_{\mathrm{c}} + \beta p - r)/(\eta + \beta)$ *and satisfies for all* $x \in Q$

$$\langle r, p^+ - x \rangle \le \eta \big[V_{\mathcal{X}}(x_{\mathrm{c}}, x) - V_{\mathcal{X}}(x_{\mathrm{c}}, p^+) - V_{\mathcal{X}}(p^+, x) \big] + \beta \big[V_{\mathcal{X}}(p, x) - V_{\mathcal{X}}(p, p^+) - V_{\mathcal{X}}(p^+, x) \big].$$

Proof. (a) With $p := \tilde{y}'/R$, $q := \tilde{y}/R$ probability vectors, $V_{\mathcal{Y}}(\tilde{y}, \tilde{y}') = R^2 \operatorname{KL}(p \| q) \ge \frac{R^2}{2} \|p - q\|_1^2 = \frac{1}{2} \|\tilde{y}' - \tilde{y}\|_1^2$ by Pinsker's inequality (Theorem A.4); dropping the coordinate 0 does not increase the ℓ_1-norm. (b) $\operatorname{KL}(p \| \text{uniform}) = \log(m+1) - \operatorname{Ent}(p) \le \log(m+1)$. (c) The objective $\tilde{y}' \mapsto \langle \tilde{q}, \tilde{y}' \rangle - \tau V_{\mathcal{Y}}(\tilde{y}, \tilde{y}')$ is strictly concave on the simplex and tends to $-\infty$ in no direction; the Lagrange conditions $\tilde{q}_i - \tau R(\log \tilde{y}_i' - \log \tilde{y}_i + 1) = \nu$ give the stated formula, which has positive coordinates. The inequality is the three-point inequality (Theorem A.6) for the concave maximization problem with the Bregman distance $V_{\mathcal{Y}}$. (d) The objective equals $\frac{\eta+\beta}{2} \|x - c\|^2 + \text{const}$ with $c = (\eta x_{\mathrm{c}} + \beta p - r)/(\eta + \beta)$, so p^+ is the projection of c. The inequality follows from the optimality condition $\langle r + \eta(p^+ - x_{\mathrm{c}}) + \beta(p^+ - p), x - p^+ \rangle \ge 0$ and the identity $2\langle p^+ - c', x - p^+ \rangle = \|x - c'\|^2 - \|p^+ - c'\|^2 - \|x - p^+\|^2$ applied with $c' = x_{\mathrm{c}}$ and $c' = p$. □

5 Baseline: batched stochastic Mirror–Prox

The saddle-point problem $\min_{x \in Q} \max_{y \in Y_R} \Phi_\gamma(x, y)$ has a convex–concave objective whose partial gradients are Lipschitz. The natural baseline is therefore the stochastic Mirror–Prox (extragradient) method of Juditsky et al. [32], Nemirovski [45] applied to the monotone operator

$$\mathcal{G}(z) := \big(\nabla_x \Phi_\gamma(x, y),\, -\nabla_y \Phi_\gamma(x, y) \big) = \big(\nabla f_\gamma(x) + J_\gamma(x)^\top y,\, -h(x) \big), \qquad z = (x, y) \in \mathcal{Z} := Q \times Y_R, \tag{22}$$

with the estimator $\hat{\mathcal{G}}(z) = (\hat{g}_y(x), -\hat{h}(x))$ of (9), averaged over mini-batches. We use the joint geometry

$$\begin{aligned} V(z, z') &:= \alpha V_{\mathcal{X}}(x, x') + \alpha^{-1} V_{\mathcal{Y}}(\tilde{y}, \tilde{y}'), \\ \|z\|_\alpha^2 := \alpha \|x\|^2 + \alpha^{-1} \|\tilde{y}\|_1^2, &\qquad \|(p, q)\|_{\alpha,*}^2 := \alpha^{-1} \|p\|^2 + \alpha \|q\|_\infty^2, \end{aligned} \tag{23}$$

where $\alpha > 0$ balances the two blocks; V is 1-strongly convex with respect to $\|\cdot\|_\alpha$ by Theorem 4.4(a), and $\|\cdot\|_{\alpha,*}$ is the dual norm. The constants that enter the analysis are

$$L_\alpha := \frac{1}{2}\Big(\frac{H}{\alpha} + \sqrt{\frac{H^2}{\alpha^2} + 4M_g^2} \Big) \le \frac{H}{\alpha} + M_g, \quad \sigma_*^2 := \frac{1}{b}\Big(\frac{2dM_R^2}{\alpha} + \alpha \min\{b, 8\kappa_m\} \sigma_h^2 \Big), \quad \mathcal{D}_\alpha^2 := \alpha \mathcal{A} + \frac{\mathcal{B}}{\alpha}. \tag{24}$$

Each iteration of Algorithm 1 consists of two sequential rounds and issues $6b$ vector calls ($2b$ paired and b single queries in each round); every call is used for the constraint vector, and the $4b$ calls of the paired queries are also used for the objective. Hence

$$N_{\mathrm{seq}} = 2N, \qquad N_{\mathrm{orc}} = N_g = 6bN, \qquad N_f = 4bN, \qquad N_{\mathrm{inner}} = N_{\mathrm{seq}}.$$

Theorem 5.1 (batched stochastic Mirror–Prox)**.** *Let Theorems 2.1 to 2.3 and 2.5 hold,* $\gamma \le \bar{\gamma}$, R *as in* (14), H *as in* (13), *and* $\eta \le 1/(\sqrt{3} L_\alpha)$. *Then the output of Algorithm 1 satisfies*

$$\mathbb{E}\, \mathsf{C}_{R,\gamma}(\bar{x}_N) \le \frac{\mathcal{D}_\alpha^2}{\eta N} + 3\eta \sigma_*^2 + \sigma_* \mathcal{D}_\alpha \sqrt{\frac{2}{N}}. \tag{25}$$

Algorithm 1 B-SMP: batched stochastic Mirror–Prox for the smoothed Lagrangian

Require: $N \geq 1$, mini-batch $b \geq 1$, balance $\alpha > 0$, step $\eta \in (0, 1/(\sqrt{3}L_\alpha)]$, radius γ, shifts b_i, bound R; $x_1 \in Q$, $\tilde{y}_1 := \tilde{y}^{\mathrm{u}}$.

1: **for** $k = 1, \dots, N$ **do**

2: **Round** $2k-1$**:** at x_k issue b paired queries and b single queries; form $\hat{g}_k := \frac{1}{b}\sum_j \hat{g}_{y_k,j}(x_k)$ and $\hat{h}_k := \frac{1}{b}\sum_j \hat{h}_j(x_k)$.

3: $x_k^w := \Pi_Q\big(x_k - \frac{\eta}{\alpha}\hat{g}_k\big)$; $\tilde{y}_k^w :=$ dual step of Theorem 4.4(c) from $\tilde{y}_k$ with $q = \eta\hat{h}_k$ and $\tau = 1/\alpha$.

4: **Round** $2k$**:** at x_k^w issue b paired and b single queries; form $\hat{g}'_k := \frac{1}{b}\sum_j \hat{g}_{y_k^w,j}(x_k^w)$ and $\hat{h}'_k := \frac{1}{b}\sum_j \hat{h}_j(x_k^w)$.

5: $x_{k+1} := \Pi_Q\big(x_k - \frac{\eta}{\alpha}\hat{g}'_k\big)$; $\tilde{y}_{k+1} :=$ dual step from $\tilde{y}_k$ with $q = \eta\hat{h}'_k$ and $\tau = 1/\alpha$.

6: **end for**

Ensure: $\bar{x}_N := \frac{1}{N}\sum_{k=1}^N x_k^w$ (and $\bar{y}_N := \frac{1}{N}\sum_k y_k^w$).

With $\eta := \min\{1/(\sqrt{3}L_\alpha),\ \mathcal{D}_\alpha/(\sigma_*\sqrt{3N})\}$,

$$\mathbb{E}\,\mathsf{C}_{R,\gamma}(\bar{x}_N) \leq \frac{\sqrt{3}L_\alpha\mathcal{D}_\alpha^2}{N} + \frac{5\sigma_*\mathcal{D}_\alpha}{\sqrt{N}}. \tag{26}$$

The proof is in Section C. It is the classical analysis of Juditsky et al. [32], written for the restricted gap of Theorem 4.3: the stochastic term $\sigma_*\mathcal{D}_\alpha\sqrt{2/N}$ comes from Theorem 3.7 applied to the dual block, and $3\eta\sigma_*^2$ from the variance of the two mini-batches of an iteration. The constant L_α is the spectral norm of the 2×2 matrix $\left(\begin{smallmatrix} H/\alpha & M_g \\ M_g & 0 \end{smallmatrix}\right)$ that dominates the block Lipschitz structure of $\mathcal{G}$.

Corollary 5.2 (complexity of B-SMP)**.** *In the setting of Theorem 5.1, let* $\varepsilon_{\mathrm{s}} > 0$*, choose* $\alpha := \sqrt{2\mathcal{B}}/D$*, the step* η *of* (26)*,*

$$N := \left\lceil \frac{2\sqrt{3}\,\big(HD^2 + M_g D\sqrt{2\mathcal{B}}\big)}{\varepsilon_{\mathrm{s}}} \right\rceil, \qquad b := \max\left\{1,\ \left\lceil \frac{200\,(dM_R^2D^2 + 8\kappa_m\mathcal{B}\sigma_h^2)}{N\varepsilon_{\mathrm{s}}^2} \right\rceil \right\}. \tag{27}$$

Then $\mathbb{E}\mathsf{C}_{R,\gamma}(\bar{x}_N) \leq \varepsilon_{\mathrm{s}}$*, and*

$$N_{\mathrm{seq}} = 2N, \qquad N_f = 4bN \leq 4N + \frac{800\,(dM_R^2D^2 + 8\kappa_m\mathcal{B}\sigma_h^2)}{\varepsilon_{\mathrm{s}}^2},$$
$$N_{\mathrm{orc}} = N_g = 6bN \leq 6N + \frac{1200\,(dM_R^2D^2 + 8\kappa_m\mathcal{B}\sigma_h^2)}{\varepsilon_{\mathrm{s}}^2}.$$

Proof. With $\alpha = \sqrt{2\mathcal{B}}/D$ and $\mathcal{A} \leq D^2/2$ we get $\mathcal{D}_\alpha^2 \leq \sqrt{2\mathcal{B}}\,D$, $L_\alpha\mathcal{D}_\alpha^2 \leq (H/\alpha + M_g)\mathcal{D}_\alpha^2 \leq HD^2 + M_gD\sqrt{2\mathcal{B}}$ and $\sigma_*^2\mathcal{D}_\alpha^2 \leq \frac{1}{b}\big(2dM_R^2D^2 + 2\mathcal{B}\min\{b, 8\kappa_m\}\sigma_h^2\big) \leq \frac{2}{b}\,(dM_R^2D^2 + 8\kappa_m\mathcal{B}\sigma_h^2)$. By (26) the two terms are at most $\varepsilon_{\mathrm{s}}/2$ each: the first because $N \geq 2\sqrt{3}(HD^2 + M_gD\sqrt{2\mathcal{B}})/\varepsilon_{\mathrm{s}}$, the second because $25\sigma_*^2\mathcal{D}_\alpha^2/N \leq 50(dM_R^2D^2 + 8\kappa_m\mathcal{B}\sigma_h^2)/(bN) \leq \varepsilon_{\mathrm{s}}^2/4$. The counts follow from $b \leq 1 + 200(\cdots)/(N\varepsilon_{\mathrm{s}}^2)$. □

Corollary 5.3 (B-SMP for smooth and for nonsmooth data)**.** *Let* $\varepsilon > 0$ *and put* $\varepsilon_{\mathrm{s}} := \varepsilon/2$.

(a) (smooth data) Under Theorem 2.4, take $\gamma \leq \min\{\bar{\gamma}, \sqrt{\varepsilon/(2\max\{L_0, L_g\})}, \varepsilon/(4M_g)\}$*,* $b_0 = L_0\gamma^2/2$*,* $b_i = L_g\gamma^2/2$ *and* $H = L_0 + RL_g$*. Then the output is an* ε*-solution in expectation with* $N_{\mathrm{seq}} = O\big((L_0{+}RL_g)D^2\varepsilon^{-1} + M_gRD\sqrt{\log(m+1)}\,\varepsilon^{-1}\big)$ *and* $N_{\mathrm{orc}} = O\big(N_{\mathrm{seq}} + (dM_R^2D^2 + \Lambda_mR^2\sigma_g^2)\varepsilon^{-2}\big)$*.*

(b) (nonsmooth data) Take $\gamma := \varepsilon/(4M)$ *(assuming* $\varepsilon \le 4M\bar{\gamma}$*),* $b_0 = \gamma M_0$*,* $b_i = \gamma M_g$ *and* $H = \sqrt{d}M_R/\gamma = 4\sqrt{d}M_R M/\varepsilon$*. Then the output is an* ε*-solution in expectation with* $N_{\mathrm{seq}} = O\big(\sqrt{d}M_R M D^2\varepsilon^{-2} + M_g R D\sqrt{\log(m+1)}\,\varepsilon^{-1}\big)$ *and* $N_{\mathrm{orc}} = O\big(N_{\mathrm{seq}} + (dM_R^2 D^2 + \Lambda_m R^2\sigma_g^2)\varepsilon^{-2}\big)$*.*

Proof. In both cases $b_0, b_g \le \varepsilon/4$, so by Theorem 4.1(c) and Theorem 4.3 the output with $\mathbb{E}\mathsf{C}_{R,\gamma} \le \varepsilon/2$ satisfies $\mathbb{E}[f(\bar{x}_N) - f^*] \le \varepsilon/2 + \varepsilon/4$ and $\mathbb{E}\mathsf{v}(\bar{x}_N) \le \varepsilon/2 + \varepsilon/4$. In both cases $\gamma \le \varepsilon/(4M_g)$, so $8\gamma^2 M_g^2 \le \varepsilon^2/2$ and $\sigma_h^2 \le 2\sigma_g^2 + \varepsilon^2/2$ by (10); the contribution of $\varepsilon^2/2$ to N_{orc} is $O(\Lambda_m R^2)$, a constant. Insert H, $\mathcal{B} = R^2\log(m+1)$ and $\varepsilon_{\mathrm{s}} = \varepsilon/2$ into Theorem 5.2. □

Remark 5.4 (what the baseline shows). The number of rounds of B-SMP is proportional to H/ε, where $H = L_{0,\gamma} + RL_{g,\gamma}$ is the smoothness of the Lagrangian. For nonsmooth data $H \propto \sqrt{d}/\varepsilon$, so the rounds scale as $\sqrt{d}/\varepsilon^2$: better than the d/ε^2 rounds of plain stochastic mirror descent with two-point estimates (which cannot be parallelized, since its rate $MD/\sqrt{N}$ does not improve with the batch size), but far from the $d^{1/4}/\varepsilon$ of the accelerated unconstrained methods of Gasnikov et al. [26]. The next section closes this gap.

6 Zeroth-order primal–dual sliding

6.1 Idea

For smooth data the round complexity $O(HD^2/\varepsilon)$ of Mirror–Prox is not optimal: for the *unconstrained* problem accelerated methods with mini-batches need only $O(\sqrt{L_0 D^2/\varepsilon})$ rounds. The obstacle is the coupling term $\langle y, h(x)\rangle$, whose Lipschitz constant M_g does not improve under acceleration [51]. Zhang and Lan [64] resolved this in the deterministic first-order setting by *sliding*: in phase t the constraints are linearized at a center u_t, and the resulting bilinear saddle-point subproblem is solved by S_t cheap primal–dual steps that only multiply vectors by the fixed Jacobian $\nabla g(u_t)$; the gradients themselves are evaluated once per phase.

In the stochastic zeroth-order setting the Jacobian at u_t is not available; it must be estimated from paired queries at u_t. Reusing a single estimate $\hat{J}_t$ during the whole inner loop is not admissible: the inner dual iterates depend on $\hat{J}_t$, and the error terms $\big\langle \bar{y}_t, (\hat{J}_t - J_t)(x_\gamma^* - u_t)\big\rangle$ are not centered: a bias of order $RM_g D\sqrt{\nu_m/b}$ per phase (Theorem 3.4) would have to be brought below ε, which forces a mini-batch of size $b = \Omega(R^2 M_g^2 D^2 \nu_m/\varepsilon^2)$ in each of the $N = O(\varepsilon^{-1/2})$ phases, i.e. a total of $O(\varepsilon^{-5/2})$ calls. Instead we draw, *in a single round*, $2S_t + 1$ independent mini-batches at the same center u_t and assign a fresh mini-batch to each inner step: one for the primal direction (bank A) and one for the dual direction (bank B). Because the locations of all queries of the round are known in advance, the round complexity is the number of phases, N, while the estimation errors of the inner steps remain conditionally centered and are absorbed by the increments of the inner iterates. The price is that every inner step consumes oracle calls, so that the number of calls is proportional to the number of inner steps $K_N = \sum_t S_t$ rather than to N; since $K_N = O(\varepsilon^{-1})$ and each step uses $b = O(\varepsilon^{-1})$ samples, the total $O(\varepsilon^{-2})$ is unchanged.

6.2 The method

Fix the parameters $L \ge H$ (see (13)), M_g, R, γ, the shifts b_i, a dual scale $\varrho > 0$, a stabilizer $\lambda \ge 0$, two mini-batch sizes $b_p, b_y \ge 1$ (for the primal banks and the reserve batch, resp. for the dual banks; the two sizes are allowed to differ because the two kinds of banks carry different information at different prices, see Theorem 6.7) and the number of phases N. The schedule is

$$\tau_t = \frac{t-1}{2}, \qquad \theta_t = \frac{t-1}{t}, \qquad S_t = \max\Big\{1, \Big\lceil \frac{M_g \varrho\, t}{L} \Big\rceil\Big\}, \qquad a_t = \frac{t}{S_t},$$
$$\eta_t = \frac{2L}{t}, \qquad \beta_t = \frac{L+\lambda}{a_t}, \qquad W_N = \frac{N(N+1)}{2}, \qquad K_N = \sum_{t=1}^{N} S_t, \qquad A_N = \sum_{t=1}^{N} \frac{t^2}{S_t}. \tag{28}$$

Algorithm 2 ZO-SLIDING: zeroth-order primal–dual sliding with fresh mini-batches

Require: parameters as in (28); $x_0 \in Q$; set $x_{-1} := u_0 := p_0 := x_0$, $\tilde{y}_0 = \tilde{y}_{-1} := \tilde{y}^{\mathrm{u}}$, $k := 0$.

for $t = 1, \dots, N$ **do**

  $u_t := \dfrac{\tau_t u_{t-1} + x_{t-1} + \theta_t (x_{t-1} - x_{t-2})}{1 + \tau_t}$. (center; $u_1 = x_0$)

  Round t**.** Issue at u_t: S_t mini-batches A_k (b_p paired queries each, $k = K_{t-1}+1, \dots, K_t$) and

  S_t mini-batches B_k (b_y paired and b_y single queries each); if $t \geq 2$, issue at u_{t-1} one reserve

  mini-batch A_t^- of b_p paired queries (constraint values only).

  for $s = 1, \dots, S_t$ **do**

    $k := k + 1$.

    if $s = 1$ **then** $\hat{r}_k := \hat{g}_0^{A_k}(u_t) + \hat{J}^{A_k}(u_t)^\top y_{k-1} + \dfrac{a_{t-1}}{a_t} \hat{J}^{A_t^-}(u_{t-1})^\top (y_{k-1} - y_{k-2})$ (the last term is absent for $t = 1$)

    else $\hat{r}_k := \hat{g}_0^{A_k}(u_t) + \hat{J}^{A_k}(u_t)^\top (2y_{k-1} - y_{k-2})$

    end if

    $p_k := \arg\min_{x \in Q} \{ \langle \hat{r}_k, x \rangle + \eta_t V_{\mathcal{X}}(x_{t-1}, x) + \beta_t V_{\mathcal{X}}(p_{k-1}, x) \}$ (Theorem 4.4(d))

    $\hat{q}_k := \hat{h}^{B_k}(u_t) + \hat{J}^{B_k}(u_t)\,(p_k - u_t) \in \mathbb{R}^m$

    $\tilde{y}_k := \arg\max_{\tilde{y} \in \tilde{Y}_R} \{ \langle (0, \hat{q}_k), \tilde{y} \rangle - \varrho^{-2} \beta_t V_{\mathcal{Y}}(\tilde{y}_{k-1}, \tilde{y}) \}$ (Theorem 4.4(c))

  end for

  $x_t := \frac{1}{S_t} \sum_{k=K_{t-1}+1}^{K_t} p_k$, $\quad \bar{y}_t := \frac{1}{S_t} \sum_{k=K_{t-1}+1}^{K_t} y_k$.

end for

Ensure: $\bar{x}_N := \frac{1}{W_N} \sum_{t=1}^{N} t\, x_t$.

Inner steps are indexed globally by $k = 1, \dots, K_N$; slot k belongs to phase $t(k)$ and the inner steps of phase t are $k = K_{t-1}+1, \dots, K_t$. For a mini-batch A of paired queries at a point u we write $\hat{g}_0^A(u)$ and $\hat{J}^A(u)$ for the averages of the estimators (9) over the batch, and for a mini-batch B of single queries we write $\hat{h}^B(u)$.

In round t the algorithm issues $(2b_p + 3b_y)S_t + 2b_p \cdot \mathbb{1}_{t \geq 2}$ vector calls: $2b_p S_t$ from the banks A_k, $3b_y S_t$ from the banks B_k and $2b_p$ from the reserve mini-batch. Every call is used for the constraint vector, and the $2b_p S_t$ calls of the banks A_k (whose pairs deliver both F and G) are also used for the objective; therefore

$$N_{\mathrm{seq}} = N, \qquad N_{\mathrm{orc}} = N_g = (2b_p + 3b_y)K_N + 2b_p(N-1), \qquad N_f = 2b_p K_N, \qquad N_{\mathrm{inner}} = K_N. \tag{29}$$

With a common size $b_p = b_y = b$ this reads $N_{\mathrm{orc}} = N_g = 5bK_N + 2b(N-1)$ and $N_f = 2bK_N$. All query locations of round t (u_t and u_{t-1}) are determined before the round; only the *assignment* of the already sampled mini-batches to the inner steps depends on the inner iterates. The centers u_t are convex combinations of the previous phase outputs (Theorem 6.2), so all queries are located in Q_γ.

6.3 The deterministic skeleton

Throughout, $J_t := J_\gamma(u_t)$, and for $(x, y) \in \mathbb{R}^d \times \mathbb{R}^m$

$$\ell_t(x, y) := f_\gamma(u_t) + \langle \nabla f_\gamma(u_t), x - u_t \rangle + \langle y, h(u_t) + J_t(x - u_t) \rangle \tag{30}$$

is the linearization of $\Phi_\gamma(\cdot, y)$ at the center; by convexity of f_γ and h_i, $\ell_t(x, y) \leq \Phi_\gamma(x, y)$ for $y \geq 0$. The exact (noise-free) inner directions are

$$r_k := \nabla f_\gamma(u_t) + J_t^\top y_{k-1} + \rho_k J_{t'(k)}^\top (y_{k-1} - y_{k-2}), \qquad q_k := h(u_t) + J_t(p_k - u_t), \tag{31}$$

where, for a slot k of phase t that is not the first one, $\rho_k := 1$ and $t'(k) := t$; for the first slot of phase $t \geq 2$, $\rho_k := a_{t-1}/a_t$ and $t'(k) := t-1$; and for $k = 1$ the last term vanishes because $y_0 = y_{-1}$. The algorithm uses $\hat{r}_k = r_k + e_k^p$ and $\hat{q}_k = q_k + e_k^y$ with the estimation errors

$$e_k^p := \hat{r}_k - r_k, \qquad e_k^y := \hat{q}_k - q_k. \tag{32}$$

Lemma 6.1 (schedule)**.** *Let $c := M_g \varrho / L$. For all $t \geq 1$: (i) $a_t M_g \varrho \leq L$; (ii) $a_{t-1}/a_t \leq 2$ for $t \geq 2$; (iii) $\max\{N, \frac{c}{2}N(N+1)\} \leq K_N \leq N + \frac{c}{2}N(N+1)$; (iv) $W_N^2/K_N \leq A_N \leq \frac{5}{3} W_N^2/K_N$.*

Lemma 6.2 (centers)**.** *For arbitrary $x_{-1} = x_0, x_1, \ldots$ and u_t defined by the recursion of Algorithm 2, $u_1 = x_0$ and for $t \geq 2$*

$$u_t = \frac{1}{t(t+1)} \Big(\sum_{j=1}^{t-2} 2j\, x_j + (4t-2)\, x_{t-1} \Big), \tag{33}$$

a convex combination of $x_1, \ldots, x_{t-1}$. Moreover, with $d_t := x_t - x_{t-1}$,

$$x_t - u_t = d_t - \theta_t d_{t-1} + \tau_t (u_t - u_{t-1}). \tag{34}$$

The next lemma is the analytic heart of the acceleration. It is a statement about the extrapolation rule alone, valid for *arbitrary* points x_t: whatever the inner loop produces, the weighted linearizations at the centers control the function value at the weighted average up to a quadratic penalty on the increments. It is a reformulation, with explicit constants, of the conjugate-based argument of Zhang and Lan [64].

Lemma 6.3 (sliding inequality)**.** *Let $\mathsf{F} : \mathbb{R}^d \to \mathbb{R}$ be convex and L-smooth, let $x_{-1} = x_0, x_1, \ldots, x_N \in \mathbb{R}^d$ be arbitrary, let u_t be given by the recursion of Algorithm 2 with $\tau_t = (t-1)/2$, $\theta_t = (t-1)/t$, let $\ell_t^{\mathsf{F}}(x) := \mathsf{F}(u_t) + \langle \nabla \mathsf{F}(u_t), x - u_t \rangle$ and $\bar{x}_N := W_N^{-1} \sum_{t=1}^N t x_t$. Then*

$$W_N\, \mathsf{F}(\bar{x}_N) \leq \sum_{t=1}^N t\, \ell_t^{\mathsf{F}}(x_t) + L \sum_{t=1}^N \|x_t - x_{t-1}\|^2 . \tag{35}$$

Lemma 6.4 (energy inequality of the inner loop)**.** *Let Theorems 2.1, 2.2 and 2.5 hold, $L \geq H$, and let the iterates be generated by Algorithm 2 with the directions $\hat{r}_k = r_k + e_k^p$, $\hat{q}_k = q_k + e_k^y$ for arbitrary vectors $e_k^p \in \mathbb{R}^d$, $e_k^y \in \mathbb{R}^m$. Then, pathwise, for every $y \in Y_R$,*

$$\begin{aligned} \sum_{t=1}^N t \big[\ell_t(x_t, y) - \ell_t(x_\gamma^*, \bar{y}_t) \big] + L \sum_{t=1}^N \|x_t - x_{t-1}\|^2 + \frac{\lambda}{2} \sum_{k=1}^{K_N} \Big[\|p_k - p_{k-1}\|^2 + \frac{\|y_k - y_{k-1}\|_1^2}{\varrho^2} \Big] \\ \leq (L+\lambda) D_A^2 - \sum_{k=1}^{K_N} a_k \langle e_k^p, p_k - x_\gamma^* \rangle + \sum_{k=1}^{K_N} a_k \langle e_k^y, y_k - y \rangle , \end{aligned} \tag{36}$$

where $a_k := a_{t(k)}$ and

$$D_A^2 := 3\mathcal{A} + \frac{\mathcal{B}}{\varrho^2}, \qquad \mathcal{A} = \tfrac{1}{2} \left\| x_0 - x_\gamma^* \right\|^2, \quad \mathcal{B} = R^2 \log(m+1). \tag{37}$$

The proofs of Theorems 6.1 to 6.4 are given in Section D. The proof of Theorem 6.4 is a careful bookkeeping of three-point inequalities: the prox terms with coefficient η_t telescope over phases and produce exactly the term $-L \sum \|d_t\|^2$ needed to cancel the penalty of (35); the prox terms with coefficient β_t telescope over slots because $a_k \beta_k = L + \lambda$ is constant; the bilinear cross terms created by the dual extrapolation telescope across phase boundaries thanks to the reserve mini-batch and the ratio a_{t-1}/a_t, and the leftovers are absorbed by Young's inequality using $a_t M_g \varrho \leq L$.

6.4 Main result

Theorem 6.5 (ZO-Sliding). *Let Theorems 2.1 to 2.3 and 2.5 hold, $\gamma \le \bar{\gamma}$, R as in (14), $L \ge H$ as in (13), and let the starting point $x_0 \in Q$ be deterministic or independent of the oracle samples. For every $\lambda > 0$, $\varrho > 0$, $b_p, b_y \ge 1$ and $N \ge 1$ the output of Algorithm 2 satisfies*

$$W_N \, \mathbb{E}\, \mathsf{C}_{R,\gamma}(\bar{x}_N) \le (L + 2\lambda)\, \mathbb{E} D_A^2 + \frac{A_N \Sigma_b^2}{\lambda}, \tag{38}$$

where

$$\Sigma_b^2 := \frac{17\, d M_R^2}{b_p} + \frac{\varrho^2 \sigma_y^2}{b_y}, \qquad \sigma_y^2 := 16\kappa_m \big(\sigma_h^2 + \nu_m M_g^2 D^2\big), \tag{39}$$

with σ_h from (10) and $\nu_m = \min\{4d, 92\kappa_m\}$ from Theorem 3.4. The first term of Σ_b^2 is the noise of the primal banks (gradient information), the second that of the dual banks (level and linearization information); with a common batch size $b_p = b_y = b$, $\Sigma_b^2 = \Sigma_A^2/b$ with $\Sigma_A^2 := 17 d M_R^2 + \varrho^2 \sigma_y^2$.

Proof. Step 1 (filtration and centering). Order the mini-batches of a round in the order in which the inner loop consumes them: A_t^- (if $t \ge 2$), then A_k, B_k for $k = K_{t-1}+1, \dots, K_t$. Let $\mathcal{F}_k^A$ be the σ-field generated by x_0 and by all mini-batches consumed up to and including A_k, and $\mathcal{F}_k^B$ the one that also includes B_k; $\mathcal{F}_0^B$ is generated by x_0. Then $p_{k-1}, y_{k-1}, y_{k-2}$ and u_t, u_{t-1} are $\mathcal{F}_{k-1}^B$-measurable, the batches A_k and A_t^- are independent of $\mathcal{F}_{k-1}^B$, p_k is $\mathcal{F}_k^A$-measurable and B_k is independent of $\mathcal{F}_k^A$. By Theorem 3.4(a) and Theorem 3.5, applied conditionally with the (measurable) weights y_{k-1}, $\rho_k(y_{k-1} - y_{k-2})$ and $w_k := p_k - u_t$,

$$\mathbb{E}[e_k^p \mid \mathcal{F}_{k-1}^B] = 0, \qquad \mathbb{E}[e_k^y \mid \mathcal{F}_k^A] = 0. \tag{40}$$

Step 2 (second moments). For a slot k that is not the first of its phase, e_k^p is the average over the b pairs of A_k of centered estimation errors of the Lagrangian gradient with weight vector $2y_{k-1} - y_{k-2}$, whose ℓ_1-norm is at most $3R$; by Theorem 3.4(a), $\mathbb{E}[\|e_k^p\|^2 \mid \mathcal{F}_{k-1}^B] \le 2d(M_0 + 3RM_g)^2/b_p \le 18 d M_R^2/b_p$. For the first slot of a phase $t \ge 2$, e_k^p is the sum of two conditionally independent centered terms, from A_k (weights y_{k-1}, ℓ_1-norm $\le R$) and from A_t^- (weights $\rho_k(y_{k-1} - y_{k-2})$, ℓ_1-norm $\le 2 \cdot 2R$ by Theorem 6.1(ii)), so $\mathbb{E}[\|e_k^p\|^2 \mid \mathcal{F}_{k-1}^B] \le [2dM_R^2 + 2d(4RM_g)^2]/b_p \le 34 d M_R^2/b_p$. Hence in all cases

$$\mathbb{E}\big[\|e_k^p\|^2 \mid \mathcal{F}_{k-1}^B\big] \le \frac{34\, d M_R^2}{b_p}. \tag{41}$$

For the dual error, $e_k^y = [\hat{h}^{B_k}(u_t) - h(u_t)] + [\hat{J}^{B_k}(u_t) - J_t] w_k$ with $\|w_k\| = \|p_k - u_t\| \le D$ (both p_k and u_t lie in Q by Theorem 6.2); both brackets are averages of b_y conditionally i.i.d. centered vectors with ℓ_∞ second moments bounded by σ_h^2 (Theorem 3.5) and $\nu_m M_g^2 D^2$ (Theorem 3.4(c),(d)) respectively. By Theorem 3.6 and $\|a + c\|_\infty^2 \le 2\|a\|_\infty^2 + 2\|c\|_\infty^2$,

$$\mathbb{E}\big[\|e_k^y\|_\infty^2 \mid \mathcal{F}_k^A\big] \le \frac{16\kappa_m}{b_y}\big(\sigma_h^2 + \nu_m M_g^2 D^2\big) = \frac{\sigma_y^2}{b_y}. \tag{42}$$

Step 3 (splitting the error terms). Write $p_k - x_\gamma^* = (p_k - p_{k-1}) + (p_{k-1} - x_\gamma^*)$ and $y_k - y = (y_k - y_{k-1}) + (y_{k-1} - y)$ in the right-hand side of (36). By Young's inequality,

$$-a_k \langle e_k^p, p_k - p_{k-1} \rangle \le \frac{a_k^2}{2\lambda} \|e_k^p\|^2 + \frac{\lambda}{2} \|p_k - p_{k-1}\|^2,$$
$$a_k \langle e_k^y, y_k - y_{k-1} \rangle \le \frac{a_k^2 \varrho^2}{2\lambda} \|e_k^y\|_\infty^2 + \frac{\lambda}{2\varrho^2} \|y_k - y_{k-1}\|_1^2,$$

and the two quadratic increments are cancelled by the left-hand side of (36). Consequently, pathwise for all $y \in Y_R$,

$$\sum_t t\big[\ell_t(x_t,y)-\ell_t(x^*_\gamma,\bar y_t)\big]+L\sum_t\|d_t\|^2$$
$$\le (L+\lambda)D_A^2+\sum_k\frac{a_k^2}{2\lambda}\big[\|e_k^p\|^2+\varrho^2\|e_k^y\|_\infty^2\big]-\sum_k a_k\langle e_k^p,p_{k-1}-x^*_\gamma\rangle+\sum_k a_k\langle e_k^y,y_{k-1}-y\rangle. \tag{43}$$

Step 4 (from linearizations to the certificate). Fix $y\in Y_R$ and apply Theorem 6.3 to $\mathsf{F}:=\Phi_\gamma(\cdot,y)$, which is convex and L-smooth on all of $\mathbb{R}^d$ by Theorem 3.2 and (13) (this is the only place where a property outside $Q_{\bar\gamma}$ is used; the points $u_t,x_t,\bar x_N$ at which the inequality is evaluated lie in Q), with $\ell_t^{\mathsf{F}}=\ell_t(\cdot,y)$: $W_N\Phi_\gamma(\bar x_N,y)\le\sum_t t\,\ell_t(x_t,y)+L\sum_t\|d_t\|^2$. Moreover $\ell_t(x^*_\gamma,\bar y_t)\le\Phi_\gamma(x^*_\gamma,\bar y_t)\le f^*_\gamma$ by convexity and (19), so $\sum_t t\,\ell_t(x^*_\gamma,\bar y_t)\le W_N f^*_\gamma$. Together with (43),

$$W_N\big[\Phi_\gamma(\bar x_N,y)-f^*_\gamma\big]\le(L+\lambda)D_A^2+\sum_k\frac{a_k^2}{2\lambda}\big[\|e_k^p\|^2+\varrho^2\|e_k^y\|_\infty^2\big]$$
$$-\sum_k a_k\langle e_k^p,p_{k-1}-x^*_\gamma\rangle+\sum_k a_k\langle e_k^y,y_{k-1}-y\rangle$$

for every $y\in Y_R$. Taking the supremum over y on both sides and recalling (17),

$$W_N\,\mathsf{C}_{R,\gamma}(\bar x_N)\le(L+\lambda)D_A^2+\sum_k\frac{a_k^2}{2\lambda}\big[\|e_k^p\|^2+\varrho^2\|e_k^y\|_\infty^2\big]$$
$$-\sum_k a_k\langle e_k^p,p_{k-1}-x^*_\gamma\rangle+\sup_{y\in Y_R}\sum_k a_k\langle e_k^y,y_{k-1}-y\rangle. \tag{44}$$

Step 5 (expectations). By (40) and the measurability of p_{k-1}, $\mathbb{E}\langle e_k^p,p_{k-1}-x^*_\gamma\rangle=0$ (all quantities are bounded, since Q is compact and the second moments are finite). Theorem 3.7 applied on $Z=\tilde Y_R$ with $\omega=R\sum_i\tilde y_i\log\tilde y_i$, the norm $\|\cdot\|_1$ (dual norm $\|\cdot\|_\infty$), $\Theta=\mathcal{B}$, $z_0=\tilde y^{\mathrm{u}}$, the filtration $(\mathcal{F}_k^B)_k$ (note that e_k^y is $\mathcal{F}_k^B$-measurable and, by (40) and the tower property with $\mathcal{F}_{k-1}^B\subseteq\mathcal{F}_k^A$, $\mathbb{E}[e_k^y\mid\mathcal{F}_{k-1}^B]=0$), $\zeta_k=\tilde y_{k-1}$ and the vectors $e_k:=-(0,e_k^y)$ (so that $a_k\langle e_k,\tilde y-\zeta_k\rangle=a_k\langle e_k^y,y_{k-1}-y\rangle$) gives, for every $\lambda'>0$,

$$\mathbb{E}\sup_{y\in Y_R}\sum_k a_k\langle e_k^y,y_{k-1}-y\rangle\le\lambda'\mathcal{B}+\frac{1}{2\lambda'}\sum_k a_k^2\,\mathbb{E}\|e_k^y\|_\infty^2.$$

Choosing $\lambda'=\lambda/\varrho^2$ and using (41), (42) and $\sum_k a_k^2=\sum_t S_t a_t^2=A_N$ in (44),

$$W_N\,\mathbb{E}\mathsf{C}_{R,\gamma}(\bar x_N)\le(L+\lambda)\mathbb{E}D_A^2+\frac{\lambda\mathcal{B}}{\varrho^2}+\frac{A_N}{2\lambda}\Big(\frac{34dM_R^2}{b_p}+\frac{\varrho^2\sigma_y^2}{b_y}\Big)+\frac{A_N\varrho^2\sigma_y^2}{2\lambda b_y}.$$

Since $\mathcal{B}/\varrho^2\le D_A^2$, the right-hand side is at most $(L+2\lambda)\mathbb{E}D_A^2+A_N\Sigma_b^2/\lambda$, which is (38). (When x_0 is random, the whole argument is carried out conditionally on x_0 and then integrated.) □

6.5 Choice of the parameters and complexity

Corollary 6.6 (optimized stabilizer)**.** *In the setting of Theorem 6.5 with deterministic x_0, the choice $\lambda:=\Sigma_b\sqrt{A_N/2}/D_A$ gives*

$$\mathbb{E}\,\mathsf{C}_{R,\gamma}(\bar x_N)\le\frac{2LD_A^2}{N(N+1)}+\frac{2}{W_N}\sqrt{2A_N\Sigma_b^2D_A^2}\le\frac{2LD_A^2}{N(N+1)}+4\,D_A\sqrt{\frac{17dM_R^2}{b_pK_N}+\frac{\varrho^2\sigma_y^2}{b_yK_N}}. \tag{45}$$

Proof. Minimizing $2\lambda D_A^2+A_N\Sigma_b^2/\lambda$ over $\lambda>0$ gives the stated λ and the value $2\sqrt{2A_N\Sigma_b^2D_A^2}$; divide (38) by W_N. By Theorem 6.1(iv), $A_N\le\frac{5}{3}W_N^2/K_N$, and $2\sqrt{10/3}\le4$. □

The bound (45) has the structure of the optimal rate for smooth stochastic convex optimization with mini-batches, $LD_A^2/N^2 + \Sigma D_A/\sqrt{\text{number of samples}}$, in which b_pK_N and b_yK_N are the numbers of primal and dual mini-batches consumed and N the number of rounds; with a common batch size the second term is $4\Sigma_A D_A/\sqrt{bK_N}$.

Corollary 6.7 (complexity of ZO-Sliding)**.** *In the setting of Theorem 6.6, let $\varepsilon_{\mathrm{s}} > 0$ with $\varepsilon_{\mathrm{s}} \le LD_A^2$, and choose*

$$N := \Big\lceil 2D_A\sqrt{L/\varepsilon_{\mathrm{s}}}\Big\rceil, \qquad b_p := \max\Big\{1, \Big\lceil \frac{2176\, dM_R^2 D_A^2}{K_N\varepsilon_{\mathrm{s}}^2}\Big\rceil\Big\}, \qquad b_y := \max\Big\{1, \Big\lceil \frac{128\,\varrho^2\sigma_y^2 D_A^2}{K_N\varepsilon_{\mathrm{s}}^2}\Big\rceil\Big\}. \tag{46}$$

Then $\mathbb{E}\mathsf{C}_{R,\gamma}(\bar{x}_N) \le \varepsilon_{\mathrm{s}}$ and

$$\begin{aligned} N_{\mathrm{seq}} = N \le 2D_A\sqrt{\frac{L}{\varepsilon_{\mathrm{s}}}} + 1, \qquad N_{\mathrm{inner}} = K_N \le N + \frac{6M_g\varrho D_A^2}{\varepsilon_{\mathrm{s}}}, \\ N_f \le 2K_N + \frac{4352\, dM_R^2 D_A^2}{\varepsilon_{\mathrm{s}}^2}, \qquad N_{\mathrm{orc}} = N_g \le 7K_N + \frac{8704\, dM_R^2 D_A^2}{\varepsilon_{\mathrm{s}}^2} + \frac{384\,\varrho^2\sigma_y^2 D_A^2}{\varepsilon_{\mathrm{s}}^2}. \end{aligned} \tag{47}$$

The objective calls thus pay only for the gradient information (dM_R^2), the constraint calls for the gradient and for the level information.

Proof. By (45), $2LD_A^2/(N(N+1)) \le 2LD_A^2/N^2 \le \varepsilon_{\mathrm{s}}/2$, and the second term of (45) is at most $\varepsilon_{\mathrm{s}}/2$ because $17dM_R^2/(b_pK_N) \le \varepsilon_{\mathrm{s}}^2/(128D_A^2)$ and $\varrho^2\sigma_y^2/(b_yK_N) \le \varepsilon_{\mathrm{s}}^2/(128D_A^2)$, so that the square root is at most $\varepsilon_{\mathrm{s}}/(8D_A)$. Put $\vartheta := D_A\sqrt{L/\varepsilon_{\mathrm{s}}} \ge 1$; then $N \le 2\vartheta + 1 \le 3\vartheta$ and $N+1 \le 4\vartheta$, so $N(N+1) \le 12\vartheta^2 = 12LD_A^2/\varepsilon_{\mathrm{s}}$ and Theorem 6.1(iii) gives $K_N \le N + \frac{M_g\varrho}{2L}N(N+1) \le N + 6M_g\varrho D_A^2/\varepsilon_{\mathrm{s}}$. Finally $b_p \le 1 + 2176dM_R^2D_A^2/(K_N\varepsilon_{\mathrm{s}}^2)$, $b_y \le 1 + 128\varrho^2\sigma_y^2D_A^2/(K_N\varepsilon_{\mathrm{s}}^2)$, $N_f = 2b_pK_N$ and $N_{\mathrm{orc}} = N_g = (2b_p + 3b_y)K_N + 2b_p(N-1) \le 4b_pK_N + 3b_yK_N$ by (29) and $N \le K_N$. □

The dual scale ϱ trades the size of $D_A^2 = 3\mathcal{A} + \mathcal{B}/\varrho^2$ against the dual noise $\varrho^2\sigma_y^2$ and the number of inner steps. A choice that depends only on known quantities and balances the two prox radii is $\varrho^2 = 2\mathcal{B}/D^2$:

Corollary 6.8 (explicit rates for $\varrho^2 = 2\mathcal{B}/D^2$)**.** *Choose $\varrho := \sqrt{2\mathcal{B}}/D$ in Theorem 6.7. Then $D_A^2 \le 2D^2$ and, with $\Lambda_m = \kappa_m\log(m+1)$,*

$$\begin{aligned} N \le 2D\sqrt{\frac{2L}{\varepsilon_{\mathrm{s}}}} + 1, \qquad K_N \le N + \frac{17\, M_gRD\sqrt{\log(m+1)}}{\varepsilon_{\mathrm{s}}}, \qquad N_f \le 2K_N + \frac{8704\, dM_R^2D^2}{\varepsilon_{\mathrm{s}}^2}, \\ N_{\mathrm{orc}} = N_g \le 7K_N + \frac{17408\, dM_R^2D^2 + 24576\,\Lambda_mR^2\big(\sigma_h^2 + \nu_mM_g^2D^2\big)}{\varepsilon_{\mathrm{s}}^2}. \end{aligned} \tag{48}$$

Consequently $N_{\mathrm{seq}} = O(\sqrt{LD^2/\varepsilon_{\mathrm{s}}})$, $N_{\mathrm{inner}} = O(N_{\mathrm{seq}} + M_gRD\sqrt{\log(m+1)}/\varepsilon_{\mathrm{s}})$ and

$$N_f = O\Big(N_{\mathrm{inner}} + \frac{dM_R^2D^2}{\varepsilon_{\mathrm{s}}^2}\Big), \qquad N_{\mathrm{orc}} = N_g = O\Big(N_{\mathrm{inner}} + \frac{dM_R^2D^2 + \Lambda_mR^2(\sigma_h^2 + \nu_mM_g^2D^2)}{\varepsilon_{\mathrm{s}}^2}\Big);$$

since $\nu_m \le 4d$ and $RM_g \le M_R$, the last expression is in turn $O\big(N_{\mathrm{inner}} + \Lambda_m(dM_R^2D^2 + R^2\sigma_h^2)/\varepsilon_{\mathrm{s}}^2\big)$, a cruder form in which the factor Λ_m also multiplies the primal term. The term $\Lambda_mR^2\nu_mM_g^2D^2$ is dimension-free ($\nu_m \le 92\kappa_m$) and is dominated by $dM_R^2D^2$ as soon as $d \ge 92\kappa_m^2\log(m+1)$.

Proof. $D_A^2 = 3\mathcal{A} + \mathcal{B}/\varrho^2 \le \frac{3}{2}D^2 + \frac{1}{2}D^2 = 2D^2$, so $dM_R^2D_A^2 \le 2dM_R^2D^2$ and $\varrho^2\sigma_y^2D_A^2 \le (2\mathcal{B}/D^2)\cdot 16\kappa_m(\sigma_h^2 + \nu_mM_g^2D^2)\cdot 2D^2 = 64\Lambda_mR^2(\sigma_h^2 + \nu_mM_g^2D^2)$; insert this into (47). $6M_g\varrho D_A^2 \le 6M_g(\sqrt{2\mathcal{B}}/D)(2D^2) = 12\sqrt{2}M_gRD\sqrt{\log(m+1)}$ and $12\sqrt{2} \le 17$. □

Corollary 6.9 (smooth data)**.** *Let Theorems 2.1 to 2.5 hold and* $\varepsilon \in (0, \min\{2\max\{L_0, L_g\}\bar{\gamma}^2,\ (L_0 + RL_g)D^2\}]$. *Take* $\gamma := \min\{\sqrt{\varepsilon/(2\max\{L_0, L_g\})},\ \varepsilon/(4M_g)\}$, $b_0 = L_0\gamma^2/2$, $b_i = L_g\gamma^2/2$, $L := L_0 + RL_g$ *and run Algorithm 2 with the parameters of Theorem 6.8 for* $\varepsilon_{\mathrm{s}} = \varepsilon/2$. *Then* $\bar{x}_N$ *is an* ε*-solution in expectation and*

$$N_{\mathrm{seq}} \le 4D\sqrt{\frac{L_0 + RL_g}{\varepsilon}} + 1, \qquad N_{\mathrm{inner}} = O\Big(N_{\mathrm{seq}} + \frac{M_g RD\sqrt{\log(m+1)}}{\varepsilon}\Big),$$
$$N_f = O\Big(N_{\mathrm{inner}} + \frac{dM_R^2 D^2}{\varepsilon^2}\Big), \qquad N_{\mathrm{orc}} = N_g = O\Big(N_{\mathrm{inner}} + \frac{dM_R^2 D^2 + \Lambda_m R^2(\sigma_g^2 + \nu_m M_g^2 D^2)}{\varepsilon^2}\Big).$$

Proof. By Theorem 3.1(c), $L_{0,\gamma} = L_0$ and $L_{g,\gamma} = L_g$ for every $\gamma > 0$, so $L = H$; the second moment of the two-point estimators does not depend on γ (Theorem 3.3), so the radius may be taken as small as convenient. The biases satisfy $b_0, b_g \le \max\{L_0, L_g\}\gamma^2/2 \le \varepsilon/4$; by Theorem 4.1(c) and Theorem 4.3, $\mathbb{E}\mathsf{C} \le \varepsilon/2$ gives $\mathbb{E}[f - f^*] \le 3\varepsilon/4$ and $\mathbb{E}\mathsf{v} \le 3\varepsilon/4$. Since $\gamma \le \varepsilon/(4M_g)$, $8\gamma^2 M_g^2 \le \varepsilon^2/2$ and $\sigma_h^2 \le 2\sigma_g^2 + \varepsilon^2/2$, so that the contribution of the smoothing part of σ_h^2 to N_{orc} is $O(\Lambda_m R^2)$, a constant. The hypothesis $\varepsilon \le LD^2$ guarantees $\varepsilon_{\mathrm{s}} = \varepsilon/2 \le LD_A^2$, because $D_A^2 \ge \mathcal{B}/\varrho^2 = D^2/2$. Insert $\varepsilon_{\mathrm{s}} = \varepsilon/2$ into (48). □

Corollary 6.10 (nonsmooth data)**.** *Let Theorems 2.1 to 2.3 and 2.5 hold and* $\varepsilon \in (0, \min\{4M\bar{\gamma},\ 2d^{1/4}\sqrt{M_R M}D\}]$ *Take* $\gamma := \varepsilon/(4M)$, $b_0 = \gamma M_0$, $b_i = \gamma M_g$, $L := \sqrt{d}M_R/\gamma = 4\sqrt{d}M_R M/\varepsilon$, *and run Algorithm 2 with the parameters of Theorem 6.8 for* $\varepsilon_{\mathrm{s}} = \varepsilon/2$. *Then* $\bar{x}_N$ *is an* ε*-solution in expectation and*

$$N_{\mathrm{seq}} \le \frac{8\, d^{1/4}\sqrt{M_R M}\, D}{\varepsilon} + 1, \qquad N_{\mathrm{inner}} = O\Big(N_{\mathrm{seq}} + \frac{M_g RD\sqrt{\log(m+1)}}{\varepsilon}\Big),$$
$$N_f = O\Big(N_{\mathrm{inner}} + \frac{dM_R^2 D^2}{\varepsilon^2}\Big), \qquad N_{\mathrm{orc}} = N_g = O\Big(N_{\mathrm{inner}} + \frac{dM_R^2 D^2 + \Lambda_m R^2(\sigma_g^2 + \nu_m M_g^2 D^2)}{\varepsilon^2}\Big).$$

Proof. $H = L_{0,\gamma} + RL_{g,\gamma} = \sqrt{d}(M_0 + RM_g)/\gamma = L$ by Theorem 3.1(c) and Theorem 3.2. The biases are $b_0 \le \gamma M = \varepsilon/4$ and $b_g \le \varepsilon/4$, and $N \le 2D\sqrt{2L/\varepsilon_{\mathrm{s}}} + 1 = 2D\sqrt{16\sqrt{d}M_R M/\varepsilon^2} + 1 = 8d^{1/4}\sqrt{M_R M}D/\varepsilon + 1$. Also $8\gamma^2 M_g^2 \le \varepsilon^2/2$, so $\sigma_h^2 \le 2\sigma_g^2 + \varepsilon^2/2$ and the contribution of $\varepsilon^2/2$ to N_{orc} is $O(\Lambda_m R^2)$, a constant. The hypothesis $\varepsilon \le 2d^{1/4}\sqrt{M_R M}D$ is equivalent to $\varepsilon/2 \le LD^2/2$ and guarantees $\varepsilon_{\mathrm{s}} \le LD_A^2$ because $D_A^2 \ge D^2/2$. □

Remark 6.11 (separate smoothing radii). Nothing forces a common smoothing radius for the objective and for the constraints. If f is smoothed with $\gamma_0 := \varepsilon/(4M_0)$ and every g_i with $\gamma_g := \varepsilon/(4M_g)$ (the estimators (9) use the respective radius, and Theorem 3.5 is applied with γ_g), the biases are still $b_0, b_g \le \varepsilon/4$ and the smoothness constant of the Lagrangian becomes

$$H = \frac{\sqrt{d}M_0}{\gamma_0} + \frac{R\sqrt{d}M_g}{\gamma_g} = \frac{4\sqrt{d}\,(M_0^2 + RM_g^2)}{\varepsilon} \le \frac{4\sqrt{d}\,M_R M}{\varepsilon},$$

since $M_0^2 \le M_0 M$ and $RM_g^2 \le RM_g M$. All statements of Sections 3 and 4 are per-function and remain valid, so Theorem 6.10 holds with $N_{\mathrm{seq}} \le 8d^{1/4}\sqrt{M_0^2 + RM_g^2}\, D/\varepsilon + 1$; when $M_0 \gg \sqrt{R}M_g$ this is within a constant of the unconstrained depth $d^{1/4}M_0 D/\varepsilon$. We keep a common radius in the main text to lighten the notation.

The round complexities of Theorems 6.9 and 6.10 coincide, up to the multiplier factor R in M_R and H, with the iteration complexities of the accelerated batched methods for *unconstrained* zeroth-order problems [26]: $\sqrt{L_0 D^2/\varepsilon}$ for smooth and $d^{1/4}MD/\varepsilon$ for nonsmooth objectives. These are the best known upper bounds for the unconstrained two-point problem; they are minimax optimal in the smooth case, while in the nonsmooth case the known depth lower bounds are weaker (see Section 9.4).

6.6 Discussion

What is paid for the depth reduction. Compared with the Mirror–Prox baseline, whose dual noise is only the level noise σ_h^2, the sliding method has the additional dual noise term $\nu_m M_g^2 D^2$ in (39). It comes from the estimate $\hat{J}^{B_k}(u_t)(p_k - u_t)$ of the linearized constraint value at the inner point p_k: the exact value $h(p_k)$ could be estimated by a single query at p_k, but p_k is not known before the round, and querying it would create a new sequential round. The linearization at the center replaces an *adaptive* query by a matrix–vector product with a fresh Jacobian sample; the variance of this product is $\nu_m M_g^2 \|p_k - u_t\|^2$. By Theorem 3.4(d), $\nu_m \le 92\kappa_m$, so the extra term is dimension-free and is dominated by the primal term $dM_R^2 D^2$ as soon as $d \gtrsim \kappa_m^2 \log(m+1)$; in low dimensions the crude bound $\nu_m \le 4d$ is the better one, and the extra term is then of the same order $dM_R^2 D^2$ as the primal one, up to the factor Λ_m.

Why the mini-batches must be fresh. If the same Jacobian estimate were reused within a phase, the inner dual iterates would be correlated with the estimation error and the term $\left\langle \bar{y}_t, (\hat{J}_t - J_t)(x_\gamma^* - u_t) \right\rangle$ in the gap could only be bounded by $R\,\mathbb{E}\left\|(\hat{J}_t - J_t)(x_\gamma^* - u_t)\right\|_\infty = O(RM_g D\sqrt{\nu_m/b_y})$, a bias that does not average out over the phases. The fresh mini-batches make the errors martingale differences with respect to the consumption order, which is exactly what (40) and Theorem 3.7 require. This is the reason why the counts $N_{\mathrm{orc}}, N_f, N_g$ in (29) are proportional to K_N rather than to N; the depth $N_{\mathrm{seq}} = N$ is unaffected.

Inner steps, calls and query points. Unlike the inner steps of ACGD-S, the inner steps of Algorithm 2 are not free in oracle calls: each of them consumes one primal and one dual mini-batch that were drawn in the current round. What they save is *rounds*: $N_{\mathrm{inner}} = K_N$ counts operations that need no new oracle round, and the $O(\varepsilon^{-2})$ calls of the method are located at only $N = O(\varepsilon^{-1/2})$ distinct centers (plus the previous center for the reserve mini-batch). For simulation oracles this concentration of the queries is an advantage of its own (state of the simulator, common random numbers) which is not reflected in the three complexity measures. In the affine case of Section 8 the inner steps are oracle-free in the literal sense.

Relation to ACGD-S. With exact gradients ($e_k^p = e_k^y = 0$, $\lambda = 0$) Theorems 6.3 and 6.4 give $\mathsf{C}_{R,\gamma}(\bar{x}_N) \le 2LD_A^2/(N(N+1))$, which is the deterministic guarantee of ACGD-S [64, Thm. 5] in our notation, with $K_N = O(N + M_g \varrho N^2/L)$ inner matrix–vector products. Our schedule (28) is theirs (τ_t, θ_t, $\eta_t = L/\tau_{t+1}$, weights $\omega_t = t$). The new ingredients are the stochastic zeroth-order estimators, the fresh-mini-batch design, the stabilizer λ, the explicit treatment of the dual noise through Theorems 3.6 and 3.7, and the transfer to the original constrained problem through the smoothing shifts.

Parameters. The method needs upper bounds L, M_g and R and the constants entering Σ_b. The stabilizer of Theorem 6.6 is the worst-case choice; in the experiments of Section 11 a much smaller stabilizer performs better, as is typical for worst-case constants.

Two batch sizes. The primal banks A_k (and the reserve batch) carry the gradient information about the Lagrangian, whose noise $17dM_R^2/b_p$ has the dimension factor d; the dual banks B_k carry the level and linearization information, whose noise $\varrho^2\sigma_y^2/b_y$ is dimension-free but does not decrease with the accuracy in the strongly convex case. Since the objective is used only in the banks A_k, $N_f = 2b_p K_N$, and the two sizes can be tuned to the two error sources independently (Theorem 6.7); this is what makes the two channels N_f and N_g of Theorem 2.8 behave differently in Theorem 7.2, where N_f inherits the $1/(\mu\varepsilon)$ rate of strongly convex optimization while N_g keeps the ε^{-2} level term. In the experiments a common size $b_p = b_y$ is used.

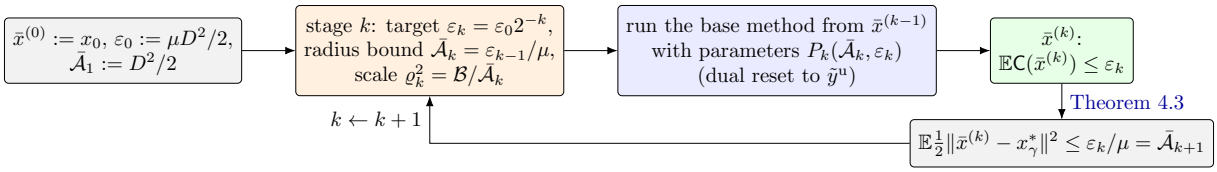


Figure 3: The restart scheme (Algorithm 3). The base method is either ZO-Sliding (Algorithm 2) or B-SMP (Algorithm 1); its parameters at stage k are chosen from the bound $\bar{\mathcal{A}}_k$ on the expected squared distance of the starting point to x_γ^*.

Algorithm 3 R-ZO-Sliding / R-B-SMP: restarts for strongly convex objectives

Require: μ, D, target $\varepsilon_{\mathrm{s}} \in (0, \mu D^2/2)$, a base method with a parameter rule $P(\bar{\mathcal{A}}, \varepsilon)$; starting point $\bar{x}^{(0)} := x_0 \in Q$.

1: $\varepsilon_0 := \mu D^2/2$, $K := \lceil \log_2(\varepsilon_0/\varepsilon_{\mathrm{s}}) \rceil$.
2: **for** $k = 1, \dots, K$ **do**
3: $\quad \varepsilon_k := \varepsilon_0 2^{-k}$, $\bar{\mathcal{A}}_k := \varepsilon_{k-1}/\mu$.
4: $\quad$ Run the base method from $\bar{x}^{(k-1)}$ with the dual variable reset to $\tilde{y}^{\mathrm{u}}$ and parameters $P(\bar{\mathcal{A}}_k, \varepsilon_k)$; let $\bar{x}^{(k)}$ be its output.
5: **end for**

Ensure: $\bar{x}^{(K)}$.

7 Strongly convex objectives: restarts

Under Theorem 2.6 the certificate controls the distance to the solution of the smoothed problem, $\frac{\mu}{2}\left\|x - x_\gamma^*\right\|^2 \le \mathsf{C}_{R,\gamma}(x)$ (Theorem 4.3). Both methods of the previous sections have guarantees that depend on the starting point only through $\mathcal{A} = \frac{1}{2}\left\|x_0 - x_\gamma^*\right\|^2$ and that are nondecreasing and concave in $\mathcal{A}$. This is exactly what is needed for the classical restart technique: halving the target accuracy at each stage halves the admissible bound on $\mathcal{A}$, so that the number of rounds per stage stays constant. The dual variable is reset to the uniform point at every stage; the dual radius $\mathcal{B}$ does not shrink, and the dual scale ϱ (or the balance α) is increased from stage to stage to keep the two prox radii balanced. Figure 3 shows the scheme.

Lemma 7.1 (restart lemma)**.** *Let Theorems 2.1, 2.2, 2.5 and 2.6 hold. Suppose that for every $\bar{\mathcal{A}} > 0$ and $\varepsilon > 0$ the base method with parameters $P(\bar{\mathcal{A}}, \varepsilon)$, started at a random point $x_0 \in Q$ independent of the oracle samples of the run, returns $\hat{x}$ with*

$$\mathbb{E}\left[\mathsf{C}_{R,\gamma}(\hat{x}) \mid x_0\right] \le \Psi_P\left(\tfrac{1}{2}\left\|x_0 - x_\gamma^*\right\|^2\right), \tag{49}$$

where $\Psi_P : [0,\infty) \to [0,\infty)$ is nondecreasing and concave with $\Psi_{P(\bar{\mathcal{A}},\varepsilon)}(\bar{\mathcal{A}}) \le \varepsilon$. Then the output of Algorithm 3 satisfies $\mathbb{E}\mathsf{C}_{R,\gamma}(\bar{x}^{(K)}) \le \varepsilon_K \le \varepsilon_{\mathrm{s}}$, and $\varepsilon_K > \varepsilon_{\mathrm{s}}/2$.

Proof. We show by induction that $\mathbb{E}\frac{1}{2}\left\|\bar{x}^{(k-1)} - x_\gamma^*\right\|^2 \le \bar{\mathcal{A}}_k$ and $\mathbb{E}\mathsf{C}(\bar{x}^{(k)}) \le \varepsilon_k$ for $k = 1, \dots, K$. For $k = 1$, $\frac{1}{2}\left\|x_0 - x_\gamma^*\right\|^2 \le D^2/2 = \varepsilon_0/\mu = \bar{\mathcal{A}}_1$. If the first claim holds for k, then by (49), Jensen's inequality (concavity) and monotonicity, $\mathbb{E}\mathsf{C}(\bar{x}^{(k)}) = \mathbb{E}\,\mathbb{E}[\mathsf{C}(\bar{x}^{(k)}) \mid \bar{x}^{(k-1)}] \le \mathbb{E}\Psi(\frac{1}{2}\left\|\bar{x}^{(k-1)} - x_\gamma^*\right\|^2) \le \Psi(\bar{\mathcal{A}}_k) \le \varepsilon_k$; here $\bar{x}^{(k-1)}$ is independent of the fresh samples of stage k. Then Theorem 4.3 gives $\mathbb{E}\frac{1}{2}\left\|\bar{x}^{(k)} - x_\gamma^*\right\|^2 \le \mathbb{E}\mathsf{C}(\bar{x}^{(k)})/\mu \le \varepsilon_k/\mu = \bar{\mathcal{A}}_{k+1}$. Finally $\varepsilon_K = \varepsilon_0 2^{-K} \in (\varepsilon_{\mathrm{s}}/2, \varepsilon_{\mathrm{s}}]$ by the definition of K. □

7.1 Restarted sliding

Theorem 7.2 (R-ZO-Sliding). *Let Theorems 2.1 to 2.3, 2.5 and 2.6 hold, $\gamma \le \bar\gamma$, $L \ge H$, $\varepsilon_{\mathrm{s}} \in (0, \mu D^2/2)$, and run Algorithm 3 with the base method Algorithm 2 and the stage-k parameters*

$$\varrho_k^2 := \frac{\mathcal{B}}{\bar{\mathcal{A}}_k}, \qquad D_k^2 := 4\bar{\mathcal{A}}_k, \qquad N_k := \left\lceil \sqrt{32L/\mu} \right\rceil,$$
$$b_{p,k} := \max\Big\{1, \Big\lceil \frac{2176\, dM_R^2 D_k^2}{K^{(k)}\varepsilon_k^2} \Big\rceil\Big\}, \qquad b_{y,k} := \max\Big\{1, \Big\lceil \frac{128\, \varrho_k^2 \sigma_y^2 D_k^2}{K^{(k)}\varepsilon_k^2} \Big\rceil\Big\}, \qquad \lambda_k := \frac{\Sigma_{b,k}}{D_k}\sqrt{\frac{A^{(k)}}{2}}, \tag{50}$$

where $\Sigma_{b,k}^2 := 17dM_R^2/b_{p,k} + \varrho_k^2\sigma_y^2/b_{y,k}$ and $K^{(k)}$, $A^{(k)}$ are the quantities K_N, A_N of (28) for $N = N_k$ and $\varrho = \varrho_k$. Then $\mathbb{E}\mathsf{C}_{R,\gamma}(\bar x^{(K)}) \le \varepsilon_{\mathrm{s}}$ and, with $K = \lceil \log_2(\mu D^2/(2\varepsilon_{\mathrm{s}})) \rceil$,

$$N_{\mathrm{seq}} = \sum_{k=1}^{K} N_k \le K\Big(\sqrt{32L/\mu} + 1\Big), \qquad N_{\mathrm{inner}} = \sum_{k=1}^{K} K^{(k)} \le N_{\mathrm{seq}} + 167\, M_g \sqrt{\frac{\mathcal{B}}{\mu\varepsilon_{\mathrm{s}}}}, \tag{51}$$

$$N_f \le 2N_{\mathrm{inner}} + \frac{1.4\cdot 10^5\, dM_R^2}{\mu\varepsilon_{\mathrm{s}}}, \qquad N_{\mathrm{orc}} = N_g \le 7N_{\mathrm{inner}} + \frac{2.8\cdot 10^5\, dM_R^2}{\mu\varepsilon_{\mathrm{s}}} + \frac{8.2\cdot 10^3\, \mathcal{B}\sigma_y^2}{\varepsilon_{\mathrm{s}}^2}, \tag{52}$$

where $\mathcal{B}\sigma_y^2 = 16\Lambda_m R^2(\sigma_h^2 + \nu_m M_g^2 D^2)$. In O-notation,

$$N_{\mathrm{seq}} = O\Big(\sqrt{\tfrac{L}{\mu}}\log\tfrac{\mu D^2}{\varepsilon_{\mathrm{s}}}\Big), \qquad N_{\mathrm{inner}} = O\Big(N_{\mathrm{seq}} + M_g R\sqrt{\tfrac{\log(m+1)}{\mu\varepsilon_{\mathrm{s}}}}\Big),$$
$$N_f = O\Big(N_{\mathrm{inner}} + \frac{dM_R^2}{\mu\varepsilon_{\mathrm{s}}}\Big), \qquad N_{\mathrm{orc}} = N_g = O\Big(N_{\mathrm{inner}} + \frac{dM_R^2}{\mu\varepsilon_{\mathrm{s}}} + \frac{\Lambda_m R^2(\sigma_h^2 + \nu_m M_g^2 D^2)}{\varepsilon_{\mathrm{s}}^2}\Big).$$

The number of calls in which the objective is used thus benefits fully from strong convexity, while the $\varepsilon_{\mathrm{s}}^{-2}$ level term is charged to the constraint calls only; this is what the two batch sizes of Algorithm 2 buy.

Proof. By Theorem 6.5 (conditionally on the starting point), the base method with parameters $(N, b_p, b_y, \lambda, \varrho)$ satisfies (49) with the nondecreasing affine function $\Psi(\mathcal{A}) = W_N^{-1}[(L+2\lambda)(3\mathcal{A} + \mathcal{B}/\varrho^2) + A_N\Sigma_b^2/\lambda]$. At stage k, $3\bar{\mathcal{A}}_k + \mathcal{B}/\varrho_k^2 = 4\bar{\mathcal{A}}_k = D_k^2$, and with λ_k as in (50) the proof of Theorem 6.6 gives $\Psi(\bar{\mathcal{A}}_k) \le 2LD_k^2/(N_k(N_k+1)) + 4D_k\Sigma_{b,k}/\sqrt{K^{(k)}}$. Since $D_k^2 = 4\varepsilon_{k-1}/\mu = 8\varepsilon_k/\mu$, the choice $N_k^2 \ge 32L/\mu = 4LD_k^2/\varepsilon_k$ makes the first term at most $\varepsilon_k/2$, and, exactly as in the proof of Theorem 6.7, $b_{p,k}K^{(k)} \ge 2176dM_R^2D_k^2/\varepsilon_k^2$ and $b_{y,k}K^{(k)} \ge 128\varrho_k^2\sigma_y^2D_k^2/\varepsilon_k^2$ make the second at most $\varepsilon_k/2$. Hence $\Psi(\bar{\mathcal{A}}_k) \le \varepsilon_k$ and Theorem 7.1 applies.

Counts. $N_{\mathrm{seq}} = KN_k \le K(\sqrt{32L/\mu}+1)$. Since $LD_k^2 = 8L\varepsilon_k/\mu \ge \varepsilon_k$ (as $L \ge H \ge \mu$ because a μ-strongly convex L-smooth function has $L \ge \mu$), Theorem 6.7 gives $K^{(k)} \le N_k + 6M_g\varrho_kD_k^2/\varepsilon_k$ with $\varrho_kD_k^2 = 4\sqrt{\mathcal{B}\bar{\mathcal{A}}_k} = 4\sqrt{2\mathcal{B}\varepsilon_k/\mu}$, i.e. $K^{(k)} \le N_k + 24M_g\sqrt{2\mathcal{B}/(\mu\varepsilon_k)}$. Now $\varepsilon_k = \varepsilon_K 2^{K-k}$ with $\varepsilon_K > \varepsilon_{\mathrm{s}}/2$, so $\sum_k \varepsilon_k^{-1/2} \le \varepsilon_K^{-1/2}\sum_{j\ge 0} 2^{-j/2} \le \frac{\sqrt{2}}{1-2^{-1/2}}\varepsilon_{\mathrm{s}}^{-1/2} \le 4.83\,\varepsilon_{\mathrm{s}}^{-1/2}$, $\sum_k \varepsilon_k^{-1} \le 2\varepsilon_K^{-1} \le 4\varepsilon_{\mathrm{s}}^{-1}$ and $\sum_k \varepsilon_k^{-2} \le \frac{4}{3}\varepsilon_K^{-2} \le \frac{16}{3}\varepsilon_{\mathrm{s}}^{-2}$. This gives $N_{\mathrm{inner}} \le N_{\mathrm{seq}} + 24\sqrt{2}\cdot 4.83\, M_g\sqrt{\mathcal{B}/(\mu\varepsilon_{\mathrm{s}})} \le N_{\mathrm{seq}} + 167M_g\sqrt{\mathcal{B}/(\mu\varepsilon_{\mathrm{s}})}$. For the calls, $dM_R^2D_k^2 = 8dM_R^2\varepsilon_k/\mu$ and $\varrho_k^2\sigma_y^2D_k^2 = (\mathcal{B}/\bar{\mathcal{A}}_k)\sigma_y^2\cdot 4\bar{\mathcal{A}}_k = 4\mathcal{B}\sigma_y^2$, so by (47) $N_f^{(k)} \le 2K^{(k)} + 4352\cdot 8\, dM_R^2/(\mu\varepsilon_k) = 2K^{(k)} + 34816\, dM_R^2/(\mu\varepsilon_k)$ and $N_g^{(k)} \le 7K^{(k)} + 8704\cdot 8\, dM_R^2/(\mu\varepsilon_k) + 384\cdot 4\,\mathcal{B}\sigma_y^2/\varepsilon_k^2 = 7K^{(k)} + 69632\, dM_R^2/(\mu\varepsilon_k) + 1536\,\mathcal{B}\sigma_y^2/\varepsilon_k^2$. Summing with $\sum_k \varepsilon_k^{-1} \le 4\varepsilon_{\mathrm{s}}^{-1}$ and $\sum_k \varepsilon_k^{-2} \le \frac{16}{3}\varepsilon_{\mathrm{s}}^{-2}$ gives $34816\cdot 4 \le 1.4\cdot 10^5$, $69632\cdot 4 \le 2.8\cdot 10^5$ and $1536\cdot\frac{16}{3} = 8192 \le 8.2\cdot 10^3$, which is (52). □

The term $\mathcal{B}\sigma_y^2/\varepsilon_{\mathrm{s}}^2$, which appears in $N_g = N_{\mathrm{orc}}$ but not in N_f, does not benefit from strong convexity. Its first part, $\Lambda_m R^2\sigma_h^2/\varepsilon_{\mathrm{s}}^2$, is unavoidable up to the factor $\kappa_m R^2$ for any method, also when the objective is strongly convex (Theorem 9.3 and Theorem 9.5: even deciding

which constraint is active at a known point costs $\Omega(\sigma_g^2 \log(1+m)/\varepsilon^2)$ calls). Its second part, $\Lambda_m R^2 \nu_m M_g^2 D^2/\varepsilon_{\mathrm{s}}^2$, is the price of the linearization discussed in Section 6.6; the restarted Mirror–Prox method below does not have it, at the cost of a larger number of rounds.

7.2 Restarted Mirror–Prox

Theorem 7.3 (R-B-SMP). *Let Theorems 2.1 to 2.3, 2.5 and 2.6 hold, $\gamma \le \bar\gamma$, $\varepsilon_{\mathrm{s}} \in (0, \mu D^2/2)$, and run Algorithm 3 with the base method Algorithm 1 and the stage-k parameters*

$$\begin{aligned} \alpha_k &:= \sqrt{\mathcal{B}/\bar{\mathcal{A}}_k}, \qquad N_k := \Big\lceil 2\sqrt{3}\Big(\frac{4H}{\mu} + 2M_g\sqrt{\frac{2\mathcal{B}}{\mu\varepsilon_k}}\Big)\Big\rceil, \\ b_k &:= \max\Big\{1, \Big\lceil \frac{200}{N_k}\Big(\frac{4dM_R^2}{\mu\varepsilon_k} + \frac{8\kappa_m\sigma_h^2\mathcal{B}}{\varepsilon_k^2}\Big)\Big\rceil\Big\}, \end{aligned} \tag{53}$$

and the step η of (26) with $\mathcal{D}_{\alpha_k}^2$ replaced by $2\sqrt{\mathcal{B}\bar{\mathcal{A}}_k}$. Then $\mathbb{E}\mathsf{C}_{R,\gamma}(\bar x^{(K)}) \le \varepsilon_{\mathrm{s}}$ and, with $K = \lceil \log_2(\mu D^2/(2\varepsilon_{\mathrm{s}}))\rceil$,

$$\begin{aligned} N_{\mathrm{seq}} &= 2\sum_k N_k \le K\Big(\frac{16\sqrt{3}H}{\mu} + 2\Big) + 95\, M_g\sqrt{\frac{\mathcal{B}}{\mu\varepsilon_{\mathrm{s}}}}, \\ N_{\mathrm{orc}} &\le 3N_{\mathrm{seq}} + \frac{2\cdot 10^4\, dM_R^2}{\mu\varepsilon_{\mathrm{s}}} + \frac{5.2\cdot 10^4\,\kappa_m R^2\log(m+1)\,\sigma_h^2}{\varepsilon_{\mathrm{s}}^2}, \end{aligned}$$

i.e. $N_{\mathrm{seq}} = O\big(\frac{H}{\mu}\log\frac{\mu D^2}{\varepsilon_{\mathrm{s}}} + M_g R\sqrt{\log(m+1)/(\mu\varepsilon_{\mathrm{s}})}\big)$ and $N_{\mathrm{orc}} = O\big(N_{\mathrm{seq}} + \frac{dM_R^2}{\mu\varepsilon_{\mathrm{s}}} + \frac{\Lambda_m R^2\sigma_h^2}{\varepsilon_{\mathrm{s}}^2}\big)$; no term proportional to $D^2/\varepsilon_{\mathrm{s}}^2$ appears.

Proof. Fix a stage and write $\alpha = \alpha_k$, $\bar{\mathcal{A}} = \bar{\mathcal{A}}_k$, $\bar{\mathcal{D}}^2 := 2\sqrt{\mathcal{B}\bar{\mathcal{A}}} = \alpha\bar{\mathcal{A}} + \mathcal{B}/\alpha$. By Theorem C.3, for a starting point x_0 with $\mathcal{A} = \frac12\|x_0 - x_\gamma^*\|^2$ and any deterministic step $\eta \le 1/(\sqrt{3}L_\alpha)$,

$$\mathbb{E}[\mathsf{C} \mid x_0] \le \Psi(\mathcal{A}) := \frac{\alpha\mathcal{A} + \mathcal{B}/\alpha}{\eta N} + 3\eta\sigma_*^2 + \sqrt{\frac{2(\alpha\mathcal{A} + \mathcal{B}/\alpha)\sigma_*^2}{N}},$$

a nondecreasing concave function of $\mathcal{A}$. With $\eta := \min\{1/(\sqrt{3}L_\alpha), \bar{\mathcal{D}}/(\sigma_*\sqrt{3N})\}$ the computation in the proof of Theorem 5.1 (with $\mathcal{D}_\alpha$ replaced by $\bar{\mathcal{D}}$) gives $\Psi(\bar{\mathcal{A}}) \le \sqrt{3}L_\alpha\bar{\mathcal{D}}^2/N + 5\sigma_*\bar{\mathcal{D}}/\sqrt{N}$. At stage k, $\bar{\mathcal{A}}_k = 2\varepsilon_k/\mu$, hence $L_{\alpha_k}\bar{\mathcal{D}}_k^2 \le (H/\alpha_k + M_g)\cdot 2\sqrt{\mathcal{B}\bar{\mathcal{A}}_k} = 2H\bar{\mathcal{A}}_k + 2M_g\sqrt{\mathcal{B}\bar{\mathcal{A}}_k} = 4H\varepsilon_k/\mu + 2M_g\sqrt{2\mathcal{B}\varepsilon_k/\mu}$, and $\sigma_*^2\bar{\mathcal{D}}_k^2 \le \frac{1}{b_k}(2dM_R^2/\alpha_k + 8\kappa_m\alpha_k\sigma_h^2)\cdot 2\sqrt{\mathcal{B}\bar{\mathcal{A}}_k} = \frac{4}{b_k}(dM_R^2\bar{\mathcal{A}}_k + 4\kappa_m\sigma_h^2\mathcal{B})$ (using $\alpha_k\sqrt{\mathcal{B}\bar{\mathcal{A}}_k} = \mathcal{B}$ and $\sqrt{\mathcal{B}\bar{\mathcal{A}}_k}/\alpha_k = \bar{\mathcal{A}}_k$). The choices (53) make both terms at most $\varepsilon_k/2$: $\sqrt{3}L_{\alpha_k}\bar{\mathcal{D}}_k^2/N_k \le \varepsilon_k/2$ iff $N_k \ge 2\sqrt{3}(4H/\mu + 2M_g\sqrt{2\mathcal{B}/(\mu\varepsilon_k)})$, and $25\sigma_*^2\bar{\mathcal{D}}_k^2/N_k \le \varepsilon_k^2/4$ iff $b_kN_k \ge 400(dM_R^2\bar{\mathcal{A}}_k + 4\kappa_m\sigma_h^2\mathcal{B})/\varepsilon_k^2 = 200(4dM_R^2/(\mu\varepsilon_k) + 8\kappa_m\sigma_h^2\mathcal{B}/\varepsilon_k^2)$. Theorem 7.1 gives the accuracy claim. The counts follow from $N_{\mathrm{seq}} = 2\sum_k N_k$, $N_{\mathrm{orc}} = 6\sum_k b_kN_k \le 6\sum_k N_k + 1200\sum_k(4dM_R^2/(\mu\varepsilon_k) + 8\kappa_m\sigma_h^2\mathcal{B}/\varepsilon_k^2)$ and the sums $\sum_k\varepsilon_k^{-1/2} \le 4.83\varepsilon_{\mathrm{s}}^{-1/2}$, $\sum_k\varepsilon_k^{-1} \le 4\varepsilon_{\mathrm{s}}^{-1}$, $\sum_k\varepsilon_k^{-2} \le \frac{16}{3}\varepsilon_{\mathrm{s}}^{-2}$ computed in the proof of Theorem 7.2: $2\sum_k N_k \le 2K(8\sqrt{3}H/\mu + 1) + 8\sqrt{6}M_g\sqrt{\mathcal{B}/\mu}\cdot 4.83\,\varepsilon_{\mathrm{s}}^{-1/2}$ with $8\sqrt{6}\cdot 4.83 \le 95$; $6\sum_k N_k = 3N_{\mathrm{seq}}$; $1200\cdot 4\cdot 4\,dM_R^2/(\mu\varepsilon_{\mathrm{s}}) \le 2\cdot 10^4 dM_R^2/(\mu\varepsilon_{\mathrm{s}})$; and $1200\cdot 8\cdot\frac{16}{3}\kappa_m\sigma_h^2\mathcal{B} = 51200\,\kappa_m\sigma_h^2\mathcal{B} \le 5.2\cdot 10^4\kappa_m R^2\log(m+1)\sigma_h^2$. □

8 Exactly known affine constraints: matrix–vector products

When the constraints are affine and known exactly, $g(x) = Ax - c$ with $A \in \mathbb{R}^{m\times d}$ and $c \in \mathbb{R}^m$, the constraint part of the oracle is not needed: $h(x) = Ax - c$ (no smoothing, no shift, $b_i = 0$), $J_t = A$ for all t, and the inner dual directions $q_k = Ap_k - c$ are exact. The only stochastic zeroth-order information concerns the objective. In this situation Algorithm 2 simplifies considerably: since

Algorithm 4 ZO-Sliding-A: sliding with exactly known affine constraints

Require: $L \geq L_{0,\gamma}$, M_A, R, γ, $\varrho > 0$, $\lambda \geq 0$, $b \geq 1$, N; schedule τ_t, θ_t as in (28), $S_t = \max\{1, \lceil M_A \varrho t / L \rceil\}$, $a_t = t/S_t$, $\eta_t = (2L+\lambda)/t$, $\beta_t = L/a_t$; initialization as in Algorithm 2.

for $t = 1, \dots, N$ **do**
  $u_t := [\tau_t u_{t-1} + x_{t-1} + \theta_t(x_{t-1} - x_{t-2})]/(1+\tau_t)$.
  Round t. Issue b paired queries of the objective at u_t and form $\hat{g}_t := \hat{g}_0^{A_t}(u_t)$.
  for $s = 1, \dots, S_t$ **do**
    $k := k+1$; $\hat{r}_k := \hat{g}_t + A^\top y_{k-1} + \rho_k A^\top (y_{k-1} - y_{k-2})$ with $\rho_k = 1$ for $s \geq 2$, $\rho_k = a_{t-1}/a_t$ for $s = 1$, $t \geq 2$ (no last term for $k = 1$).
    $p_k := \arg\min_{x \in Q}\{\langle \hat{r}_k, x\rangle + \eta_t V_{\mathcal{X}}(x_{t-1}, x) + \beta_t V_{\mathcal{X}}(p_{k-1}, x)\}$; $q_k := Ap_k - c$;
    $\tilde{y}_k := \arg\max_{\tilde{y} \in \tilde{Y}_R}\{\langle (0, q_k), \tilde{y}\rangle - \varrho^{-2}\beta_t V_{\mathcal{Y}}(\tilde{y}_{k-1}, \tilde{y})\}$.
  end for
  $x_t := \frac{1}{S_t}\sum_k p_k$ over the slots of phase t.
end for

Ensure: $\bar{x}_N := W_N^{-1} \sum_t t\, x_t$.

the dual step has no noise, the inner stabilizer is not needed, and since the Jacobian is exact, *one* mini-batch of paired queries per phase suffices for the primal directions (the reuse of the objective gradient estimate within a phase is harmless, as in accelerated stochastic gradient methods, because the primal error only has to be centered). Each inner step then costs one multiplication by $A^\top$ and one by A, which is the matrix–vector product count of ACGD-S [64]; here the inner steps are oracle-free in the literal sense. In the accounting of Theorem 2.8, round t issues $2b$ vector calls, all of them used for the objective only, so that $N_{\mathrm{orc}} = N_f = 2bN$ and $N_g = 0$. We write $M_A := \|A\|_{2\to\infty}$.

Theorem 8.1 (ZO-Sliding-A)**.** *Let Theorems 2.1, 2.2 and 2.5 hold for the problem with $g(x) = Ax - c$, let $\gamma \leq \bar{\gamma}$, R as in (14), $L \geq L_{0,\gamma}$ and $\lambda > 0$. Then the output of Algorithm 4 satisfies*

$$W_N\, \mathbb{E}\, \mathsf{C}_{R,\gamma}(\bar{x}_N) \leq (3L+\lambda)\mathcal{A} + \frac{L\mathcal{B}}{\varrho^2} + \frac{\sigma_f^2}{2\lambda b}\sum_{t=1}^{N} t^2, \qquad \sigma_f^2 := 2dM_0^2, \tag{54}$$

and with $\lambda := \sigma_f \sqrt{\sum_t t^2/(2b\mathcal{A})}$

$$\mathbb{E}\, \mathsf{C}_{R,\gamma}(\bar{x}_N) \leq \frac{2LD_A^2}{N(N+1)} + 2\sigma_f\sqrt{\frac{\mathcal{A}}{b(N+1)}}, \qquad D_A^2 = 3\mathcal{A} + \frac{\mathcal{B}}{\varrho^2}. \tag{55}$$

Proof. We specialize the argument of Theorem 6.4 to the affine case; the details are given in Section D.5. The dual error vanishes, $e_k^y = 0$, and the primal error of every slot of phase t equals $\hat{e}_t := \hat{g}_t - \nabla f_\gamma(u_t)$, which is independent of $\mathcal{F}_{t-1}$ (the information before round t) with $\mathbb{E}\hat{e}_t = 0$ and $\mathbb{E}\,\|\hat{e}_t\|^2 \leq \sigma_f^2/b$ (Theorem 3.3). With $a_k\beta_k = L$ the inner primal terms telescope to $L[V_{\mathcal{X}}(p_0, x) - V_{\mathcal{X}}(p_K, x)] - L\sum_k V_{\mathcal{X}}(p_{k-1}, p_k)$, which is exactly what is needed to absorb the bilinear cross terms (they need $L/2$ per increment, see (57)); the dual terms telescope to $\varrho^{-2}L[V_{\mathcal{Y}}(\tilde{y}_0, \tilde{y}) - V_{\mathcal{Y}}(\tilde{y}_K, \tilde{y}) - \sum_k V_{\mathcal{Y}}(\tilde{y}_{k-1}, \tilde{y}_k)]$; and the outer terms, with $t\eta_t = 2L + \lambda$, telescope to $(2L+\lambda)[V_{\mathcal{X}}(x_0, x) - V_{\mathcal{X}}(x_N, x)] - (L + \frac{\lambda}{2})\sum_t \|d_t\|^2$. Hence, pathwise for every $y \in Y_R$,

$$\sum_t t\big[\ell_t(x_t, y) - \ell_t(x_\gamma^*, \bar{y}_t)\big] + \Big(L + \frac{\lambda}{2}\Big)\sum_t \|d_t\|^2 \leq (3L+\lambda)\mathcal{A} + \frac{L\mathcal{B}}{\varrho^2} - \sum_k a_k \left\langle \hat{e}_{t(k)}, p_k - x_\gamma^* \right\rangle.$$

The error sum equals $\sum_t t\left\langle \hat{e}_t, x_t - x_\gamma^*\right\rangle$ because $\sum_{k \in \text{phase } t} a_t p_k = t x_t$. Splitting $x_t - x_\gamma^* = d_t + (x_{t-1} - x_\gamma^*)$ and using Young's inequality, $-t\,\langle \hat{e}_t, d_t\rangle \leq \frac{t^2}{2\lambda}\|\hat{e}_t\|^2 + \frac{\lambda}{2}\|d_t\|^2$, where the increment

term is cancelled by the left-hand side. The remaining error terms $-t\langle \hat{e}_t, x_{t-1} - x^*_\gamma \rangle$ have zero mean because x_{t-1} is $\mathcal{F}_{t-1}$-measurable. Since the right-hand side no longer depends on y, taking the supremum over $y \in Y_R$, using Theorem 6.3 for $\mathsf{F} = \Phi_\gamma(\cdot, y)$ (which is L-smooth because $L_{g,\gamma} = 0$) and (19) exactly as in Step 4 of the proof of Theorem 6.5, and taking expectations gives (54). Minimizing $\lambda\mathcal{A} + \sigma_f^2 \sum_t t^2/(2\lambda b)$ over λ gives $\sigma_f\sqrt{2\mathcal{A}\sum_t t^2/b}$; since $\sum_{t=1}^N t^2 = \frac{N(N+1)(2N+1)}{6} \le NW_N$ and $3L\mathcal{A} + L\mathcal{B}/\varrho^2 \le LD_A^2$, dividing by W_N yields (55). □

Corollary 8.2 (complexity of ZO-SLIDING-A)**.** *In the setting of Theorem 8.1, let* $\varepsilon_{\mathrm{s}} \in (0, LD_A^2]$ *and choose* $N := \lceil 2D_A\sqrt{L/\varepsilon_{\mathrm{s}}}\rceil$, $b := \max\{1, \lceil 16\sigma_f^2\mathcal{A}/((N+1)\varepsilon_{\mathrm{s}}^2)\rceil\}$ *(with* $\mathcal{A}$ *replaced by the known bound* $D^2/2$ *if necessary). Then* $\mathbb{E}\mathsf{C}_{R,\gamma}(\bar{x}_N) \le \varepsilon_{\mathrm{s}}$*,*

$$N_{\mathrm{seq}} = N \le 2D_A\sqrt{\frac{L}{\varepsilon_{\mathrm{s}}}} + 1, \qquad N_{\mathrm{orc}} = N_f = 2bN \le 2N + \frac{64\, dM_0^2\mathcal{A}}{\varepsilon_{\mathrm{s}}^2},$$
$$N_g = 0, \qquad N_{\mathrm{inner}} = 2K_N \le 2N + \frac{12\, M_A\varrho D_A^2}{\varepsilon_{\mathrm{s}}}.$$

With $\varrho = \sqrt{2\mathcal{B}}/D$*:* $D_A^2 \le 2D^2$ *and* $N_{\mathrm{inner}} \le 2N + 34\, M_A RD\sqrt{\log(m+1)}/\varepsilon_{\mathrm{s}}$*. For a nonsmooth objective (*$\gamma = \varepsilon/(4M_0)$*,* $L = \sqrt{d}M_0/\gamma$*,* $\varepsilon_{\mathrm{s}} = \varepsilon/2$*,* $\varepsilon \le 2d^{1/4}M_0D$ *so that* $\varepsilon_{\mathrm{s}} \le LD_A^2$*) this gives* $N_{\mathrm{seq}} \le 8d^{1/4}M_0D/\varepsilon + 1$ *and* $N_f = O(N_{\mathrm{seq}} + dM_0^2D^2/\varepsilon^2)$*; for an* L_0*-smooth objective (*γ *arbitrarily small,* $L = L_0$*,* $\varepsilon \le L_0D^2$*)* $N_{\mathrm{seq}} \le 4D\sqrt{L_0/\varepsilon} + 1$*.*

Proof. As in Theorem 6.7: the two terms of (55) are at most $\varepsilon_{\mathrm{s}}/2$ each, $K_N \le N + \frac{M_A\varrho}{2L}N(N+1) \le N + 6M_A\varrho D_A^2/\varepsilon_{\mathrm{s}}$ (each inner step performs two matrix–vector products), $2bN \le 2N + 32\sigma_f^2\mathcal{A}N/((N+1)\varepsilon_{\mathrm{s}}^2) \le 2N + 32\sigma_f^2\mathcal{A}/\varepsilon_{\mathrm{s}}^2$, and $6\sqrt{2}\cdot 2 \le 17\cdot 2$. □

The objective-call complexity $O(dM_0^2D^2/\varepsilon^2)$ and the round complexities $O(\sqrt{L_0D^2/\varepsilon})$ and $O(d^{1/4}M_0D/\varepsilon)$ are those of the *unconstrained* problem [26]; the constraints cost only $O(M_ARD\sqrt{\log(m+1)}/\varepsilon)$ matrix–vector products, which is the ACGD-S accounting [64, Cor. 3] with the Euclidean dual norm replaced by the entropy geometry (whence $\sqrt{\log(m+1)}$ instead of $\sqrt{m}$ for the dual radius). The order $O(1/\varepsilon)$ of N_{inner} cannot be improved by methods that access A only through matrix–vector products [51]; that lower bound concerns the exponent of ε and does not address the exact dependence on M_A, R, D and m.

9 Lower bounds

We now show that the ε^{-2} terms in the call complexities of Sections 6 and 7 are unavoidable, and we identify which ingredients are responsible for them: the dimension factor d for the two-point information about gradients (objective *and* constraints), and the factor $\sigma_g^2\log(1+m)$ for the information about the constraint levels. All lower bounds are minimax statements over classes of problems satisfying Theorems 2.1 to 2.3 and 2.5 (with the indicated constants); they hold for arbitrary, possibly randomized and adaptive, algorithms whose output $\hat{x}$ is any measurable function of the observations. Theorems 9.1 and 9.2 count paired queries (one paired query is two vector calls, so the bounds hold for N_{orc} with the constant halved), Theorem 9.3 counts vector calls and Theorem 9.4 scalar calls. The proofs are given in Section E.

9.1 Two-point calls for the objective

Theorem 9.1 (objective information)**.** *Let* $d \ge 1$*,* $M > 0$*,* $R_x > 0$*,* $Q = R_x\mathbb{B}$*,* $m \ge 1$ *and* $T \ge 1$*. For every algorithm that makes at most* T *paired queries and outputs* $\hat{x} \in Q$ *there is a problem in the class of linear objectives* $F(x,\xi) = \langle \xi, x\rangle$ *with* $\xi \sim \mathcal{N}(\theta, s^2I_d)$*,* $\|\theta\| \le M/2$*,* $s = M/\sqrt{2d}$

(which satisfy Theorem 2.2 with $M_0 = M$) and inactive constraints $G_i \equiv -1$ (Theorem 2.3 with $\sigma_g = 0$, Theorem 2.5 with $\rho = 1$) such that

$$\mathbb{E}\big[f(\hat{x}) - f^*\big] \;\geq\; \frac{MR_x}{32}\,\min\Big\{1, \sqrt{\frac{d}{T}}\Big\}.$$

Consequently, an ε-solution in expectation with $\varepsilon < MR_x/32$ requires $T \geq dM^2R_x^2/(1024\varepsilon^2)$ paired queries.

The proof is Assouad's method with d binary parameters and the chain rule for the Kullback–Leibler divergence; it is the two-point (and thus noise-cancelling) version of the argument of Duchi et al. [22], Shamir [56]: even though the difference $F(x^+,\xi) - F(x^-,\xi) = \langle \xi, x^+ - x^-\rangle$ removes the noise in the value, a paired query reveals only the projection of the unknown mean onto the span of the two query points, a linear measurement of rank at most two, and d coordinates of the mean have to be estimated to accuracy ε/R_x. Together with Theorem 8.2 this shows that the term $dM_0^2D^2/\varepsilon^2$ in the objective calls is optimal up to constants, and in the affine case (Theorem 8.1) the whole objective complexity is optimal up to the lower-order term N_{seq}; in Theorem 6.10 the objective and the constraint information are estimated jointly and enter through $M_R = M_0 + RM_g$, so there the match is up to the multiplier factor.

9.2 Two-point calls for the constraints

Theorem 9.2 (constraint information)**.** *Let $d \geq 1$, $M > 0$, $R_x > 0$, $T \geq 1$. Consider problems in the variables $z = (x,t) \in Q' := R_x\mathbb{B} \times [-2R_x, 2R_x]$ with the known deterministic objective $f(z) = Mt$ and the single stochastic constraint $G(z,\xi) = \langle \xi, x\rangle - Mt$, $\xi \sim \mathcal{N}(\theta, s^2 I_d)$, $\|\theta\| \leq M/2$, $s = M/\sqrt{2d}$. The class has diameter $D = 2\sqrt{5}R_x$ and satisfies Theorem 2.2 with $M_0 = M$ and $M_g = \frac{\sqrt{7}}{2}M$ (since $\mathbb{E}\,\|(\xi,-M)\|^2 \leq \frac{3}{4}M^2 + M^2$), Theorem 2.3 on $Q'_{\bar\gamma}$ with $\sigma_g^2 = s^2(R_x + \bar\gamma)^2$, and Theorem 2.5 with $\rho = MR_x$ at $z = (0, R_x)$; the active multiplier equals 1 in every problem of the class. For every algorithm that makes at most T paired queries of the constraint and outputs $\hat{z} \in Q'$ there is a problem in this class with*

$$\mathbb{E}\Big[\big(f(\hat{z}) - f^*\big) + [g(\hat{z})]_+\Big] \;\geq\; \frac{MR_x}{32}\,\min\Big\{1, \sqrt{\frac{d}{T}}\Big\} = \frac{M_gD}{32\sqrt{35}}\,\min\Big\{1, \sqrt{\frac{d}{T}}\Big\}.$$

Consequently, an ε-solution in expectation in the sense of (3) with $\varepsilon < MR_x/64$ requires $T \geq dM^2R_x^2/(4096\,\varepsilon^2) = dM_g^2D^2/(143360\,\varepsilon^2)$ paired queries.

The reduction is a scaled epigraph transformation, $\min_x \varphi(x) = \min\{Mt : \varphi(x) \leq Mt\}$: an algorithm for the constrained problem is an algorithm for minimizing the unknown linear function $\varphi(x) = \langle \theta, x\rangle$ with the same information, and $\varphi(\hat{x}) - \varphi^* \leq (M\hat{t} - f^*) + [g(\hat{z})]_+$. The scaling of the epigraph variable by M is what makes the constants universal: it keeps the diameter of Q' of the order of R_x and the Lipschitz constant of the constraint of the order of M, so that the lower bound can be expressed in the parameters M_g and D of the class itself (an unscaled epigraph $\{\vartheta : \varphi(x) \leq \vartheta\}$ would have diameter of order MR_x and Lipschitz constant $\sqrt{M^2+1}$, and the resulting bound would not be of the form $dM_g^2D^2/\varepsilon^2$ uniformly in M). (The bound is stated for the sum of gap and violation because $f(\hat{z}) - f^* = M\hat{t} - \varphi^*$ may be negative at infeasible points; (3) bounds the expectation of this sum by 2ε.) Thus the two-point information about *constraint* gradients has the same $dM_g^2D^2/\varepsilon^2$ price as the information about the objective, with universal constants, and the term $dR^2M_g^2D^2/\varepsilon^2$ of $M_R^2 = (M_0 + RM_g)^2$ in Theorem 6.10 cannot be removed in general: the upper bound matches the lower bound in d, M_g, D and ε. The multiplier factor R^2 is not addressed by this construction, in which the active multiplier equals 1; see Theorem 12.4.

9.3 Calls for the constraint levels

The third source of ε^{-2} complexity is the noise in the constraint *values*, which does not cancel in differences. The following theorems concern the very simple one-dimensional problem of finding, among m constraints of the form $x \le \beta_i$, the one with the smallest level. They show that the entropy dual geometry, which produces the factors $\log(m+1)$ in Theorems 5.1 and 6.5, is not an artifact: $\log(1+m)$ vector calls are necessary, and if a call returns only one component, m calls are necessary.

Theorem 9.3 (level information, vector oracle)**.** *Let $m \ge 1$, $\sigma > 0$ and $0 < \varepsilon \le \min\{\sigma/16,\, 1/16\}$. Consider the one-dimensional problems with $Q = [-1, 1]$ (so $D = 2$), known objective $f(x) = -x$, and constraints $G_i(x, \xi) = x - \xi_i$, where $\xi = (\xi_1, \dots, \xi_m)$ has independent coordinates taking the values $\pm\sigma$ with $\mathbb{E}\xi_i = \beta_i \in \{-a, a\}$, $a := 8\varepsilon \le 1/2$; the class consists of the $m+1$ vectors $\beta^{(0)} = (a, \dots, a)$ and $\beta^{(j)} = \beta^{(0)} - 2ae_j$, $j = 1, \dots, m$ (Theorem 2.2 with $M_0 = M_g = 1$, Theorem 2.3 with $\sigma_g = 2\sigma$, Theorem 2.5 at $x^{\mathrm{s}} = -1$ with $\rho = 1 - a \ge 1/2$; the active multiplier is $y^* = 1$ in every problem of the class, so $R = 2$ is admissible in (14)). Any algorithm that makes at most T vector calls and outputs $\hat{x}$ with $\mathbb{E}[f(\hat{x}) - f^*] \le \varepsilon$ and $\mathbb{E}\mathsf{v}(\hat{x}) \le \varepsilon$ for every problem of the class satisfies*

$$T \;\ge\; \frac{3\,\sigma^2 \log(1+m)}{1024\,\varepsilon^2} = \frac{3\,\sigma_g^2 \log(1+m)}{4096\,\varepsilon^2}.$$

Theorem 9.4 (level information, scalar oracle)**.** *In the setting of Theorem 9.3, suppose that each call returns a single component $G_i(x, \xi)$ for an index i chosen by the algorithm (a fresh ξ per call). Then any algorithm with the same guarantee satisfies $\mathbb{E}_0[T] \ge 25\, m\sigma^2/(8192\,\varepsilon^2) \ge m\sigma^2/(328\,\varepsilon^2)$, where T is the (possibly random) total number of calls and $\mathbb{E}_0$ is the expectation under $\beta^{(0)}$.*

Remark 9.5 (strongly convex objectives). The level information remains an ε^{-2} bottleneck when the objective is strongly convex. Replace $f(x) = -x$ in Theorem 9.3 by $f(x) = -x + \frac{\mu}{2}x^2$ with $\mu \in (0, 1]$, which is μ-strongly convex and decreasing on $[-1, 1]$, and take $a := 16\varepsilon$ with $0 < \varepsilon \le \min\{\sigma/32, 1/32\}$. Then every algorithm with the guarantee of Theorem 9.3 satisfies $T \ge 3\sigma^2 \log(1+m)/(4096\,\varepsilon^2)$, and in the scalar model of Theorem 9.4 $\mathbb{E}_0[T] \ge 25m\sigma^2/(32768\,\varepsilon^2)$. The proof is given in Section E.

Theorem 9.3 is proved with a mixture argument (χ^2-divergence between the uniform mixture of the m alternatives and the null hypothesis); Theorem 9.4 with the chain rule for the Kullback–Leibler divergence applied to the number of observations of each coordinate. In both cases the optimization guarantee is converted into a statistical test: a solution with gap and violation below ε reveals the sign of the smallest level.

9.4 Depth

The lower bounds above concern the number of calls. Lower bounds on the number of *rounds* of adaptive queries are much harder to obtain and are known only for the unconstrained problem: for nonsmooth M-Lipschitz convex minimization on the unit ball with polynomially many gradient (or value) queries per round, no algorithm can guarantee accuracy ε in fewer than $\tilde{\Omega}(\min\{\varepsilon^{-2}, d^{1/3}\varepsilon^{-2/3}\})$ rounds [7, 14, 19, 44]. The best known upper bounds are $O(d^{1/4}\varepsilon^{-1})$ by smoothing [21, 26] and, in the regime $\varepsilon \lesssim d^{-1/4}$, $\tilde{O}(d^{1/3}\varepsilon^{-2/3})$ by ball acceleration [14, 15]; the latter methods use (stochastic) gradient queries, and whether they can be implemented with a two-point value oracle without losing their depth is open. Since an unconstrained problem is a special case of (1) (with inactive constraints), the lower bounds apply verbatim to N_{seq} in our setting, and Theorem 6.10 matches the best known depth of the unconstrained two-point problem up to the factor $\sqrt{M_R/M}$; it is not a minimax statement. For smooth problems, $\Omega(\sqrt{L_0 D^2/\varepsilon})$ rounds are necessary even with exact gradients [47], so the depth of Theorem 6.9 is optimal up to the factor $\sqrt{(L_0 + RL_g)/L_0}$. The smoothness of the constraints cannot simply be dropped from this factor: for the scaled epigraph problem $\min\{Mt : \varphi(x) - Mt \le 0,\ x \in R_x\mathbb{B},\ |t| \le 2R_x\}$ of

Table 4: Complexity bounds for an ε-solution in expectation. $H = L_0 + RL_g$, $M_R = M_0 + RM_g$, $M = \max\{M_0, M_g\}$, $\Lambda_m = (1 + \log 2m)\log(m+1)$; σ_g^2 bounds the variance of the constraint values; $\nu_m = \min\{4d, 92\kappa_m\} = O(\min\{d, 1+\log m\})$. The three call channels of Theorem 2.8 are listed separately: N_f (calls in which the objective is used), N_g (calls in which the constraint vector is used) and N_{orc} (all calls); the call columns list the terms beyond $O(N_{\mathrm{inner}})$; "$N_f + \dots$" in the N_g column means the N_f term plus the listed one. For Algorithms 1 and 2 every call carries the constraint vector, so $N_{\mathrm{orc}} = N_g$; for Algorithm 4, $N_g = 0$ and $N_{\mathrm{orc}} = N_f$.

Result	Data	N_{seq}	N_f	N_g $(= N_{\mathrm{orc}})$	N_{inner}
Theorem 5.3(a)	smooth	$\frac{HD^2}{\varepsilon} + \frac{M_g RD\sqrt{\log m}}{\varepsilon}$	$\frac{dM_R^2D^2+\Lambda_m R^2\sigma_g^2}{\varepsilon^2}$	$= N_f$	N_{seq}
Theorem 5.3(b)	nonsmooth	$\frac{\sqrt{d}M_R MD^2}{\varepsilon^2} + \frac{M_g RD\sqrt{\log m}}{\varepsilon}$	$\frac{dM_R^2D^2+\Lambda_m R^2\sigma_g^2}{\varepsilon^2}$	$= N_f$	N_{seq}
Theorem 6.9	smooth	$\sqrt{\frac{HD^2}{\varepsilon}}$	$\frac{dM_R^2D^2}{\varepsilon^2}$	$N_f + \frac{\Lambda_m R^2(\sigma_g^2+\nu_m M_g^2D^2)}{\varepsilon^2}$	$N_{\mathrm{seq}} + \frac{M_g RD\sqrt{\log m}}{\varepsilon}$
Theorem 6.10	nonsmooth	$\frac{d^{1/4}\sqrt{M_R MD}}{\varepsilon}$	$\frac{dM_R^2D^2}{\varepsilon^2}$	$N_f + \frac{\Lambda_m R^2(\sigma_g^2+\nu_m M_g^2D^2)}{\varepsilon^2}$	$N_{\mathrm{seq}} + \frac{M_g RD\sqrt{\log m}}{\varepsilon}$
Theorem 7.2	μ-s.c., smooth	$\sqrt{\frac{H}{\mu}}\log\frac{\mu D^2}{\varepsilon}$	$\frac{dM_R^2}{\mu\varepsilon}$	$N_f + \frac{\Lambda_m R^2(\sigma_g^2+\nu_m M_g^2D^2)}{\varepsilon^2}$	$N_{\mathrm{seq}} + M_g R\sqrt{\frac{\log m}{\mu\varepsilon}}$
Theorem 7.2	μ-s.c., nonsmooth	$d^{1/4}\sqrt{\frac{M_R M}{\mu\varepsilon}}\log\frac{\mu D^2}{\varepsilon}$	$\frac{dM_R^2}{\mu\varepsilon}$	$N_f + \frac{\Lambda_m R^2(\sigma_g^2+\nu_m M_g^2D^2)}{\varepsilon^2}$	$N_{\mathrm{seq}} + M_g R\sqrt{\frac{\log m}{\mu\varepsilon}}$
Theorem 7.3	μ-s.c., smooth	$\frac{H}{\mu}\log\frac{\mu D^2}{\varepsilon} + M_g R\sqrt{\frac{\log m}{\mu\varepsilon}}$	$\frac{dM_R^2}{\mu\varepsilon} + \frac{\Lambda_m R^2\sigma_g^2}{\varepsilon^2}$	$= N_f$	N_{seq}
Theorem 8.2	affine g, smooth f	$\sqrt{\frac{L_0D^2}{\varepsilon}}$	$\frac{dM_0^2D^2}{\varepsilon^2}$ $(= N_{\mathrm{orc}})$	0	$N_{\mathrm{seq}} + \frac{\|A\|_{2\to\infty} RD\sqrt{\log m}}{\varepsilon}$
Theorem 8.2	affine g, nonsmooth f	$\frac{d^{1/4}M_0D}{\varepsilon}$	$\frac{dM_0^2D^2}{\varepsilon^2}$ $(= N_{\mathrm{orc}})$	0	$N_{\mathrm{seq}} + \frac{\|A\|_{2\to\infty} RD\sqrt{\log m}}{\varepsilon}$

Theorem 9.2 with an M-Lipschitz, L_g-smooth convex φ (here $L_0 = 0$, the active multiplier is 1, and the diameter is $2\sqrt{5}R_x = \Theta(D)$), an ε-solution in the sense of (3) is a 2ε-minimizer of φ over $R_x\mathbb{B}$ in expectation, so $\Omega(\sqrt{L_gD^2/\varepsilon})$ rounds are necessary as well. Whether the coupling constant M_gR must appear in the depth of any zeroth-order method, as it does in N_{inner}, is open (Theorem 12.3).

10 Summary of the rates

Table 4 collects the complexity bounds of Sections 5 to 8 for the target accuracy ε of the original problem (1) (the smoothed target is $\varepsilon_{\mathrm{s}} = \varepsilon/2$ and the biases are at most $\varepsilon/4$). Constants are omitted; the exact statements with constants are the theorems and corollaries referenced in the first column. Table 5 confronts the ε^{-2} terms with the lower bounds of Section 9.

Three structural conclusions can be drawn. First, for both smooth and nonsmooth data the sliding method has the round complexity of the unconstrained accelerated batched methods, with the constraints entering only through the smoothness H of the Lagrangian and the multiplier scale R; the coupling constant M_gR appears only in the number of inner steps, which need no new oracle rounds. Second, the ε^{-2} call complexity is intrinsic and is composed of two different pieces, the gradient information (with the factor d) and the level information (with the factor $\log(1+m)$); the two pieces are charged to different channels, N_f paying only for the gradient information and N_g for both, which is visible in the strongly convex case, where $N_f = O(N_{\mathrm{inner}} + dM_R^2/(\mu\varepsilon))$ while N_g keeps the ε^{-2} level term that Theorem 9.5 shows to be necessary. Third, Table 5 states precisely in which parameters the lower bounds match: the objective term is matched up to constants, the constraint-gradient term in all parameters except the multiplier factor R^2, and the level term up to $\kappa_m R^2$, where κ_m stems from the ℓ_∞ bound on the dual noise (Theorem 3.6) and the entropy radius $\mathcal{B}$ contributes the necessary $\log(m+1)$. This is an optimality map rather than a blanket optimality statement: the dependence on the size of the multipliers, the extra κ_m, the linearization term and the nonsmooth depth are open (Section 12.1), and only the exponent of ε is known to be optimal for the depth and for the inner operations.

Table 5: Upper bounds of this paper against the lower bounds of Section 9. In Theorem 9.1 the class is a ball of radius R_x and $D = 2R_x$; in Theorem 9.2 the bound is stated directly in the diameter D and the Lipschitz constant M_g of the class; in all lower-bound constructions the active multiplier equals 1. "Matching" means matching up to universal constants in the listed parameters. The depth rows record what is known about N_{seq}.

Resource	Upper bound (this paper)	Lower bound	Status
objective gradient, two-point calls	$dM_0^2D^2/\varepsilon^2$ (Theorem 8.2)	dM^2D^2/ε^2 (Theorem 9.1)	matching up to constants
constraint gradients, two-point calls	$dR^2M_g^2D^2/\varepsilon^2$ (Theorem 6.10)	$dM_g^2D^2/\varepsilon^2$ (Theorem 9.2)	matching in d, M_g, D, ε at unit multipliers; factor R^2 open
constraint levels, vector oracle	$\Lambda_m R^2\sigma_g^2/\varepsilon^2$ (Theorems 5.1 and 6.5)	$\sigma_g^2\log(1+m)/\varepsilon^2$ (Theorem 9.3, Theorem 9.5)	one log necessary; factor $\kappa_m R^2$ open
constraint levels, scalar oracle	$m\Lambda_m R^2\sigma_g^2/\varepsilon^2$ (one vector call $= m$ scalar calls)	$m\sigma_g^2/\varepsilon^2$ (Theorem 9.4)	linear in m necessary; factor $\Lambda_m R^2$ open
linearization term (Theorem 6.5)	$\Lambda_m R^2\nu_m M_g^2D^2/\varepsilon^2$	—	dimension-free; below the objective term if $d \geq 92\kappa_m^2\log(m+1)$
depth, smooth data	$\sqrt{(L_0 + RL_g)D^2/\varepsilon}$ (Theorem 6.9)	$\sqrt{L_0D^2/\varepsilon}$, and $\sqrt{L_gD^2/\varepsilon}$ at unit multipliers (Section 9.4)	order in ε optimal; factor $\sqrt{1 + RL_g/L_0}$ open
depth, nonsmooth data	$d^{1/4}\sqrt{M_R M}D/\varepsilon$ (Theorem 6.10)	$\tilde{\Omega}(\min\{\varepsilon^{-2}, d^{1/3}\varepsilon^{-2/3}\})$, unconstrained (Section 9.4)	best known two-point depth up to $\sqrt{M_R/M}$; not minimax
inner operations, affine g	$\|A\|_{2\to\infty} RD\sqrt{\log(m+1)}/\varepsilon$ (Theorem 8.2)	$\Omega(1/\varepsilon)$ for matrix–vector methods [51]	order in ε optimal; parameter dependence open

11 Numerical experiments

The experiments have three purposes: to observe the separation between rounds and calls predicted by Theorems 5.1 and 6.5, to illustrate the restart scheme of Section 7, and to visualize the two information-theoretic effects behind the analysis (the γ-independence of the common-sample two-point estimator, Theorem 3.3, and the $\log(1+m)$ scaling of the level information, Theorem 9.3). All numbers in this section are produced by the code distributed with the paper from stored raw records; nothing is entered by hand. Details on the implementation, the exact parameter values and the reproducibility protocol are given in Section F.

11.1 Test problems

We use $d = 24$, $m = 8$ and $Q = \mathbb{B}$ (unit ball, $D = 2$). Two objectives are considered: the *smooth* quadratic $f(x) = \frac{1}{2}\|x - a\|^2$ with $a = 0.75\,e_1$, which is 1-smooth and 1-strongly convex, and the *nonsmooth* "cusp" $f(x) = -x_1 + \frac{1}{4}\|x_{2:d}\|$, which is not differentiable on the line $x_{2:d} = 0$ through the solution. The constraints are $g_1(x) = x_1 - 0.2$ (active at the solution), $g_i(x) = \langle A_i, x\rangle - 0.3$ for $i = 2, \dots, m-1$ with random unit rows $A_i \perp e_1$ (inactive), and $g_m(x) = \frac{1}{2}\sum_{j=2}^{11} x_j^2 - 0.05$ (smooth, inactive). Both problems have the solution $x^* = 0.2\,e_1$ with $f^* = 0.15125$ (smooth) and $f^* = -0.2$ (cusp) and the single active multiplier $y_1^* = 0.55$, respectively 1; we use $R = 2$. Both instances satisfy the standing assumptions in the form stated in Section 2: the data are convex on $\mathbb{R}^d$ and Lipschitz on $Q_{\bar{\gamma}}$ (Theorem 2.2, with the constants of Table 7 computed on the ball of radius $1+\bar{\gamma}$), the quadratic objective and g_m are 1-smooth on $\mathbb{R}^d$ (Theorem 2.4), and the quadratic objective is 1-strongly convex (Theorem 2.6); no extension or modification of the instance is needed, and the cusp objective, which is only Lipschitz, is handled by Theorem 3.2, which does not change any oracle answer. The oracle returns $F(x,\xi) = f(x) + 0.02\,s_0\,\langle q_0, x\rangle$ and $G_i(x,\xi) = g_i(x) + 0.02\,s_i\,\langle q_i, x\rangle + 0.02\,o_i$ with independent Rademacher variables $s_0, \dots, s_m, o_1, \dots, o_m$ and fixed random unit vectors q_i: the slope noise perturbs the Lipschitz constants of the samples (Theorem 2.2), the offset noise perturbs the constraint levels (Theorem 2.3) and does not cancel

Table 6: Final results at equal budgets (30 seeds). N_{orc} is the number of vector calls (for both methods every call carries the constraint vector, so $N_g = N_{\text{orc}}$), N_f the number of calls whose objective component is used ($2b_pK_N$ for ZO-Sliding, $4bN$ for B-SMP), N_{seq} the number of sequential oracle rounds. Means over seeds of the objective gap, the maximal violation and the joint error J at the last iterate; the last column is a 95% percentile bootstrap confidence interval for the mean of J.

Instance	Method	N_{orc}	N_f	N_{seq}	f-gap	Violation	J	95% CI for J
Smooth	ZO-Sliding, $b = 8$, $\lambda = 0.1$	28,064	10,848	60	0.0016	0.0000	0.0016	[0.0013, 0.0020]
Smooth	ZO-Sliding, $b = 8$, worst-case λ	28,064	10,848	60	-0.0632	0.3355	0.3355	[0.3341, 0.3369]
Smooth	B-SMP, $b = 1$	28,062	18,708	9354	-0.0403	0.0806	0.0806	[0.0775, 0.0838]
Smooth	B-SMP, $b = 8$	28,032	18,688	1168	0.0014	0.0000	0.0014	[0.0013, 0.0015]
Cusp	ZO-Sliding, $b = 8$, $\lambda = 0.1$	6,824	2,352	60	0.0078	0.0093	0.0228	[0.0186, 0.0275]
Cusp	ZO-Sliding, $b = 8$, worst-case λ	6,824	2,352	60	-0.7021	0.7422	0.7422	[0.7409, 0.7434]
Cusp	B-SMP, $b = 1$	6,822	4,548	2274	0.0363	0.0029	0.0363	[0.0324, 0.0405]
Cusp	B-SMP, $b = 8$	6,816	4,544	284	0.0856	0.0000	0.0856	[0.0759, 0.0950]

in differences. The smoothing radii are $\gamma = 0.04$ (smooth) and $\gamma = 0.08$ (cusp); the smoothing shift of g_m is its exact bias $\gamma^2 \cdot 10/(2(d+2))$, all other constraints are affine and need no shift. The joint error of a point x is $J(x) := \max\{0,\, f(x) - f^*,\, \mathsf{v}(x)\}$; note that $f(x) - f^*$ may be negative for infeasible points, which is why both quantities are reported.

11.2 Rounds versus calls

We compare ZO-Sliding (Algorithm 2) with B-SMP (Algorithm 1) at *equal total numbers of vector calls* N_{orc} (Theorem 2.8; for both methods every call carries the constraint vector, so $N_{\text{orc}} = N_g$, while $N_f = 2b_pK_N$, resp. $4bN$, is smaller). ZO-Sliding uses $N = 60$ rounds, mini-batches $b_p = b_y = 8$, dual scale $\varrho = 1$, the inner schedule of (28), $S_t = \max\{1, \lceil M_g \varrho t/L \rceil\}$, with the same constants M_g and L that enter the theoretical bounds ($M_g = 1.06$, $L = 3$ for the smooth instance; $M_g = 1.10$, $L = 2 + \sqrt{d}/(4\gamma) \approx 17.3$ for the cusp; see Section F), so that the condition $a_t M_g \varrho \le L$ of Theorem 6.1 holds in every phase, and the stabilizer $\lambda = 0.1$; the resulting budgets are 28,064 calls with 678 inner steps (smooth) and 6,824 calls with 147 inner steps (cusp). We also run ZO-Sliding with the worst-case stabilizer of Theorem 6.6 ($\lambda = 865.1$ and $\lambda = 1896.0$). B-SMP uses $\alpha = \sqrt{20}$, $\eta = 0.45/L_\alpha$ and mini-batches $b \in \{1, 8\}$, with the number of iterations chosen so that the total number of calls equals the budget of ZO-Sliding. The parameters of both methods were fixed before the runs and not tuned to the outcome; the experiment is designed to exhibit the separation between the three resources predicted by the theory, not to establish the best empirical performance of either method, for which a tuning protocol with equal budgets would be needed. Each configuration is run with 30 independent seeds; we report means, and percentile bootstrap confidence intervals for the mean of the final joint error. Table 6 and Figure 4 show the results.

The picture is the one predicted by the theory. On the smooth instance ZO-Sliding and B-SMP ($b = 8$) reach comparable final accuracy ($J \approx 0.0016$ with 95% interval [0.0013, 0.0020] versus 0.0014 with [0.0013, 0.0015]; the small difference is within the sampling error of the experiment), but ZO-Sliding needs 60 rounds against 1,168, of which 992 are needed to reach the final error of ZO-Sliding: the inner loop performs 678 steps without new oracle rounds that a parallel implementation of B-SMP would have to spend as sequential rounds. B-SMP with $b = 1$ performs 9,354 rounds and stalls at $J \approx 0.0806$, with dual iterates oscillating around the active constraint; this is consistent with, though not implied by, the worst-case bound of Theorem 3.6, which does not certify a reduction of the ℓ_∞ noise for $b < 8\kappa_m$. (ZO-Sliding also uses $b = 8 < 8\kappa_m$; its dual noise is additionally averaged over the S_t slots of a phase and over the phases through the weights a_k, which is what the bound (45) in terms of bK_N expresses.) On the cusp instance the smoothed Lagrangian has a large smoothness constant $L \approx 17.3$ (Theorem 3.1(c)); B-SMP with $b = 8$ can afford only 142 iterations at the given

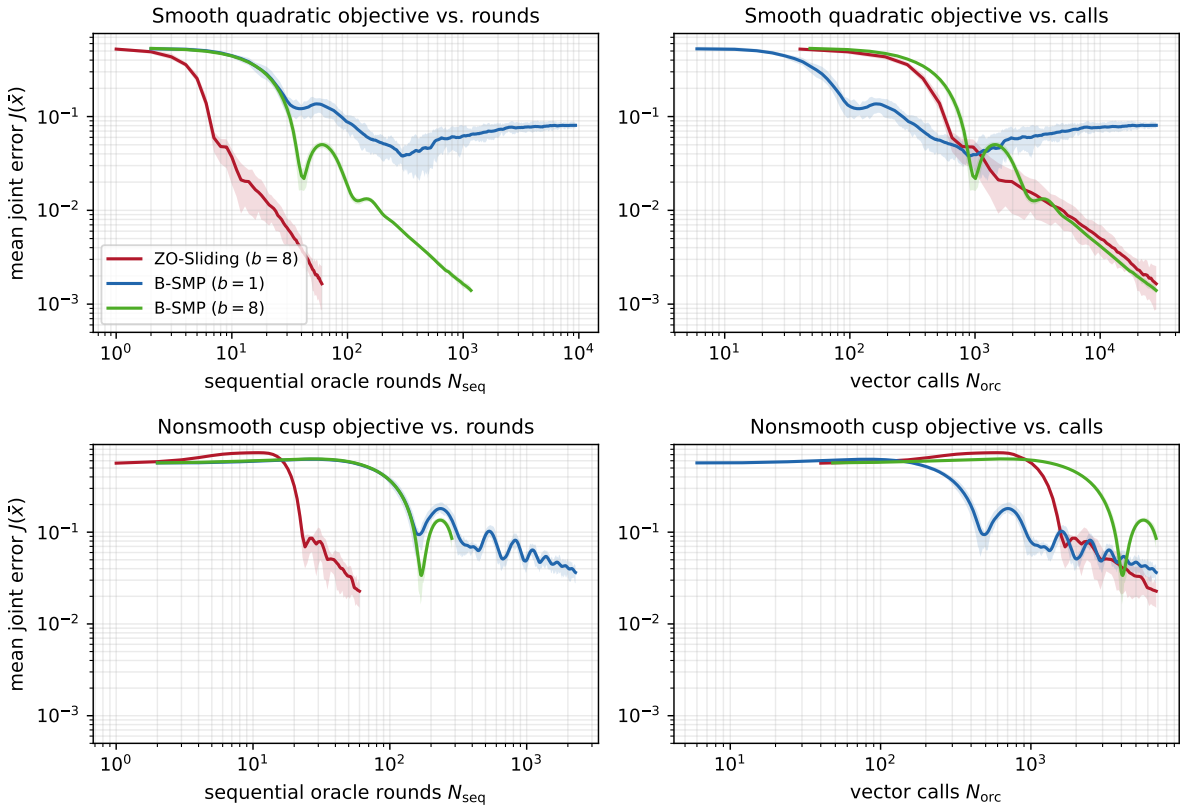


Figure 4: Mean joint error against sequential rounds (left) and against vector calls N_{orc} (right), smooth instance (top) and cusp instance (bottom); bands are the interquartile range over 30 seeds. On the smooth instance B-SMP with $b = 8$ (green) needs 992 rounds to reach the final mean error of ZO-Sliding (red), which uses 60; on the cusp instance ZO-Sliding reaches the final mean error of B-SMP with $b = 8$ after 24 rounds (against 284) and that of B-SMP with $b = 1$ (blue) after 47 rounds (against 2,274). B-SMP with $b = 1$ stalls on the smooth instance at the level of its dual noise; the worst-case bound of Theorem 3.6 does not certify a reduction of the ℓ_∞ noise for $b < 8\kappa_m \approx 30$, but the stall itself is an empirical observation.

budget and remains far from the solution ($J \approx 0.0856$), B-SMP with $b = 1$ reaches 0.0363, and ZO-Sliding reaches 0.0228 in 60 rounds. The remaining error is the stochastic error scale of this budget (the constraint offsets have standard deviation 0.02 and the optimal point lies on the active constraint); it decreases with the budget, as Figure 4 (right) shows, and the present data do not identify an asymptotic floor.

The runs with the worst-case stabilizer show the usual weakness of worst-case constants: with $\lambda \approx 10^3$ the inner proximal steps almost do not move ($\beta_t = (L + \lambda)/a_t$), the primal iterates stay close to the starting point $x_0 = a$ and the output is infeasible with $J \approx 0.34$ (smooth) and $J \approx 0.74$ (cusp). The value $\lambda = 0.1$ was chosen once and not tuned; Theorem 6.5 holds for every $\lambda > 0$, and the experiments indicate that the actual variance is far below the worst-case bound $b\,\Sigma_b^2 = \Sigma_A^2 \approx 3 \cdot 10^4$ that determines the worst-case stabilizer.

11.3 Restarts on the strongly convex instance

Figure 5 shows R-ZO-Sliding (Algorithm 3 with the base method Algorithm 2) on the smooth instance ($\mu = 1$): 7 stages of 10 rounds, mini-batch $b_{p,k} = b_{y,k} = 2^{k-1}$ and dual scale $\varrho_k = 2^{(k-1)/2}$ (so that the number of inner steps $S_t = \max\{1, \lceil M_g \varrho_k t/L \rceil\}$ grows with the stage), dual reset at every stage, compared with the non-restarted method with $N = 60$ rounds and three mini-batch sizes: $b = 8$ and $b = 16$, and $b = 23$, for which the plain run has (up to rounding) the same

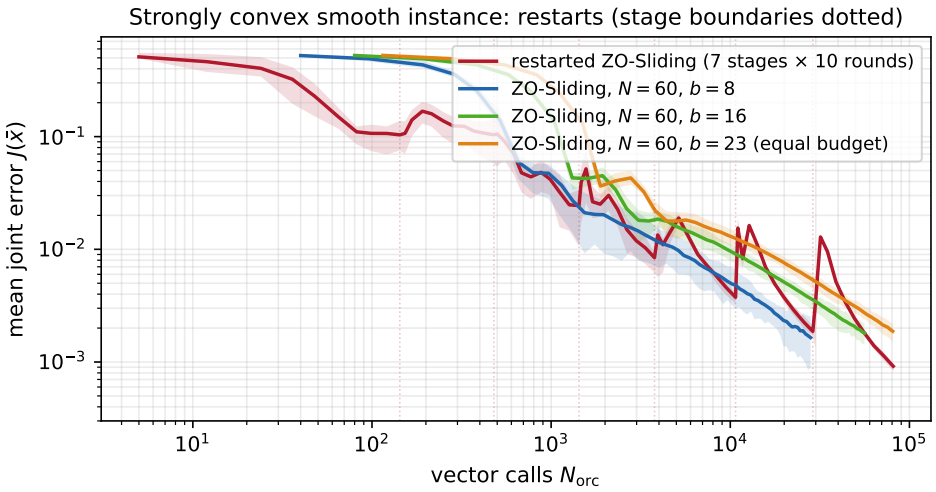


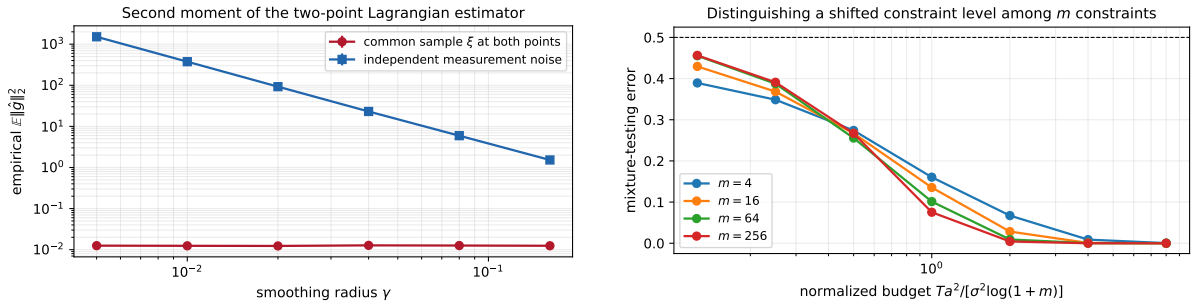
Figure 5: Restarted versus plain ZO-SLIDING on the strongly convex smooth instance; mean joint error against vector calls, interquartile bands over 30 seeds. Dotted vertical lines mark the stage boundaries of the restarted method; the plain run with $b = 23$ has the same budget as the restarted run.

number of vector calls as the restarted run, so that the last comparison is at *equal budget*.

The restarted method reaches $J = 0.0009$ (95% bootstrap interval [0.0009, 0.0010]) after 81,231 calls and 70 rounds, whereas the plain method reaches $J = 0.0016$ ([0.0013, 0.0020]) with $b = 8$ (28,064 calls), $J = 0.0018$ ([0.0016, 0.0020]) with $b = 16$ (56,128 calls) and $J = 0.0019$ ([0.0017, 0.0020]) with $b = 23$ (80,684 calls, the budget of the restarted run). Increasing the mini-batch of the plain method from 8 to 23 does not improve the final accuracy, which indicates that at $N = 60$ rounds the plain method is limited by its deterministic part and by the fixed dual scale rather than by the number of samples. At equal budget the restarted method ends at about half the error of the plain method, with disjoint confidence intervals, so the advantage is attributable to the schedule (the shrinking primal radius and the growing dual scale and inner loop) rather than to additional samples. This is the behaviour predicted by Theorem 7.2; the comparison uses the same fixed, untuned parameters for all runs. The saw-tooth shape of the restarted curve is the visible effect of the two resets at each stage boundary: the dual variable returns to the uniform point and the weighted average $\bar{x}$ restarts from the last stage output, so the first rounds of a stage are dominated by the dual transient. At the budgets considered here the level-noise term $\Lambda_m R^2 \sigma_h^2 / \varepsilon^2$, which is not improved by strong convexity, is already comparable to the primal term, so the observed advantage of the restarts is moderate, consistent with the theory.

11.4 Two diagnostics

Figure 6a verifies Theorem 3.3: the empirical second moment of the two-point estimator $\hat{g}_y(x)$ of the Lagrangian gradient (at $x = x^*$, $y = y^*$) is flat in γ over two orders of magnitude when both points share the sample ξ, while an estimator built from independent measurement noise of standard deviation 0.01 at the two points has second moment growing as γ^{-2}, which would force γ to stay large and the bias γM to dominate. Figure 6b illustrates Theorem 9.3: for the hypothesis-testing problem of its proof (one of m constraint levels is shifted by $-2a$, $a = 0.04$, $\sigma = 0.2$), the error of the likelihood-ratio test between the null hypothesis $\beta^{(0)}$ and the uniform mixture of the m alternatives, plotted against the normalized budget $Ta^2/(\sigma^2 \log(1+m))$, follows a single curve for $m \in \{4, 16, 64, 256\}$ and drops below 10% only for normalized budgets above 1; the number of vector calls needed for a fixed testing error therefore scales as $\sigma^2 \log(1+m)/a^2$, which is the source of the $\log(m+1)$ factors in the dual geometry.



(a) Common sample versus independent noise.

(b) Level information and $\log(1+m)$.

Figure 6: Monte Carlo diagnostics (12 000 samples per point in (a), 4 000 trials per point in (b)).

12 Conclusion and open problems

We have given a complete account of two-point zeroth-order methods for convex stochastic optimization with stochastic functional constraints, in which the three resources of a method—sequential oracle rounds, oracle calls (with a separate record of the calls used for the objective and for the constraints), and inner operations that require no new round—are accounted for separately. Randomized smoothing with an explicit shift of the constraint levels reduces the problem to a smooth saddle-point problem; a batched stochastic Mirror–Prox method serves as a baseline with $O(\varepsilon^{-1})$ rounds for smooth data; the zeroth-order primal–dual sliding method reduces the number of rounds to the level of the best known unconstrained methods, $O(\sqrt{HD^2/\varepsilon})$ (smooth, minimax optimal in ε) and $O(d^{1/4}\sqrt{M_R\bar{M}D/\varepsilon})$ (nonsmooth, the best known depth of the two-point model but not a minimax statement), while keeping $\tilde{O}(\varepsilon^{-2})$ calls, with the coupling cost paid in inner steps only; the separate primal and dual mini-batches charge the gradient information (factor d) to the objective calls and the level information (factor $\log(1+m)$) to the constraint calls only; restarts give $O(\sqrt{L/\mu}\log(1/\varepsilon))$ rounds and $O(dM_R^2/(\mu\varepsilon))$ objective calls for strongly convex objectives, while the level term stays at ε^{-2}, where the lower bound of Theorem 9.5 places it; and for exactly known affine constraints the constraints cost nothing in calls and $O(\varepsilon^{-1})$ in matrix–vector products. Information-theoretic lower bounds show that the ε^{-2} call terms are unavoidable, with the dimension factor d attached to the gradient information and the factor $\log(1+m)$ attached to the level information; the objective term is matched up to constants, the constraint-gradient term in d, M_g, D and ε at unit multipliers, and the level term up to $\kappa_m R^2$ (Table 5). Which parameters are closed and which are open is recorded in Table 5; the open ones are the subject of the questions below.

12.1 Open problems

The results raise the following questions, which are not addressed in this paper.

Open problem 12.1 (Slater-free variant through a level-set reformulation)**.** Consider the level-set function $\ell(\vartheta) := \min_{x\in Q}\max\{f_\gamma(x)-\vartheta,\ h_1(x),\dots,h_m(x)\}$, whose root is f_γ^*. Does a zeroth-order sliding method applied to the inner minimax problem at a fixed level ϑ, combined with a root-finding scheme in ϑ [3, 12, 40, 41], attain the round complexity of Theorem 6.5 with the multiplier bound R replaced by a constant depending only on a condition measure of ℓ, without Theorem 2.5? The inner problem has to be solved to an accuracy that depends on the unknown slope of ℓ at its root, and the zeroth-order estimates of the level values enter the root-finding test with non-vanishing noise.

Open problem 12.2 (dual noise without the diameter term)**.** Can the term $\Lambda_m R^2\nu_m M_g^2 D^2/\varepsilon^2$ in the call complexity of Theorems 6.5 and 7.2 be replaced by $\Lambda_m R^2\sigma_h^2/\varepsilon^2 +$ lower order by a method with the same round complexity? A natural candidate estimates the linearized constraint

value at p_k not by $\hat{J}(u_t)(p_k - u_t)$ but by a single query at a point that is announced one round ahead (a prediction of p_k); the control of the resulting bias requires techniques beyond those of this paper. In the strongly convex case a different route is to run stage k of Algorithm 3 on $Q \cap B(\bar{x}^{(k-1)}, r_k)$ with $r_k^2 = O(\varepsilon_{k-1}/\mu)$, which would replace D^2 by r_k^2 in (42) and make the term of the same order $\Lambda_m R^2 \nu_m M_g^2/(\mu\varepsilon)$ as the primal one; this needs the inclusion $x_\gamma^* \in B(\bar{x}^{(k-1)}, r_k)$ with high probability rather than in expectation.

Open problem 12.3 (depth lower bounds with constraints)**.** Is the coupling constant $M_g R$ necessarily present in the number of *rounds* of a zeroth-order method for (1), or only in the number of inner operations, as in Theorem 6.5? Is the factor $\sqrt{1 + RL_g/L_0}$ in the depth of Theorem 6.9 necessary beyond the unit-multiplier observation of Section 9.4? Lower bounds for parallel randomized algorithms [14, 19] are known only for unconstrained problems. A related question is whether the ball-acceleration methods with depth $\tilde{O}(d^{1/3}\varepsilon^{-2/3})$ [14, 15] can be implemented with a two-point value oracle, which would improve the nonsmooth depth of Theorem 6.10 in the regime $\varepsilon \lesssim d^{-1/4}$.

Open problem 12.4 (optimality of the multiplier factor)**.** The upper bounds contain R^2 in front of the ε^{-2} terms, whereas in the constructions of Theorems 9.2 and 9.3 the active multiplier equals 1. Determine the minimax dependence on the size of the multipliers (or on the Slater margin ρ). Two remarks delimit the question. First, the criterion (3) is not invariant under a rescaling $g \mapsto g/\kappa$ of the constraints, which multiplies the multipliers by κ and divides M_g and σ_g by κ while relaxing the violation requirement; a meaningful statement about the dependence on R must therefore fix the normalization of g through M_g and σ_g. Second, part of the gap is between the a priori bound $R = 1 + M_0 D/\rho$ of (14) and the actual size $1 + \|y_\gamma^*\|_1$ of the multipliers, i.e. a question of adaptivity to an unknown multiplier scale.

Open problem 12.5 (high probability and heavy tails)**.** All guarantees are in expectation. Bounds in probability can be obtained by running $O(\log(1/\delta))$ independent copies and selecting the best one, but the selection itself requires estimates of f and of v at the candidate points, i.e. additional single queries whose noise has variance σ_g^2 (and an analogous assumption on the values of F, which we did not need). A direct high-probability analysis under light-tailed noise, and an analysis under heavy-tailed noise with gradient clipping [33], are open for the constrained setting.

Open problem 12.6 (adversarial noise and one-point feedback)**.** For unconstrained problems, Gasnikov et al. [26] showed that the accelerated batched scheme tolerates bounded adversarial noise of level $\delta = O(\varepsilon^2/(d^{1/2}MD))$ (up to logarithms) and treated one-point feedback with $O(d^2/\varepsilon^3)$ or worse call complexities. The extension to constraints requires an analysis of the propagation of the bias δ/γ of the Jacobian estimates through the dual dynamics.

Open problem 12.7 (sharper constants and the factor κ_m)**.** What are the sharp constants in Theorem 6.5 (in place of 17, $16\kappa_m$, $92\kappa_m$) and in the restart theorems? The factor κ_m in the level term enters only through the worst-case ℓ_∞ bound of Theorem 3.6 on the dual noise. An analysis of the entropic dual step through local norms (as in the analysis of exponential-weights algorithms, where the residual of a step is $\frac{1}{2\tau}\sum_i \tilde{y}_i e_i^2$ rather than $\frac{1}{2\tau}\|e\|_\infty^2$) would replace $\mathbb{E}\,\|e\|_\infty^2$ by $R \max_i \mathbb{E} e_i^2$ and remove κ_m from the level term, leaving only the $\log(m+1)$ of the entropy radius; this requires a control of $\max_i |\hat{q}_{k,i}|/\tau$, e.g. by clipping the level estimates or by a boundedness assumption on the noise. We do not pursue this here.

References

[1] A. Agarwal, O. Dekel, and L. Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Proceedings of the 23rd Conference on Learning Theory (COLT)*, pages 28–40, 2010.

[2] A. Akhavan, M. Pontil, and A. B. Tsybakov. Exploiting higher order smoothness in derivative-free optimization and continuous bandits. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[3] A. Y. Aravkin, J. V. Burke, D. Drusvyatskiy, M. P. Friedlander, and S. Roy. Level-set methods for convex optimization. *Mathematical Programming*, 174:359–390, 2019.

[4] F. Bach and V. Perchet. Highly-smooth zero-th order online optimization. In *Proceedings of the 29th Conference on Learning Theory (COLT)*, volume 49 of *Proceedings of Machine Learning Research*, pages 257–283, 2016.

[5] D. Bakry, I. Gentil, and M. Ledoux. *Analysis and Geometry of Markov Diffusion Operators*. Springer, 2014.

[6] K. Balasubramanian and S. Ghadimi. Zeroth-order nonconvex stochastic optimization: handling constraints, high dimensionality, and saddle points. *Foundations of Computational Mathematics*, 22(1):35–76, 2022.

[7] E. Balkanski and Y. Singer. Parallelization does not accelerate convex optimization: adaptivity lower bounds for non-smooth convex minimization, 2018. arXiv:1808.03880.

[8] A. Bayandina, P. Dvurechensky, A. Gasnikov, F. Stonyakin, and A. Titov. Mirror descent and convex optimization problems with non-smooth inequality constraints. In *Large-Scale and Distributed Optimization*, volume 2227 of *Lecture Notes in Mathematics*, pages 181–213. Springer, 2018.

[9] A. S. Bayandina, A. V. Gasnikov, and A. A. Lagunovskaya. Gradient-free two-point methods for solving stochastic nonsmooth convex optimization problems with small non-random noises. *Automation and Remote Control*, 79(8):1399–1408, 2018.

[10] A. Beznosikov, A. Sadiev, and A. Gasnikov. Gradient-free methods with inexact oracle for convex-concave stochastic saddle-point problem. In *Mathematical Optimization Theory and Operations Research (MOTOR 2020)*, volume 1275 of *Communications in Computer and Information Science*, pages 105–119. Springer, 2020.

[11] D. Boob, Q. Deng, and G. Lan. Stochastic first-order methods for convex and nonconvex functional constrained optimization. *Mathematical Programming*, 197:215–279, 2023.

[12] D. Boob, Q. Deng, and G. Lan. Level constrained first order methods for function constrained optimization. *Mathematical Programming*, 209:1–61, 2025.

[13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[14] S. Bubeck, Q. Jiang, Y. T. Lee, Y. Li, and A. Sidford. Complexity of highly parallel non-smooth convex optimization. In *Advances in Neural Information Processing Systems*, volume 32, pages 13900–13909, 2019.

[15] Y. Carmon, A. Jambulapati, Y. Jin, Y. T. Lee, D. Liu, A. Sidford, and K. Tian. ReSQueing parallel and private stochastic convex optimization. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 2031–2058, 2023.

[16] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[17] A. R. Conn, K. Scheinberg, and L. N. Vicente. *Introduction to Derivative-Free Optimization*. SIAM, 2009.

[18] Q. Deng, G. Lan, and Z. Lin. Uniformly optimal and parameter-free first-order methods for convex and function-constrained optimization. *INFORMS Journal on Computing*, 2026. Published online, https://doi.org/10.1287/ijoc.2025.1177; arXiv:2412.06319.

[19] J. Diakonikolas and C. Guzmán. Lower bounds for parallel and randomized convex optimization. *Journal of Machine Learning Research*, 21(5):1–31, 2020.

[20] J. Duchi, F. Ruan, and C. Yun. Minimax bounds on stochastic batched convex optimization. In *Proceedings of the 31st Conference on Learning Theory (COLT)*, volume 75 of *Proceedings of Machine Learning Research*, pages 3065–3162, 2018.

[21] J. C. Duchi, P. L. Bartlett, and M. J. Wainwright. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012.

[22] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono. Optimal rates for zero-order convex optimization: the power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

[23] D. Dvinskikh, V. Tominin, I. Tominin, and A. Gasnikov. Noisy zeroth-order optimization for non-smooth saddle point problems. In *Mathematical Optimization Theory and Operations Research (MOTOR 2022)*, volume 13367 of *Lecture Notes in Computer Science*, pages 18–33. Springer, 2022.

[24] P. Dvurechensky, E. Gorbunov, and A. Gasnikov. An accelerated directional derivative method for smooth stochastic convex optimization. *European Journal of Operational Research*, 290(2):601–621, 2021.

[25] A. D. Flaxman, A. T. Kalai, and H. B. McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 385–394, 2005.

[26] A. Gasnikov, A. Novitskii, V. Novitskii, F. Abdukhakimov, D. Kamzolov, A. Beznosikov, M. Takáč, P. Dvurechensky, and B. Gu. The power of first-order smooth optimization for black-box non-smooth problems. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pages 7241–7265, 2022.

[27] A. Gasnikov, D. Dvinskikh, P. Dvurechensky, E. Gorbunov, A. Beznosikov, and A. Lobanov. Randomized gradient-free methods in convex optimization. In *Encyclopedia of Optimization*. Springer, 2023. arXiv:2211.13566.

[28] A. V. Gasnikov, E. A. Krymova, A. A. Lagunovskaya, I. N. Usmanova, and F. A. Fedorenko. Stochastic online optimization. Single-point and multi-point non-linear multi-armed bandits. Convex and strongly-convex case. *Automation and Remote Control*, 78(2):224–234, 2017.

[29] A. V. Gasnikov, A. V. Lobanov, and F. S. Stonyakin. Highly smooth zeroth-order methods for solving optimization problems under the PL condition. *Computational Mathematics and Mathematical Physics*, 64(4):739–770, 2024.

[30] E. Gorbunov, D. Dvinskikh, and A. Gasnikov. Optimal decentralized distributed algorithms for stochastic convex optimization, 2019. arXiv:1911.07363.

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

[32] A. Juditsky, A. Nemirovski, and C. Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

[33] N. Kornilov, O. Shamir, A. Lobanov, D. Dvinskikh, A. Gasnikov, I. Shibaev, E. Gorbunov, and S. Horváth. Accelerated zeroth-order method for non-smooth stochastic convex optimization problem with infinite variance. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[34] G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133:365–397, 2012.

[35] G. Lan. Gradient sliding for composite optimization. *Mathematical Programming*, 159: 201–235, 2016.

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

[37] J. Larson, M. Menickelly, and S. M. Wild. Derivative-free optimization methods. *Acta Numerica*, 28:287–404, 2019.

[38] M. Ledoux. *The Concentration of Measure Phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, 2001.

[39] Z. Li, P.-Y. Chen, S. Liu, S. Lu, and Y. Xu. Zeroth-order optimization for composite problems with functional constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7453–7461, 2022.

[40] Q. Lin, R. Ma, and T. Yang. Level-set methods for finite-sum constrained convex optimization. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pages 3112–3121, 2018.

[41] Q. Lin, S. Nadarajah, N. Soheili, and T. Yang. A data efficient and feasible level set method for stochastic convex optimization with expectation constraints. *Journal of Machine Learning Research*, 21(143):1–45, 2020.

[42] A. Lobanov. Stochastic adversarial noise in the "black box" optimization problem. In *Optimization and Applications (OPTIMA 2023)*, volume 14395 of *Lecture Notes in Computer Science*, pages 60–71. Springer, 2023.

[43] A. Lobanov, N. Bashirov, and A. Gasnikov. The "black-box" optimization problem: zero-order accelerated stochastic method via kernel approximation, 2023. arXiv:2310.02371.

[44] A. Nemirovski. On parallel complexity of nonsmooth convex optimization. *Journal of Complexity*, 10(4):451–463, 1994.

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

[46] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[47] A. S. Nemirovsky and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.

[48] Y. Nesterov and V. Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

[49] A. Nguyen and K. Balasubramanian. Stochastic zeroth-order functional constrained optimization: oracle complexity and applications. *INFORMS Journal on Optimization*, 5(3): 256–272, 2023.

[50] V. Novitskii and A. Gasnikov. Improved exploitation of higher order smoothness in derivative-free optimization. *Optimization Letters*, 16:2059–2071, 2022.

[51] Y. Ouyang and Y. Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, 185:1–35, 2021.

[52] B. T. Polyak. A general method for solving extremal problems. *Soviet Mathematics Doklady*, 8:593–597, 1967.

[53] B. T. Polyak and A. B. Tsybakov. Optimal order of accuracy of search algorithms in stochastic optimization. *Problems of Information Transmission*, 26(2):126–133, 1990.

[54] K. Ren and L. Luo. A parameter-free and near-optimal zeroth-order algorithm for stochastic convex optimization. In *Proceedings of the 42nd International Conference on Machine Learning (ICML 2025)*, volume 267 of *Proceedings of Machine Learning Research*. PMLR, 2025. arXiv:2502.05600.

[55] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[56] O. Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *Journal of Machine Learning Research*, 18(52):1–11, 2017.

[57] U. V. Shanbhag and F. Yousefian. Zeroth-order randomized block methods for constrained minimization of expectation-valued Lipschitz continuous functions. In *2021 Seventh Indian Control Conference (ICC)*, pages 7–12, 2021.

[58] J. C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37(3):332–341, 1992.

[59] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

[60] I. Usmanova, A. Krause, and M. Kamgarpour. Safe convex learning under uncertain constraints. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89 of *Proceedings of Machine Learning Research*, pages 2106–2114, 2019.

[61] Y. Xu. First-order methods for constrained convex programming based on linearized augmented Lagrangian function. *INFORMS Journal on Optimization*, 3(1):89–117, 2021.

[62] Z. Yu, D. W. C. Ho, and D. Yuan. Distributed randomized gradient-free mirror descent algorithm for constrained optimization. *IEEE Transactions on Automatic Control*, 67(2): 957–964, 2022.

[63] H. Zhang, C. Zhang, Z. Qi, and Z. Lin. Near-optimal lower bounds for randomized algorithms in exact value zeroth-order convex optimization, 2026. arXiv:2607.16558.

[64] Z. Zhang and G. Lan. Solving convex smooth function constrained optimization is almost as easy as unconstrained optimization, 2022. arXiv:2210.05807.

A Auxiliary facts

Fact A.1 (Poincaré inequality on the sphere)**.** *Let $d \geq 2$, $u \sim U(\mathbb{S})$ and $q : \mathbb{S} \to \mathbb{R}$ be Λ-Lipschitz with respect to the Euclidean distance of $\mathbb{R}^d$. Then* $\operatorname{Var} q(u) = \mathbb{E}(q(u) - \mathbb{E}q(u))^2 \leq \Lambda^2/(d-1)$.

Reference. The uniform probability measure on $\mathbb{S} = \mathbb{S}^{d-1}$ satisfies the Poincaré inequality $\operatorname{Var}(q) \leq \frac{1}{d-1}\mathbb{E}\left\|\nabla_{\mathbb{S}} q\right\|^2$, where $d-1$ is the first nonzero eigenvalue of the Laplace–Beltrami operator (e.g. 38, Sect. 2.1 and Prop. 2.10, 5, Sect. 4.5). A function that is Λ-Lipschitz for the Euclidean (chordal) distance is Λ-Lipschitz for the geodesic distance, hence $\|\nabla_{\mathbb{S}} q\| \leq \Lambda$ almost everywhere by Rademacher's theorem. □

Fact A.2 (Gaussian concentration)**.** *Let $\mathsf{g} \sim \mathcal{N}(0, I_d)$ and $\mathsf{Q} : \mathbb{R}^d \to \mathbb{R}$ be Λ-Lipschitz. Then for all $t \geq 0$, $\mathbb{P}(\mathsf{Q}(\mathsf{g}) - \mathbb{E}\mathsf{Q}(\mathsf{g}) \geq t) \leq e^{-t^2/(2\Lambda^2)}$ [13, Thm. 5.6].*

Fact A.3 (moments on the sphere and for Gaussians)**.** *Let $u \sim U(\mathbb{S})$, $\mathsf{g} \sim \mathcal{N}(0, I_d)$ and $w \in \mathbb{R}^d$. Then $\mathbb{E}\langle u, w\rangle^2 = \|w\|^2/d$, $\mathbb{E}\langle u, w\rangle^4 = 3\|w\|^4/(d(d+2))$, $\mathbb{E}\|\mathsf{g}\|^2 = d$, $\mathbb{E}\|\mathsf{g}\|^4 = d(d+2)$, and $\mathsf{g}/\|\mathsf{g}\| \sim U(\mathbb{S})$ is independent of $\|\mathsf{g}\|$.*

Proof. By rotation invariance $\mathbb{E}\langle u, w\rangle^{2k} = \|w\|^{2k}\mathbb{E}u_1^{2k}$, and $\sum_i \mathbb{E}u_i^2 = 1$ gives $\mathbb{E}u_1^2 = 1/d$. Since $\mathsf{g} = \|\mathsf{g}\|\, u$ with independent factors, $\mathbb{E}\mathsf{g}_1^4 = \mathbb{E}\|\mathsf{g}\|^4\, \mathbb{E}u_1^4$, and $\mathbb{E}\mathsf{g}_1^4 = 3$, $\mathbb{E}\|\mathsf{g}\|^4 = \operatorname{Var}(\chi_d^2) + d^2 = 2d + d^2$ give $\mathbb{E}u_1^4 = 3/(d(d+2))$. □

Fact A.4 (Pinsker's inequality and χ^2)**.** *For probability measures P, Q on a common space, $\mathrm{TV}(P,Q) := \sup_A |P(A) - Q(A)| \leq \sqrt{\mathrm{KL}(P\|Q)/2}$ and $\mathrm{TV}(P,Q) \leq \frac{1}{2}\sqrt{\chi^2(P\|Q)}$, where $\chi^2(P\|Q) = \mathbb{E}_Q[(dP/dQ)^2] - 1$. For probability vectors $p, q \in \mathbb{R}^n$, $\mathrm{TV} = \frac{1}{2}\|p - q\|_1$, hence $\mathrm{KL}(p\|q) \geq \frac{1}{2}\|p - q\|_1^2$ [59, Lemma 2.5].*

Fact A.5 (chain rule and data processing for adaptive experiments)**.** *Consider an adaptive algorithm that, for $t = 1, \ldots, T$, chooses a query X_t as a measurable function of the previous observations $O_1, \ldots, O_{t-1}$ and of an internal random seed ω, and receives an observation O_t drawn from a distribution P_{X_t} (respectively P'_{X_t}) independently of the past given X_t. Let P, P' be the laws of the transcript $(\omega, X_1, O_1, \ldots, X_T, O_T)$ under the two families. Then*

$$\mathrm{KL}(P\|P') = \sum_{t=1}^{T} \mathbb{E}_P\big[\mathrm{KL}(P_{X_t}\|P'_{X_t})\big],$$

and for every measurable function ψ of the transcript, $\mathrm{KL}(\mathrm{law}_P(\psi)\|\mathrm{law}_{P'}(\psi)) \leq \mathrm{KL}(P\|P')$ and $\mathrm{TV}(\mathrm{law}_P(\psi), \mathrm{law}_{P'}(\psi)) \leq \mathrm{TV}(P, P')$ [59, Sect. 2.4]. Moreover, if $P_{X_t} = P'_{X_t}$ does not depend on X_t (the observations are i.i.d. regardless of the queries), the likelihood ratio dP/dP' is the product of the likelihood ratios of the observations.

Proof. The chain rule follows from the factorization of the transcript density as $p(\omega)\prod_t p_{X_t}(O_t)$ (the query X_t being a deterministic function of the past, its conditional law is a point mass under both hypotheses) and the tower property; the data-processing inequalities are the monotonicity of f-divergences under measurable maps. □

Fact A.6 (three-point inequality)**.** *Let $Z \subseteq Z'$ be closed convex sets, ω a convex function on Z' that is 1-strongly convex with respect to a norm $\|\cdot\|_Z$ on Z' and differentiable on an open set containing the relative interior Z'° of Z' (or the whole Z'), and $V(z, z') = \omega(z') - \omega(z) - \langle \nabla\omega(z), z' - z\rangle$. For $\bar{z} \in Z'^{\circ}$ (not necessarily in Z), a vector ξ and $\tau > 0$, let $z^+ := \arg\min_{z \in Z}\{\langle \xi, z\rangle + \tau V(\bar{z}, z)\}$ and assume $z^+ \in Z'^{\circ}$ (this holds automatically for the Euclidean distance and for the entropy on the simplex, where the minimizer has positive coordinates, see the explicit formula of Theorem 4.4(c)). Then for all $z \in Z$,*

$$\langle \xi, z^+ - z\rangle \leq \tau\big[V(\bar{z}, z) - V(z^+, z) - V(\bar{z}, z^+)\big], \qquad V(\bar{z}, z^+) \geq \tfrac{1}{2}\left\|z^+ - \bar{z}\right\|_Z^2.$$

Proof. The optimality condition of the convex problem defining z^+ is

$$\left\langle \xi + \tau(\nabla\omega(z^+) - \nabla\omega(\bar{z})), z - z^+ \right\rangle \geq 0 \quad \text{for all } z \in Z,$$

and the identity $\langle \nabla\omega(z^+) - \nabla\omega(\bar{z}), z - z^+\rangle = V(\bar{z}, z) - V(z^+, z) - V(\bar{z}, z^+)$ is verified by expanding the definitions. Strong convexity gives the last claim. □

Fact A.7 (conjugates of smooth convex functions)**.** *Let $\mathsf{F} : \mathbb{R}^d \to \mathbb{R}$ be convex and L-smooth, and $\mathsf{F}^*(\pi) := \sup_x\{\langle \pi, x\rangle - \mathsf{F}(x)\}$ its Fenchel conjugate. For every $x' \in \mathbb{R}^d$, with $g' := \nabla\mathsf{F}(x')$, we have $\mathsf{F}^*(g') = \langle g', x'\rangle - \mathsf{F}(x')$ (Fenchel–Young equality) and, for every $\pi \in \operatorname{dom}\mathsf{F}^*$,*

$$\mathsf{F}^*(\pi) \;\geq\; \mathsf{F}^*(g') + \left\langle x', \pi - g' \right\rangle + \frac{1}{2L}\left\|\pi - g'\right\|^2.$$

Proof. The equality is the first-order optimality condition of the concave maximization defining $\mathsf{F}^*(g')$, attained at x'. By L-smoothness, $\mathsf{F}(x) \leq \mathsf{F}(x') + \langle g', x - x'\rangle + \frac{L}{2}\|x - x'\|^2$ for all x, hence

$$\begin{aligned}
\mathsf{F}^*(\pi) &\geq \sup_x\Big\{\langle \pi, x\rangle - \mathsf{F}(x') - \left\langle g', x - x'\right\rangle - \frac{L}{2}\left\|x - x'\right\|^2\Big\} \\
&= \left\langle \pi, x'\right\rangle - \mathsf{F}(x') + \sup_x\Big\{\left\langle \pi - g', x - x'\right\rangle - \frac{L}{2}\left\|x - x'\right\|^2\Big\} \\
&= \mathsf{F}^*(g') + \left\langle x', \pi - g'\right\rangle + \frac{1}{2L}\left\|\pi - g'\right\|^2.
\end{aligned}$$

□

B Proofs for Section 3

Proof of Theorem 3.1. (a) Convexity is inherited by averaging: $\varphi_\gamma(\theta x + (1-\theta)x') = \mathbb{E}\varphi(\theta(x + \gamma v) + (1-\theta)(x' + \gamma v)) \leq \theta\varphi_\gamma(x) + (1-\theta)\varphi_\gamma(x')$. If φ is M-Lipschitz on C_γ and $x, x' \in C$, then $x + \gamma v, x' + \gamma v \in C_\gamma$ and $|\varphi_\gamma(x) - \varphi_\gamma(x')| \leq \mathbb{E}|\varphi(x + \gamma v) - \varphi(x' + \gamma v)| \leq M\|x - x'\|$. Since $\mathbb{E}v = 0$, Jensen's inequality gives $\varphi(x) = \varphi(\mathbb{E}(x + \gamma v)) \leq \mathbb{E}\varphi(x + \gamma v) = \varphi_\gamma(x)$ for every x; and for $x \in C$, $\varphi(x + \gamma v) \leq \varphi(x) + \gamma M\|v\| \leq \varphi(x) + \gamma M$ because the segment $[x, x + \gamma v]$ lies in C_γ. If φ is L-smooth, $\varphi(x + \gamma v) \leq \varphi(x) + \gamma\langle \nabla\varphi(x), v\rangle + \frac{L}{2}\gamma^2\|v\|^2$, and taking expectations with $\mathbb{E}v = 0$ and $\mathbb{E}\|v\|^2 = d/(d+2)$ (for $v \sim U(\mathbb{B})$, $\mathbb{P}(\|v\| \leq r) = r^d$, so $\mathbb{E}\|v\|^2 = \int_0^1 r^2\, d(r^d) = d/(d+2)$) gives the second bound. Affine functions are their own averages.

(b) Let first $\psi \in C^1(\mathbb{R}^d)$. Then $\psi_\gamma(x) = |\gamma\mathbb{B}|^{-1}\int_{\gamma\mathbb{B}}\psi(x+z)\,dz$ is C^1 (differentiation under the integral sign over the compact ball, the gradient being continuous), and applying the divergence theorem on the ball $\gamma\mathbb{B}$ (outer unit normal z/γ) to the C^1 vector fields $\psi(x + \cdot)e_i$,

$$\begin{aligned}
\nabla\psi_\gamma(x) = \frac{1}{|\gamma\mathbb{B}|}\int_{\gamma\mathbb{B}}\nabla\psi(x+z)\,dz &= \frac{1}{|\gamma\mathbb{B}|}\int_{\gamma\mathbb{S}}\psi(x+z)\frac{z}{\gamma}\,d\sigma(z) \\
&= \frac{|\gamma\mathbb{S}|}{|\gamma\mathbb{B}|}\cdot\frac{1}{\gamma}\,\mathbb{E}_{u\sim U(\mathbb{S})}[\psi(x+\gamma u)u] = \frac{d}{\gamma}\,\mathbb{E}[\psi(x+\gamma u)u],
\end{aligned}$$

because $|\gamma\mathbb{S}|/|\gamma\mathbb{B}| = d/\gamma$ (the $(d-1)$-dimensional measure of $\gamma\mathbb{S}$ is $d\gamma^{d-1}|\mathbb{B}|$). This settles the case of an L-smooth φ. For a general M-Lipschitz convex φ, let $\varphi^\delta := \varphi * \rho_\delta$ be the convolution with a smooth probability density supported in $\delta\mathbb{B}$; then $\varphi^\delta \in C^\infty$ is M-Lipschitz and $\sup_x|\varphi^\delta(x) - \varphi(x)| \leq M\delta$. Consequently $\sup_x|(\varphi^\delta)_\gamma(x) - \varphi_\gamma(x)| \leq M\delta$ and $\sup_x\left\|\nabla(\varphi^\delta)_\gamma(x) - \frac{d}{\gamma}\mathbb{E}[\varphi(x+\gamma u)u]\right\| \leq \frac{d}{\gamma}M\delta$. Thus $(\varphi^\delta)_\gamma \to \varphi_\gamma$ and $\nabla(\varphi^\delta)_\gamma \to G(x) := \frac{d}{\gamma}\mathbb{E}[\varphi(x+\gamma u)u]$ uniformly on $\mathbb{R}^d$ as $\delta \to 0$, where G is continuous (indeed Md-Lipschitz, since $\|G(x) - G(x')\| \leq \frac{d}{\gamma}\mathbb{E}|\varphi(x+\gamma u) - \varphi(x'+\gamma u)| \leq \frac{dM}{\gamma}\|x - x'\|$). Uniform convergence of

C^1 functions together with uniform convergence of their gradients to a continuous limit implies that the limit is C^1 with the limit gradient: for every x, h, $(\varphi^\delta)_\gamma(x+h) - (\varphi^\delta)_\gamma(x) = \int_0^1 \langle \nabla(\varphi^\delta)_\gamma(x+sh), h\rangle\, ds \to \int_0^1 \langle G(x+sh), h\rangle\, ds$, so $\varphi_\gamma(x+h) - \varphi_\gamma(x) = \int_0^1 \langle G(x+sh), h\rangle\, ds = \langle G(x), h\rangle + o(\|h\|)$ by continuity of G. The symmetric form follows from $u \mapsto -u$ invariance of $U(\mathbb{S})$.

(c) For x, x' and any unit vector e, by (7), $\langle \nabla\varphi_\gamma(x) - \nabla\varphi_\gamma(x'), e\rangle = \frac{d}{\gamma}\mathbb{E}[(\varphi(x+\gamma u) - \varphi(x'+\gamma u))\,\langle u, e\rangle]$, whose absolute value is at most $\frac{d}{\gamma} M\,\|x - x'\|\,\mathbb{E}|\langle u, e\rangle| \le \frac{d}{\gamma} M\,\|x - x'\|\,\sqrt{\mathbb{E}\,\langle u, e\rangle^2} = \frac{\sqrt{d}M}{\gamma}\,\|x - x'\|$ by Theorem A.3. Taking the supremum over e gives the Lipschitz constant $\sqrt{d}M/\gamma$. If φ is L-smooth, differentiation under the expectation is justified by dominated convergence with the local bound $\|\nabla\varphi(x+\gamma v)\| \le \|\nabla\varphi(x)\| + L\gamma$ on the compact ball $x + \gamma\mathbb{B}$ (no Lipschitz bound on φ is needed), and gives $\nabla\varphi_\gamma(x) = \mathbb{E}\nabla\varphi(x+\gamma v)$, which is L-Lipschitz.

(d) If $\psi := \varphi - \frac{\mu}{2}\|\cdot\|^2$ is convex on C_γ, then for $x, x' \in C$ and $\theta \in [0,1]$ the points $x+\gamma v, x'+\gamma v$ and their convex combination lie in C_γ, so $\psi_\gamma(x) := \mathbb{E}\psi(x+\gamma v)$ is convex on C; and $\varphi_\gamma(x) = \psi_\gamma(x) + \frac{\mu}{2}\mathbb{E}\,\|x+\gamma v\|^2 = \psi_\gamma(x) + \frac{\mu}{2}\|x\|^2 + \frac{\mu}{2}\gamma^2\mathbb{E}\,\|v\|^2$, so $\varphi_\gamma - \frac{\mu}{2}\|\cdot\|^2$ equals on C the convex function ψ_γ plus a constant.

(e) $\tilde\varphi$ is the infimal convolution of the convex function $\varphi + \iota_{C_{\bar\gamma}}$ (where ι is the indicator, 0 on the set and $+\infty$ outside) with the convex function $M\|\cdot\|$, hence convex; it is finite because $C_{\bar\gamma} \neq \emptyset$ and φ is bounded below on the bounded set $C_{\bar\gamma}$ (for unbounded C finiteness follows from the Lipschitz bound $\varphi(z) \ge \varphi(z_0) - M\|z - z_0\|$ on $C_{\bar\gamma}$, which makes $z \mapsto \varphi(z) + M\|x - z\|$ bounded below by $\varphi(z_0) - M\|x - z_0\|$); and it is M-Lipschitz as an infimum of M-Lipschitz functions of x. For $x \in C_{\bar\gamma}$ the choice $z = x$ gives $\tilde\varphi(x) \le \varphi(x)$, while $\varphi(z) + M\|x - z\| \ge \varphi(x)$ for every $z \in C_{\bar\gamma}$ by the Lipschitz property on the convex set $C_{\bar\gamma}$; hence $\tilde\varphi = \varphi$ on $C_{\bar\gamma}$. Finally, for $x \in C$ and $\gamma \le \bar\gamma$ the points $x + \gamma v$ lie in $C_{\bar\gamma}$, so $\tilde\varphi_\gamma(x) = \mathbb{E}\tilde\varphi(x+\gamma v) = \mathbb{E}\varphi(x+\gamma v) = \varphi_\gamma(x)$. □

Proof of Theorem 3.4(d). Fix ξ and $w \neq 0$, and put $q_i(u) := \frac{1}{2\gamma}[G_i(x+\gamma u, \xi) - G_i(x-\gamma u, \xi)]$ for $u \in \mathbb{S}$. Each q_i is odd and $M_g(\xi)$-Lipschitz on $\mathbb{S}$ for the Euclidean distance, and $|q_i(u)| = \frac{1}{2}|q_i(u) - q_i(-u)| \le M_g(\xi)$. Extend q_i to $\mathbb{R}^d$ by positive homogeneity: $\mathsf{Q}_i(\mathsf{g}) := \|\mathsf{g}\|\, q_i(\mathsf{g}/\|\mathsf{g}\|)$ for $\mathsf{g} \neq 0$, $\mathsf{Q}_i(0) := 0$. We claim that Q_i is $3M_g(\xi)$-Lipschitz on $\mathbb{R}^d$. Indeed, for $\mathsf{g}, \mathsf{g}' \neq 0$ with unit vectors $\hat{\mathsf{g}}, \hat{\mathsf{g}}'$,

$$|\mathsf{Q}_i(\mathsf{g}) - \mathsf{Q}_i(\mathsf{g}')| \le \big|\|\mathsf{g}\| - \|\mathsf{g}'\|\big|\,|q_i(\hat{\mathsf{g}})| + \|\mathsf{g}'\|\,|q_i(\hat{\mathsf{g}}) - q_i(\hat{\mathsf{g}}')| \le M_g(\xi)\,\|\mathsf{g} - \mathsf{g}'\| + M_g(\xi)\,\|\mathsf{g}'\|\,\|\hat{\mathsf{g}} - \hat{\mathsf{g}}'\|,$$

and $\|\mathsf{g}'\|\,\|\hat{\mathsf{g}} - \hat{\mathsf{g}}'\| = \left\|\frac{\|\mathsf{g}'\|}{\|\mathsf{g}\|}\mathsf{g} - \mathsf{g}'\right\| \le \left\|\frac{\|\mathsf{g}'\|}{\|\mathsf{g}\|}\mathsf{g} - \mathsf{g}\right\| + \|\mathsf{g} - \mathsf{g}'\| = \big|\|\mathsf{g}'\| - \|\mathsf{g}\|\big| + \|\mathsf{g} - \mathsf{g}'\| \le 2\,\|\mathsf{g} - \mathsf{g}'\|$; the case $\mathsf{g}' = 0$ is $|\mathsf{Q}_i(\mathsf{g})| \le M_g(\xi)\,\|\mathsf{g}\|$. Now let $\mathsf{g} \sim \mathcal{N}(0, I_d)$, so that $u := \mathsf{g}/\|\mathsf{g}\| \sim U(\mathbb{S})$ is independent of $\|\mathsf{g}\|$ (Theorem A.3). Since Q_i is odd, $\mathbb{E}\mathsf{Q}_i(\mathsf{g}) = 0$, and Theorem A.2 applied to $\pm\mathsf{Q}_i$ gives $\mathbb{P}(|\mathsf{Q}_i(\mathsf{g})| \ge t) \le 2e^{-t^2/(18M_g(\xi)^2)}$. Writing $c := 18M_g(\xi)^2$ and using $\mathbb{P}(\max_i \mathsf{Q}_i^4 > t) \le \min\{1, 2me^{-\sqrt{t}/c}\}$,

$$\mathbb{E}\max_i \mathsf{Q}_i(\mathsf{g})^4 \le \alpha + \int_\alpha^\infty 2me^{-\sqrt{t}/c}\,dt = \alpha + 4mc\,(\sqrt{\alpha} + c)\,e^{-\sqrt{\alpha}/c} \qquad (\alpha > 0),$$

because $\int_\alpha^\infty e^{-\sqrt{t}/c}dt = 2\int_{\sqrt{\alpha}}^\infty se^{-s/c}ds = 2c(\sqrt{\alpha} + c)e^{-\sqrt{\alpha}/c}$. With $\sqrt{\alpha} := c\log(2m)$ this gives $\mathbb{E}\max_i \mathsf{Q}_i(\mathsf{g})^4 \le c^2[\log^2(2m) + 2\log(2m) + 2] \le 2c^2(1 + \log(2m))^2 = 648\,M_g(\xi)^4\kappa_m^2$. On the other hand $\max_i \mathsf{Q}_i(\mathsf{g})^4 = \|\mathsf{g}\|^4 \max_i q_i(u)^4$ with independent factors, so $\mathbb{E}_u \max_i q_i(u)^4 = \mathbb{E}\max_i \mathsf{Q}_i(\mathsf{g})^4/\mathbb{E}\,\|\mathsf{g}\|^4 \le 648M_g(\xi)^4\kappa_m^2/(d(d+2))$. Since $(\hat J(x)w)_i = d\,q_i(u)\,\langle u, w\rangle$, the Cauchy–Schwarz inequality and Theorem A.3 yield, still for fixed ξ,

$$\mathbb{E}_u\left\|\hat J(x)w\right\|_\infty^2 = d^2\,\mathbb{E}_u\Big[\max_i q_i(u)^2\,\langle u, w\rangle^2\Big] \le d^2\sqrt{\mathbb{E}_u \max_i q_i(u)^4}\sqrt{\mathbb{E}_u\,\langle u, w\rangle^4}$$
$$\le d^2 \cdot \frac{\sqrt{648}\,M_g(\xi)^2\kappa_m}{\sqrt{d(d+2)}} \cdot \frac{\sqrt{3}\,\|w\|^2}{\sqrt{d(d+2)}} \le \sqrt{1944}\,\kappa_m M_g(\xi)^2\,\|w\|^2.$$

Taking the expectation over ξ with $\mathbb{E}M_g(\xi)^2 \le M_g^2$ and $\sqrt{1944} < 45$ gives the first claim. The second follows from $\mathbb{E}\left\|(\hat{J}-J_\gamma)w\right\|_\infty^2 \le 2\mathbb{E}\left\|\hat{J}w\right\|_\infty^2 + 2\left\|J_\gamma w\right\|_\infty^2 \le (90\kappa_m+2)M_g^2\left\|w\right\|^2 \le 92\kappa_m M_g^2\left\|w\right\|^2$ by part (b). (All steps up to the Cauchy–Schwarz inequality are performed for fixed ξ, so that only the second moment of $M_g(\xi)$ is needed at the end; no assumption on higher moments of the Lipschitz modulus is required.) □

Proof of Theorem 3.6. The bound H is Jensen's inequality: $\left\|\frac{1}{b}\sum_j Z_j\right\|_\infty^2 \le \frac{1}{b}\sum_j \left\|Z_j\right\|_\infty^2$. For the second bound let $Z_1',\dots,Z_b'$ be an independent copy of $Z_1,\dots,Z_b$ and $\varepsilon_1,\dots,\varepsilon_b$ independent Rademacher signs, independent of everything else. Since $\mathbb{E}Z_j' = 0$ and $\|\cdot\|_\infty^2$ is convex, $\left\|\sum_j Z_j\right\|_\infty^2 = \left\|\mathbb{E}'[\sum_j(Z_j-Z_j')]\right\|_\infty^2 \le \mathbb{E}'\left\|\sum_j(Z_j-Z_j')\right\|_\infty^2$, where $\mathbb{E}'$ integrates over the copy; hence $\mathbb{E}\left\|\sum_j Z_j\right\|_\infty^2 \le \mathbb{E}\left\|\sum_j(Z_j-Z_j')\right\|_\infty^2 = \mathbb{E}\left\|\sum_j \varepsilon_j(Z_j-Z_j')\right\|_\infty^2$, the last equality because $Z_j - Z_j'$ is symmetric and the pairs are independent. Condition on $(Z_j,Z_j')_j$ and put $a_{ji} := (Z_j-Z_j')_i$, $X_i := \sum_j \varepsilon_j a_{ji}$ and $s^2 := \sum_j \left\|Z_j - Z_j'\right\|_\infty^2 \ge \sum_j a_{ji}^2$ for every i. By Hoeffding's lemma, $\mathbb{E}[e^{\theta X_i} \mid Z,Z'] \le e^{\theta^2 s^2/2}$ for all $\theta\in\mathbb{R}$, hence $\mathbb{P}(X_i^2 > t \mid Z,Z') \le 2e^{-t/(2s^2)}$. Therefore, for every $\alpha > 0$,

$$\mathbb{E}\Big[\max_i X_i^2 \;\Big|\; Z,Z'\Big] \le \alpha + \sum_{i=1}^m \int_\alpha^\infty \mathbb{P}(X_i^2 > t \mid Z,Z')\,dt \le \alpha + 4ms^2 e^{-\alpha/(2s^2)},$$

and the choice $\alpha = 2s^2\log(2m)$ gives $\mathbb{E}[\max_i X_i^2 \mid Z,Z'] \le 2s^2(1+\log(2m)) = 2\kappa_m s^2$. Finally $\mathbb{E}s^2 \le \sum_j 2(\mathbb{E}\left\|Z_j\right\|_\infty^2 + \mathbb{E}\left\|Z_j'\right\|_\infty^2) \le 4b\mathsf{H}$, so $\mathbb{E}\left\|\sum_j Z_j\right\|_\infty^2 \le 8\kappa_m b\mathsf{H}$; dividing by b^2 proves the claim. The conditional version is proved identically. □

C Proof of Theorem 5.1

Throughout this appendix $\mathcal{Z} := Q \times \tilde{Y}_R$ (we identify $y \in Y_R$ with its lifting $\tilde{y} \in \tilde{Y}_R$), V is the joint Bregman distance (23), $\|\cdot\|_\alpha$ and $\|\cdot\|_{\alpha,*}$ are the norms of (23), and the operator $\mathcal{G}$ of (22) is extended to $\mathcal{Z}$ by padding its dual component with a zero slack coordinate: $\mathcal{G}(z) = (\nabla f_\gamma(x) + J_\gamma(x)^\top y,\ -(0,h(x)))$. The estimator of an iteration is $\hat{\mathcal{G}}(z) = (\hat{g}, -(0,\hat{h}))$ with the mini-batch averages of Algorithm 1. By Theorem 4.4(a), $V(z,z') \ge \frac{1}{2}\|z-z'\|_\alpha^2$, and $\|\cdot\|_{\alpha,*}$ is the dual norm of $\|\cdot\|_\alpha$. The proximal steps of Algorithm 1 are $w = \arg\min_{z'\in\mathcal{Z}}\{\left\langle \eta\hat{\mathcal{G}}(z), z'\right\rangle + V(z,z')\}$ (the two blocks separate).

Lemma C.1 (Lipschitz constant and monotonicity)**.** *For all $z=(x,y), z'=(x',y') \in \mathcal{Z}$: $\|\mathcal{G}(z)-\mathcal{G}(z')\|_{\alpha,*} \le L_\alpha\|z-z'\|_\alpha$ with L_α from (24), and $\langle \mathcal{G}(z), z-z'\rangle \ge \Phi_\gamma(x,y') - \Phi_\gamma(x',y)$.*

Proof. By (4), Theorem 3.4(b) and $\|y\|_1 \le R$, $\|\nabla_x\Phi_\gamma(z) - \nabla_x\Phi_\gamma(z')\| \le \|\nabla f_\gamma(x) - \nabla f_\gamma(x')\| + \left\|(J_\gamma(x)-J_\gamma(x'))^\top y\right\| + \left\|J_\gamma(x')^\top(y-y')\right\| \le H\|x-x'\| + M_g\|y-y'\|_1$, and $\|h(x)-h(x')\|_\infty \le M_g\|x-x'\|$. With $\mathsf{a} := \sqrt{\alpha}\|x-x'\|$ and $\mathsf{c} := \|\tilde{y}-\tilde{y}'\|_1/\sqrt{\alpha} \ge \|y-y'\|_1/\sqrt{\alpha}$,

$$\begin{aligned}\left\|\mathcal{G}(z)-\mathcal{G}(z')\right\|_{\alpha,*}^2 &\le \frac{(H\|x-x'\| + M_g\|y-y'\|_1)^2}{\alpha} + \alpha M_g^2\left\|x-x'\right\|^2 \\ &\le \Big(\frac{H}{\alpha}\mathsf{a} + M_g\mathsf{c}\Big)^2 + (M_g\mathsf{a})^2 = \left\|\mathsf{B}(\mathsf{a},\mathsf{c})^\top\right\|^2,\end{aligned}$$

where $\mathsf{B} := \left(\begin{smallmatrix} H/\alpha & M_g \\ M_g & 0\end{smallmatrix}\right)$; hence $\|\mathcal{G}(z)-\mathcal{G}(z')\|_{\alpha,*} \le \|\mathsf{B}\|_2\sqrt{\mathsf{a}^2+\mathsf{c}^2} = \|\mathsf{B}\|_2\|z-z'\|_\alpha$, and the spectral norm of the symmetric matrix B is its largest eigenvalue $\frac{1}{2}(H/\alpha + \sqrt{H^2/\alpha^2 + 4M_g^2}) = L_\alpha$. For the second claim, convexity of $\Phi_\gamma(\cdot,y)$ gives $\langle\nabla_x\Phi_\gamma(x,y), x-x'\rangle \ge \Phi_\gamma(x,y) - \Phi_\gamma(x',y)$, and linearity in y gives $\langle -h(x), y-y'\rangle = \Phi_\gamma(x,y') - \Phi_\gamma(x,y)$; add. □

Lemma C.2 (one iteration)**.** *Let* $z_k \in \mathcal{Z}$, $w_k := \arg\min_{z\in\mathcal{Z}}\{\left\langle \eta\hat{\mathcal{G}}(z_k), z\right\rangle + V(z_k,z)\}$ *and* $z_{k+1} := \arg\min_{z\in\mathcal{Z}}\{\left\langle \eta\hat{\mathcal{G}}'(w_k), z\right\rangle + V(z_k,z)\}$, *where* $\hat{\mathcal{G}}(z_k) = \mathcal{G}(z_k)+\Delta_k$ *and* $\hat{\mathcal{G}}'(w_k) = \mathcal{G}(w_k)+\Delta'_k$. *If* $\eta \le 1/(\sqrt{3}L_\alpha)$, *then for every* $z \in \mathcal{Z}$

$$\eta\left\langle \hat{\mathcal{G}}'(w_k), w_k - z\right\rangle \le V(z_k,z) - V(z_{k+1},z) + \frac{3\eta^2}{2}\left(\|\Delta_k\|_{\alpha,*}^2 + \|\Delta'_k\|_{\alpha,*}^2\right).$$

Proof. Theorem A.6 for the two proximal steps gives, for all $z \in \mathcal{Z}$, $\eta\left\langle \hat{\mathcal{G}}'(w_k), z_{k+1} - z\right\rangle \le V(z_k,z) - V(z_{k+1},z) - V(z_k,z_{k+1})$ and, with $z = z_{k+1}$ in the first step, $\eta\left\langle \hat{\mathcal{G}}(z_k), w_k - z_{k+1}\right\rangle \le V(z_k,z_{k+1}) - V(w_k,z_{k+1}) - V(z_k,w_k)$. Adding and rearranging,

$$\eta\left\langle \hat{\mathcal{G}}'(w_k), w_k - z\right\rangle \le V(z_k,z) - V(z_{k+1},z) - V(w_k,z_{k+1}) - V(z_k,w_k) + \eta\left\langle \hat{\mathcal{G}}(z_k) - \hat{\mathcal{G}}'(w_k), w_k - z_{k+1}\right\rangle.$$

Put $\mathsf{A} := \|z_k - w_k\|_\alpha$, $\mathsf{B} := \|w_k - z_{k+1}\|_\alpha$ and $\delta := \|\Delta_k\|_{\alpha,*} + \|\Delta'_k\|_{\alpha,*}$. Since $\hat{\mathcal{G}}(z_k) - \hat{\mathcal{G}}'(w_k) = [\mathcal{G}(z_k) - \mathcal{G}(w_k)] + \Delta_k - \Delta'_k$, Theorem C.1 and the definition of the dual norm give $\eta\left\langle \hat{\mathcal{G}}(z_k) - \hat{\mathcal{G}}'(w_k), w_k - z_{k+1}\right\rangle \le \eta L_\alpha \mathsf{A}\mathsf{B} + \eta\delta\mathsf{B} \le \frac{1}{2}\mathsf{A}^2 + \frac{\eta^2 L_\alpha^2}{2}\mathsf{B}^2 + \frac{c}{2}\mathsf{B}^2 + \frac{\eta^2\delta^2}{2c}$ with $c := 1 - \eta^2 L_\alpha^2 \ge \frac{2}{3}$. Using $V(w_k,z_{k+1}) \ge \frac{1}{2}\mathsf{B}^2$ and $V(z_k,w_k) \ge \frac{1}{2}\mathsf{A}^2$, the four terms $-V(w_k,z_{k+1}) - V(z_k,w_k) + \frac{1}{2}\mathsf{A}^2 + \frac{1}{2}(\eta^2 L_\alpha^2 + c)\mathsf{B}^2$ are at most 0, and $\frac{\eta^2\delta^2}{2c} \le \frac{3\eta^2}{4}\delta^2 \le \frac{3\eta^2}{2}(\|\Delta_k\|_{\alpha,*}^2 + \|\Delta'_k\|_{\alpha,*}^2)$. □

Proof of Theorem 5.1. Let $\mathcal{F}_k$ denote the σ-field generated by x_1 and by the mini-batches of the iterations $1,\dots,k$, and $\mathcal{F}_{k-1/2}$ the one that additionally contains the first mini-batch of iteration k. Then z_k is $\mathcal{F}_{k-1}$-measurable, w_k is $\mathcal{F}_{k-1/2}$-measurable, $\mathbb{E}[\Delta_k \mid \mathcal{F}_{k-1}] = 0$ and $\mathbb{E}[\Delta'_k \mid \mathcal{F}_{k-1/2}] = 0$ by Theorem 3.4(a) and Theorem 3.5, and

$$\mathbb{E}\left[\|\Delta_k\|_{\alpha,*}^2 \mid \mathcal{F}_{k-1}\right] = \alpha^{-1}\mathbb{E}[\|\hat{g}_k - \nabla_x\Phi_\gamma(z_k)\|^2 \mid \mathcal{F}_{k-1}] + \alpha\,\mathbb{E}[\left\|\hat{h}_k - h(x_k)\right\|_\infty^2 \mid \mathcal{F}_{k-1}] \le \frac{2dM_R^2}{\alpha b} + \alpha\min\Big\{1, \frac{8\kappa_m}{b}\Big\}\sigma_h^2 = \sigma_*^2$$

by Theorem 3.4(a) (the b paired samples are conditionally i.i.d. and centered, with weight vector y_k, $\|y_k\|_1 \le R$), Theorem 3.5 and Theorem 3.6; the same bound holds for Δ'_k given $\mathcal{F}_{k-1/2}$.

By Theorem C.2 and Theorem C.1, for every $z = (x,\tilde{y}) \in \mathcal{Z}$,

$$\eta\left[\Phi_\gamma(x_k^w, y) - \Phi_\gamma(x, y_k^w)\right] \le \eta\left\langle \mathcal{G}(w_k), w_k - z\right\rangle = \eta\left\langle \hat{\mathcal{G}}'(w_k), w_k - z\right\rangle - \eta\left\langle \Delta'_k, w_k - z\right\rangle \le V(z_k,z) - V(z_{k+1},z) + \frac{3\eta^2}{2}\left(\|\Delta_k\|_{\alpha,*}^2 + \|\Delta'_k\|_{\alpha,*}^2\right) + \eta\left\langle \Delta'_k, z - w_k\right\rangle.$$

Summing over $k = 1,\dots,N$, dividing by ηN, and using convexity of $\Phi_\gamma(\cdot,y)$ and linearity of $\Phi_\gamma(x,\cdot)$,

$$\Phi_\gamma(\bar{x}_N, y) - \Phi_\gamma(x, \bar{y}_N) \le \frac{1}{N}\sum_{k=1}^N\left[\Phi_\gamma(x_k^w, y) - \Phi_\gamma(x, y_k^w)\right] \le \frac{V(z_1,z)}{\eta N} + \frac{3\eta}{2N}\sum_{k=1}^N\left(\|\Delta_k\|_{\alpha,*}^2 + \|\Delta'_k\|_{\alpha,*}^2\right) + \frac{1}{N}\sum_{k=1}^N\left\langle \Delta'_k, z - w_k\right\rangle.$$

Now fix $x := x_\gamma^*$ and let $Z := \{x_\gamma^*\} \times \tilde{Y}_R$. By (19), $\mathsf{C}_{R,\gamma}(\bar{x}_N) \le \sup_{y \in Y_R} \Phi_\gamma(\bar{x}_N, y) - \Phi_\gamma(x_\gamma^*, \bar{y}_N)$, and $\sup_{z \in Z} V(z_1, z) = \alpha\mathcal{A} + \alpha^{-1}\mathcal{B} = \mathcal{D}_\alpha^2$ by (21) ($\tilde{y}_1 = \tilde{y}^{\mathrm{u}}$). Hence

$$\mathsf{C}_{R,\gamma}(\bar{x}_N) \le \frac{\mathcal{D}_\alpha^2}{\eta N} + \frac{3\eta}{2N} \sum_{k=1}^{N} \left(\|\Delta_k\|_{\alpha,*}^2 + \left\|\Delta_k'\right\|_{\alpha,*}^2 \right) + \frac{1}{N} \sup_{z \in Z} \sum_{k=1}^{N} \left\langle \Delta_k', z - w_k \right\rangle .$$

Taking expectations, the second term is at most $3\eta\sigma_*^2$. For the third we apply Theorem 3.7 with the compact set Z, the ambient set $Z' := \mathcal{Z} = Q \times \tilde{Y}_R$ (the distance-generating function $\alpha\frac{1}{2}\|x\|^2 + \alpha^{-1}R\sum_i \tilde{y}_i \log \tilde{y}_i$ is 1-strongly convex for $\|\cdot\|_\alpha$ on $\mathcal{Z}$), the starting point $z_0 := z_1 = (x_1, \tilde{y}^{\mathrm{u}}) \in \mathcal{Z}$, which lies in Z' but in general not in Z (its primal part is $x_1 \neq x_\gamma^*$), and $\Theta := \sup_{z \in Z} V(z_1, z) = \mathcal{D}_\alpha^2$, with the filtration $(\mathcal{F}_{k-1/2})_k$, $e_k := \Delta_k'$ (which is $\mathcal{F}_{k+1/2}$-measurable and centered given $\mathcal{F}_{k-1/2}$), $\zeta_k := w_k$ (bounded and $\mathcal{F}_{k-1/2}$-measurable), $a_k := 1$ and $v_k^2 := \sigma_*^2$. It gives $\mathbb{E}\sup_{z \in Z} \sum_k \langle \Delta_k', z - w_k \rangle \le \sqrt{2\mathcal{D}_\alpha^2 N \sigma_*^2}$, and dividing by N yields (25).

For (26), if $\eta = \mathcal{D}_\alpha/(\sigma_*\sqrt{3N}) \le 1/(\sqrt{3}L_\alpha)$ then $\mathcal{D}_\alpha^2/(\eta N) = 3\eta\sigma_*^2 = \sqrt{3}\sigma_*\mathcal{D}_\alpha/\sqrt{N}$ and the right-hand side of (25) equals $(2\sqrt{3} + \sqrt{2})\sigma_*\mathcal{D}_\alpha/\sqrt{N} \le 5\sigma_*\mathcal{D}_\alpha/\sqrt{N}$. Otherwise $\eta = 1/(\sqrt{3}L_\alpha) < \mathcal{D}_\alpha/(\sigma_*\sqrt{3N})$, i.e. $\sigma_*\sqrt{N} < L_\alpha\mathcal{D}_\alpha$, and then $\mathcal{D}_\alpha^2/(\eta N) = \sqrt{3}L_\alpha\mathcal{D}_\alpha^2/N$, $3\eta\sigma_*^2 = \sqrt{3}\sigma_*^2/L_\alpha \le \sqrt{3}\sigma_*\mathcal{D}_\alpha/\sqrt{N}$, so the right-hand side is at most $\sqrt{3}L_\alpha\mathcal{D}_\alpha^2/N + (\sqrt{3} + \sqrt{2})\sigma_*\mathcal{D}_\alpha/\sqrt{N}$. □

Remark C.3 (random starting point). If x_1 is random and independent of the mini-batches of the run, the proof above applies conditionally on x_1 with the random radius $\mathcal{D}_\alpha^2 = \alpha\mathcal{A} + \alpha^{-1}\mathcal{B}$, $\mathcal{A} = \frac{1}{2}\left\|x_1 - x_\gamma^*\right\|^2$, and any deterministic $\eta \le 1/(\sqrt{3}L_\alpha)$:

$$\mathbb{E}\left[\mathsf{C}_{R,\gamma}(\bar{x}_N) \mid x_1\right] \le \frac{\alpha\mathcal{A} + \mathcal{B}/\alpha}{\eta N} + 3\eta\sigma_*^2 + \sqrt{\frac{2(\alpha\mathcal{A} + \mathcal{B}/\alpha)\sigma_*^2}{N}}.$$

The right-hand side is a nondecreasing concave function of $\mathcal{A}$; this is the form used in the restart analysis of Theorem 7.3.

D Proofs for Section 6

D.1 Proof of Lemma 6.1

Recall $S_t = \max\{1, \lceil ct \rceil\}$ with $c = M_g\varrho/L$, $a_t = t/S_t$.

(i) $S_t \ge \lceil ct \rceil \ge ct$, hence $a_t \le 1/c = L/(M_g\varrho)$.

(ii) Let $t \ge 2$. If $\lceil ct \rceil \le 1$ then $S_t = S_{t-1} = 1$ and $a_{t-1}/a_t = (t-1)/t < 1$. Otherwise $ct > 1$ and $S_t = \lceil ct \rceil < ct + 1$. If $c(t-1) \ge 1$, then $S_{t-1} \ge c(t-1)$ and $a_{t-1}/a_t = \frac{(t-1)S_t}{tS_{t-1}} \le \frac{(t-1)(ct+1)}{t\,c(t-1)} = 1 + \frac{1}{ct} < 2$. If $c(t-1) < 1$, then $S_{t-1} = 1$, $c < 1/(t-1)$, and $a_{t-1}/a_t \le \frac{(t-1)(ct+1)}{t} < \frac{(t-1)}{t}\left(\frac{t}{t-1} + 1\right) = \frac{2t-1}{t} < 2$.

(iii) From $\max\{1, ct\} \le S_t \le 1 + ct$ we get $\max\{N, \frac{c}{2}N(N+1)\} \le K_N \le N + \frac{c}{2}N(N+1)$.

(iv) The lower bound is the Cauchy–Schwarz inequality $W_N^2 = (\sum_t t)^2 = (\sum_t \frac{t}{\sqrt{S_t}}\sqrt{S_t})^2 \le A_N K_N$. For the upper bound, note $t^2/S_t \le \min\{t^2, t/c\}$. If $cN \le 1$, then $S_t = 1$ for all $t \le N$, $K_N = N$, $A_N = \frac{N(N+1)(2N+1)}{6}$ and $A_N K_N / W_N^2 = \frac{2(2N+1)}{3(N+1)} \le \frac{4}{3}$. If $cN > 1$, let $T_0 := \lfloor 1/c \rfloor \le N - 1$. Then

$$A_N \le \sum_{t \le T_0} t^2 + \frac{1}{c} \sum_{t=T_0+1}^{N} t \le \frac{T_0(T_0+1)(2T_0+1)}{6} + \frac{N(N+1)}{2c}$$
$$\le \frac{(T_0+1)(2T_0+1)}{6c} + \frac{N(N+1)}{2c} \le \frac{5N(N+1)}{6c},$$

using $T_0 \le 1/c$, $T_0 + 1 \le N$ and $2T_0 + 1 \le 2N$. Since $K_N \ge \frac{c}{2}N(N+1)$, i.e. $\frac{1}{c} \le \frac{N(N+1)}{2K_N}$, we get $A_N \le \frac{5N^2(N+1)^2}{12K_N} = \frac{5}{3}\frac{W_N^2}{K_N}$.

D.2 Proof of Lemma 6.2

For $t = 1$, $\tau_1 = \theta_1 = 0$ gives $u_1 = x_0$. Multiplying the recursion by $t(1+\tau_t) = t(t+1)/2$ and using $t\tau_t = t(t-1)/2$, $t\theta_t = t-1$,

$$t(t+1)u_t = t(t-1)u_{t-1} + 2t\,x_{t-1} + 2(t-1)(x_{t-1} - x_{t-2}).$$

For $t = 2$ this gives $6u_2 = 2x_0 + 4x_1 + 2(x_1 - x_0) = 6x_1$, which is (33) (empty sum). Assume (33) for $t-1 \geq 2$, i.e. $(t-1)t\,u_{t-1} = \sum_{j\leq t-3} 2jx_j + (4t-6)x_{t-2}$. Then $t(t+1)u_t = \sum_{j\leq t-3} 2jx_j + (4t-6)x_{t-2} - 2(t-1)x_{t-2} + (2t+2t-2)x_{t-1} = \sum_{j\leq t-3} 2jx_j + 2(t-2)x_{t-2} + (4t-2)x_{t-1}$, which is (33) for t. The coefficients are nonnegative and sum to $(t-2)(t-1) + 4t - 2 = t(t+1)$, so u_t is a convex combination of $x_1, \ldots, x_{t-1}$. Finally, the recursion gives $x_{t-1} = (1+\tau_t)u_t - \tau_t u_{t-1} - \theta_t d_{t-1}$, hence $x_t - u_t = d_t + x_{t-1} - u_t = d_t - \theta_t d_{t-1} + \tau_t(u_t - u_{t-1})$.

D.3 Proof of Lemma 6.3

Write $g_t := \nabla \mathsf{F}(u_t)$, $\pi := \nabla \mathsf{F}(\bar{x}_N)$, and let F^* be the conjugate of F. By Theorem A.7, $\mathsf{F}^*(g_t) = \langle g_t, u_t\rangle - \mathsf{F}(u_t)$, so that $\ell_t^{\mathsf{F}}(x) = \langle g_t, x\rangle - \mathsf{F}^*(g_t)$, and the "Bregman distances" $D_t(\pi') := \mathsf{F}^*(\pi') - \mathsf{F}^*(g_t) - \langle u_t, \pi' - g_t\rangle$ satisfy

$$D_t(\pi') \geq \frac{1}{2L}\left\|\pi' - g_t\right\|^2 \geq 0 \qquad \text{for all } \pi' \in \operatorname{dom}\mathsf{F}^*. \tag{56}$$

Fenchel–Young at $\bar{x}_N$ gives $\mathsf{F}(\bar{x}_N) = \langle \pi, \bar{x}_N\rangle - \mathsf{F}^*(\pi)$, and $W_N\bar{x}_N = \sum_t t x_t$; therefore

$$W_N\mathsf{F}(\bar{x}_N) - \sum_{t=1}^N t\,\ell_t^{\mathsf{F}}(x_t) = \sum_{t=1}^N t\big[\langle \pi - g_t, x_t\rangle - \mathsf{F}^*(\pi) + \mathsf{F}^*(g_t)\big] = \sum_{t=1}^N t\big[\langle \pi - g_t, x_t - u_t\rangle - D_t(\pi)\big],$$

where we used $\mathsf{F}^*(\pi) - \mathsf{F}^*(g_t) = D_t(\pi) + \langle u_t, \pi - g_t\rangle$. Insert (34), $x_t - u_t = d_t - \theta_t d_{t-1} + \tau_t(u_t - u_{t-1})$, and the identity

$$\langle \pi - g_t, u_t - u_{t-1}\rangle = D_{t-1}(\pi) - D_t(\pi) - D_{t-1}(g_t) \qquad (t \geq 2),$$

which is checked by expanding the three definitions (the terms $\mathsf{F}^*(\pi)$, $\mathsf{F}^*(g_{t-1})$ and $\langle u_{t-1}, \pi\rangle$ cancel). For $t = 1$ we have $\tau_1 = 0$, so the identity is not needed. We obtain

$$\begin{aligned} W_N\mathsf{F}(\bar{x}_N) - \sum_t t\,\ell_t^{\mathsf{F}}(x_t) &= \sum_{t=1}^N t\,\langle \pi - g_t, d_t - \theta_t d_{t-1}\rangle \\ &\quad + \sum_{t=2}^N \big[t\tau_t D_{t-1}(\pi) - t(1+\tau_t)D_t(\pi) - t\tau_t D_{t-1}(g_t)\big] - 1\cdot(1+\tau_1)D_1(\pi). \end{aligned}$$

Since $t(1+\tau_t) = t(t+1)/2 = (t+1)\tau_{t+1}$, the $D(\pi)$-terms telescope: $\sum_{t=1}^N [t\tau_t D_{t-1}(\pi) - (t+1)\tau_{t+1}D_t(\pi)] = -(N+1)\tau_{N+1}D_N(\pi) = -W_N D_N(\pi)$ (recall $\tau_1 = 0$). For the first sum, $t\theta_t = t-1$ and a shift of the index give

$$\sum_{t=1}^N t\,\langle \pi - g_t, d_t\rangle - \sum_{t=2}^N (t-1)\,\langle \pi - g_t, d_{t-1}\rangle = N\,\langle \pi - g_N, d_N\rangle + \sum_{t=1}^{N-1} t\,\langle g_{t+1} - g_t, d_t\rangle\,.$$

Hence

$$\begin{aligned} &W_N\mathsf{F}(\bar{x}_N) - \sum_t t\,\ell_t^{\mathsf{F}}(x_t) \\ &\qquad = \Big[N\,\langle \pi - g_N, d_N\rangle - W_N D_N(\pi)\Big] + \sum_{t=1}^{N-1}\Big[t\,\langle g_{t+1} - g_t, d_t\rangle - (t+1)\tau_{t+1}D_t(g_{t+1})\Big]. \end{aligned}$$

By (56) and the elementary bound $\max_{r\ge 0}\{\mathsf{p}r-\mathsf{q}r^2\}=\mathsf{p}^2/(4\mathsf{q})$ for $\mathsf{p},\mathsf{q}>0$,

$$N\langle \pi-g_N,d_N\rangle - W_N D_N(\pi) \le N\|\pi-g_N\|\,\|d_N\| - \frac{W_N}{2L}\|\pi-g_N\|^2 \\ \le \frac{LN^2\|d_N\|^2}{2W_N} = \frac{LN}{N+1}\|d_N\|^2 \le L\|d_N\|^2,$$

and, with $(t+1)\tau_{t+1}=t(t+1)/2$,

$$t\langle g_{t+1}-g_t,d_t\rangle - \frac{t(t+1)}{2}D_t(g_{t+1}) \le t\|g_{t+1}-g_t\|\,\|d_t\| - \frac{t(t+1)}{4L}\|g_{t+1}-g_t\|^2 \\ \le \frac{Lt}{t+1}\|d_t\|^2 \le L\|d_t\|^2.$$

Summing proves (35).

D.4 Proof of Lemma 6.4

Fix $y\in Y_R$ and write $x:=x^*_\gamma$, $\Delta y_k:=y_k-y_{k-1}$ ($\Delta y_0=0$), $\Delta p_k:=p_k-p_{k-1}$, $K:=K_N$, and $t(k)$ for the phase of slot k. All Bregman distances below are $V_{\mathcal{X}}$ and $V_{\mathcal{Y}}$ of (20).

Step 1: three-point inequalities. By Theorem 4.4(d) applied to the primal step with the direction $\hat r_k=r_k+e^p_k$, for all $x'\in Q$,

$$\langle r_k,p_k-x'\rangle \le \eta_{t(k)}\big[V_{\mathcal{X}}(x_{t(k)-1},x') - V_{\mathcal{X}}(x_{t(k)-1},p_k) - V_{\mathcal{X}}(p_k,x')\big] \\ + \beta_{t(k)}\big[V_{\mathcal{X}}(p_{k-1},x') - V_{\mathcal{X}}(p_{k-1},p_k) - V_{\mathcal{X}}(p_k,x')\big] - \langle e^p_k,p_k-x'\rangle,$$

and by Theorem 4.4(c) applied to the dual step with $\hat q_k=q_k+e^y_k$ and $\tau=\beta_{t(k)}/\varrho^2$, for all $y'\in Y_R$,

$$\langle q_k,y'-y_k\rangle \le \frac{\beta_{t(k)}}{\varrho^2}\big[V_{\mathcal{Y}}(\tilde y_{k-1},\tilde y') - V_{\mathcal{Y}}(\tilde y_k,\tilde y') - V_{\mathcal{Y}}(\tilde y_{k-1},\tilde y_k)\big] + \langle e^y_k,y_k-y'\rangle.$$

Step 2: the linearized gap and the cross terms. From (30) and (31), a direct computation gives, for every slot k of phase t,

$$\ell_t(p_k,y)-\ell_t(x,y_k) = \langle r_k,p_k-x\rangle + \langle q_k,y-y_k\rangle + \langle \Delta y_k, J_t(p_k-x)\rangle - \rho_k\big\langle \Delta y_{k-1}, J_{t'(k)}(p_k-x)\big\rangle.$$

(Indeed, $\ell_t(p_k,y)-\ell_t(x,y_k)=\langle\nabla f_\gamma(u_t),p_k-x\rangle+\langle y,q_k\rangle-\langle y_k,h(u_t)+J_t(p_k-u_t)\rangle+\langle y_k,J_t(p_k-x)\rangle$, while $\langle r_k,p_k-x\rangle+\langle q_k,y-y_k\rangle$ equals the same expression with $\langle y_k,J_t(p_k-x)\rangle$ replaced by $\langle y_{k-1},J_t(p_k-x)\rangle+\rho_k\big\langle\Delta y_{k-1},J_{t'(k)}(p_k-x)\big\rangle$.) Define $C_k:=a_k\big\langle\Delta y_k,J_{t(k)}(p_k-x)\big\rangle$ and $C_0:=0$. For $k\ge 2$ we have $a_k\rho_k=a_{k-1}$ and $J_{t'(k)}=J_{t(k-1)}$ in both cases of (31) (for a non-first slot $\rho_k=1$, $a_k=a_{k-1}$, $t'(k)=t(k)=t(k-1)$; for the first slot of phase $t\ge 2$, $a_k\rho_k=a_t\cdot a_{t-1}/a_t=a_{t-1}=a_{k-1}$ and $t'(k)=t-1=t(k-1)$), so that $a_k\rho_k\big\langle\Delta y_{k-1},J_{t'(k)}(p_k-x)\big\rangle = C_{k-1}+a_{k-1}\big\langle\Delta y_{k-1},J_{t(k-1)}\Delta p_k\big\rangle$; for $k=1$ both sides vanish. Multiplying the identity by a_k and summing over k,

$$\sum_{k=1}^{K} a_k\big[\ell_{t(k)}(p_k,y)-\ell_{t(k)}(x,y_k)\big] \\ = \sum_k a_k\langle r_k,p_k-x\rangle + \sum_k a_k\langle q_k,y-y_k\rangle + C_K - \sum_{k=2}^{K} a_{k-1}\big\langle\Delta y_{k-1},J_{t(k-1)}\Delta p_k\big\rangle.$$

By (4), Theorem 3.4(b), Theorem 6.1(i) and Young's inequality,

$$\begin{aligned}\left|a_{k-1}\left\langle \Delta y_{k-1}, J_{t(k-1)}\Delta p_k\right\rangle\right| &\le a_{k-1}M_g\varrho\cdot\frac{\|\Delta y_{k-1}\|_1}{\varrho}\|\Delta p_k\| \le \frac{L}{2}\Big[\frac{\|\Delta y_{k-1}\|_1^2}{\varrho^2}+\|\Delta p_k\|^2\Big],\\ |C_K| &\le \frac{L}{2}\Big[\frac{\|\Delta y_K\|_1^2}{\varrho^2}+\|p_K-x\|^2\Big].\end{aligned} \tag{57}$$

Since ℓ_t is affine in each argument and $\sum_{k\in\text{phase }t} a_t = t$, the left-hand side equals $\sum_t t[\ell_t(x_t,y)-\ell_t(x,\bar y_t)]$. Therefore

$$\begin{aligned}\sum_t t\big[\ell_t(x_t,y)-\ell_t(x,\bar y_t)\big] \le \sum_k a_k\langle r_k, p_k-x\rangle + \sum_k a_k\langle q_k, y-y_k\rangle\\ + \frac{L}{2\varrho^2}\sum_{k=1}^K\|\Delta y_k\|_1^2 + \frac{L}{2}\sum_{k=2}^K\|\Delta p_k\|^2 + \frac{L}{2}\|p_K-x\|^2.\end{aligned} \tag{58}$$

Step 3: telescoping. Insert Step 1 with $x'=x$ and $y'=y$. Since $a_k\beta_{t(k)} = L+\lambda$ for all k and $p_0=x_0$,

$$\sum_k a_k\beta_{t(k)}\big[V_{\mathcal X}(p_{k-1},x)-V_{\mathcal X}(p_k,x)-V_{\mathcal X}(p_{k-1},p_k)\big] = (L+\lambda)\big[\mathcal A - V_{\mathcal X}(p_K,x)\big] - \frac{L+\lambda}{2}\sum_{k=1}^K\|\Delta p_k\|^2.$$

For the outer terms, $a_t\eta_t S_t = t\eta_t = 2L$ and the convexity of $V_{\mathcal X}$ in each argument (Jensen over the S_t slots of phase t, with x_t the average of the p_k) give

$$\begin{aligned}&\sum_k a_k\eta_{t(k)}\big[V_{\mathcal X}(x_{t-1},x)-V_{\mathcal X}(x_{t-1},p_k)-V_{\mathcal X}(p_k,x)\big]\\ &\quad\le 2L\sum_{t=1}^N\big[V_{\mathcal X}(x_{t-1},x)-V_{\mathcal X}(x_{t-1},x_t)-V_{\mathcal X}(x_t,x)\big] = 2L\big[\mathcal A - V_{\mathcal X}(x_N,x)\big] - L\sum_{t=1}^N\|d_t\|^2.\end{aligned}$$

For the dual terms, $a_k\beta_{t(k)}/\varrho^2 = (L+\lambda)/\varrho^2$, $V_{\mathcal Y}(\tilde y_0,\tilde y)\le\mathcal B$ and Theorem 4.4(a) give

$$\sum_k\frac{a_k\beta_{t(k)}}{\varrho^2}\big[V_{\mathcal Y}(\tilde y_{k-1},\tilde y)-V_{\mathcal Y}(\tilde y_k,\tilde y)-V_{\mathcal Y}(\tilde y_{k-1},\tilde y_k)\big] \le \frac{L+\lambda}{\varrho^2}\mathcal B - \frac{L+\lambda}{2\varrho^2}\sum_{k=1}^K\|\Delta y_k\|_1^2.$$

Step 4: collecting. Substituting into (58) and dropping $-2LV_{\mathcal X}(x_N,x)\le 0$, the terms in $\|p_K-x\|^2$ combine to $-(L+\lambda)V_{\mathcal X}(p_K,x)+\frac{L}{2}\|p_K-x\|^2 = -\frac{\lambda}{2}\|p_K-x\|^2\le 0$, the terms in $\|\Delta p_k\|^2$ to $-\frac{L+\lambda}{2}\sum_{k=1}^K\|\Delta p_k\|^2+\frac{L}{2}\sum_{k=2}^K\|\Delta p_k\|^2\le-\frac{\lambda}{2}\sum_{k=1}^K\|\Delta p_k\|^2$, and the terms in $\|\Delta y_k\|_1^2$ to $-\frac{\lambda}{2\varrho^2}\sum_{k=1}^K\|\Delta y_k\|_1^2$. What remains is

$$\begin{aligned}&\sum_t t\big[\ell_t(x_t,y)-\ell_t(x,\bar y_t)\big] + L\sum_t\|d_t\|^2 + \frac{\lambda}{2}\sum_k\Big[\|\Delta p_k\|^2+\frac{\|\Delta y_k\|_1^2}{\varrho^2}\Big]\\ &\qquad\le (3L+\lambda)\mathcal A + \frac{(L+\lambda)\mathcal B}{\varrho^2} - \sum_k a_k\langle e_k^p, p_k-x\rangle + \sum_k a_k\langle e_k^y, y_k-y\rangle,\end{aligned}$$

and $(3L+\lambda)\mathcal A+(L+\lambda)\mathcal B/\varrho^2\le(L+\lambda)(3\mathcal A+\mathcal B/\varrho^2) = (L+\lambda)D_A^2$. This is (36).

D.5 Proof details for Theorem 8.1

In Algorithm 4 the exact directions are $r_k=\nabla f_\gamma(u_t)+A^\top y_{k-1}+\rho_k A^\top\Delta y_{k-1}$ and $q_k = Ap_k-c$, the errors are $e_k^p=\hat e_{t(k)} := \hat g_{t(k)}-\nabla f_\gamma(u_{t(k)})$ and $e_k^y=0$. Steps 1 and 2 of the previous proof

apply verbatim with $J_t = A$, $M_g = M_A$ (Theorem 6.1(i) holds with $S_t = \max\{1, \lceil M_A \varrho t/L \rceil\}$), and (57) requires $\frac{L}{2}$ per increment on both sides. In Step 3 the inner coefficients are $a_k \beta_{t(k)} = L$, so the primal inner terms telescope to $L[\mathcal{A} - V_{\mathcal{X}}(p_K, x)] - \frac{L}{2}\sum_k \|\Delta p_k\|^2$ and the dual terms to $\frac{L}{\varrho^2}\mathcal{B} - \frac{L}{2\varrho^2}\sum_k \|\Delta y_k\|_1^2$; these exactly cancel the cross terms of (57) (no negative increment terms remain, which is why the stabilizer is not needed in the inner loop). The outer terms have $t\eta_t = 2L + \lambda$ and telescope to $(2L+\lambda)[\mathcal{A} - V_{\mathcal{X}}(x_N, x)] - (L + \frac{\lambda}{2})\sum_t \|d_t\|^2$, so that Step 4 yields

$$\sum_t t\big[\ell_t(x_t, y) - \ell_t(x, \bar{y}_t)\big] + \Big(L + \frac{\lambda}{2}\Big)\sum_t \|d_t\|^2 \le (3L+\lambda)\mathcal{A} + \frac{L\mathcal{B}}{\varrho^2} - \sum_k a_k \left\langle e_k^p, p_k - x\right\rangle,$$

as claimed in the proof of Theorem 8.1. Finally, $\sum_{k \in \text{phase } t} a_t \left\langle \hat{e}_t, p_k - x\right\rangle = t\left\langle \hat{e}_t, x_t - x\right\rangle$ because e_k^p is the same vector for all slots of the phase.

E Proofs for Section 9

E.1 Proof of Theorem 9.1

The family. For $v \in \{-1,1\}^d$ let $\theta_v := av$ with a parameter $a \in (0, M/(2\sqrt{d})]$ fixed below, so that $\|\theta_v\| = a\sqrt{d} \le M/2$, and let $\xi \sim \mathcal{N}(\theta_v, s^2 I_d)$ with $s^2 = M^2/(2d)$. Then $F(x,\xi) = \langle \xi, x\rangle$ is $\|\xi\|$-Lipschitz with $\mathbb{E}\,\|\xi\|^2 = a^2 d + s^2 d \le M^2/4 + M^2/2 < M^2$, so Theorem 2.2 holds with $M_0 = M$; the constraints $G_i \equiv -1$ are trivially Lipschitz, noiseless and satisfy Slater's condition with $\rho = 1$. The objective is $f_v(x) = a\langle v, x\rangle$, and on $Q = R_x\mathbb{B}$ its minimum is $f_v^* = -aR_x\sqrt{d}$, attained at $-R_x v/\sqrt{d}$.

From the gap to a Hamming distance. For $\hat{x} \in Q$ let $\hat{v}_j := -\operatorname{sign}(\hat{x}_j)$ (with $\operatorname{sign}(0) := 1$) and $\mathsf{J} := \{j : \hat{v}_j = v_j\}$, so that $v_j \hat{x}_j \le 0$ for $j \in \mathsf{J}$ and $v_j\hat{x}_j \ge 0$ otherwise. Then $\langle v, \hat{x}\rangle \ge \sum_{j\in\mathsf{J}} v_j\hat{x}_j \ge -\sum_{j\in\mathsf{J}} |\hat{x}_j| \ge -\sqrt{|\mathsf{J}|}\,\|\hat{x}\| \ge -R_x\sqrt{d - \operatorname{Ham}(\hat{v}, v)}$, whence

$$f_v(\hat{x}) - f_v^* = a\big(\langle v, \hat{x}\rangle + R_x\sqrt{d}\big) \ge aR_x\big(\sqrt{d} - \sqrt{d - \operatorname{Ham}(\hat{v}, v)}\big) \ge \frac{aR_x}{2\sqrt{d}}\operatorname{Ham}(\hat{v}, v),$$

using $\sqrt{d} - \sqrt{d-k} = k/(\sqrt{d} + \sqrt{d-k}) \ge k/(2\sqrt{d})$.

Assouad's argument. Let P_v be the law of the transcript of the algorithm (internal seed, queries and observations) under θ_v; $\hat{v}$ is a measurable function of the transcript. For $j \in \{1, \dots, d\}$ and v let $v^{(j)}$ denote v with the j-th coordinate flipped, and let $P_{+j} := 2^{-(d-1)}\sum_{v : v_j = 1} P_v$, $P_{-j} := 2^{-(d-1)}\sum_{v : v_j = -1} P_v$. Then

$$\frac{1}{2^d}\sum_v P_v(\hat{v}_j \ne v_j) = \frac{1}{2}\big[P_{+j}(\hat{v}_j = -1) + P_{-j}(\hat{v}_j = 1)\big] \ge \frac{1}{2}\big(1 - \mathrm{TV}(P_{+j}, P_{-j})\big),$$

because $P(A) + P'(A^c) \ge 1 - \mathrm{TV}(P, P')$ for every event A. Summing over j and using the joint convexity of TV (the two mixtures are averages over the matched pairs $(v, v^{(j)})$, $v_j = 1$),

$$\frac{1}{2^d}\sum_v \mathbb{E}_v \operatorname{Ham}(\hat{v}, v) \ge \frac{d}{2} - \frac{1}{2}\sum_{j=1}^d \mathrm{TV}(P_{+j}, P_{-j}) \ge \frac{d}{2} - \frac{1}{2}\cdot\frac{1}{2^d}\sum_v\sum_{j=1}^d \mathrm{TV}\big(P_v, P_{v^{(j)}}\big),$$

where each unordered pair $\{v, v^{(j)}\}$ appears twice in the double sum, which accounts for the factor 2^{-d} (instead of $2^{-(d-1)}$) and the symmetry of TV.

Information per query. Fix v and j. The observation of a paired query at (x^+, x^-) is $O = (\langle \xi, x^+\rangle, \langle \xi, x^-\rangle) = X^\top \xi$ with $X := [x^+\ x^-] \in \mathbb{R}^{d\times 2}$ (a single query is the case $x^+ = x^-$), a Gaussian vector with mean $X^\top\theta$ and covariance $s^2 X^\top X$. For two means θ, θ' the Kullback–Leibler divergence between the corresponding observation laws is $\frac{1}{2s^2}\|\Pi_X(\theta - \theta')\|^2$, where $\Pi_X = X(X^\top X)^+ X^\top$ is the orthogonal projector onto the span of x^+, x^- (this is the standard formula for

Gaussians with a common, possibly singular, covariance and a mean difference in its range). For $\theta = \theta_v$, $\theta' = \theta_{v^{(j)}}$ the difference is $\pm 2ae_j$, so the divergence equals $\frac{2a^2}{s^2}\|\Pi_X e_j\|^2$. By the chain rule (Theorem A.5), $\mathrm{KL}(P_v\|P_{v^{(j)}}) = \frac{2a^2}{s^2}\mathbb{E}_v\sum_{t=1}^{T}\|\Pi_{X_t}e_j\|^2$, where X_t is the (random) t-th query. For fixed v, Pinsker's inequality, the Cauchy–Schwarz inequality over j, and $\sum_j\|\Pi_X e_j\|^2 = \operatorname{tr}\Pi_X \le 2$ give

$$\sum_{j=1}^{d}\mathrm{TV}\big(P_v, P_{v^{(j)}}\big) \le \sum_{j=1}^{d}\sqrt{\tfrac{1}{2}\,\mathrm{KL}(P_v\|P_{v^{(j)}})} \le \sqrt{\frac{d}{2}\sum_{j=1}^{d}\mathrm{KL}(P_v\|P_{v^{(j)}})}$$
$$= \sqrt{\frac{da^2}{s^2}\,\mathbb{E}_v\sum_{t=1}^{T}\operatorname{tr}\Pi_{X_t}} \le \frac{a\sqrt{2dT}}{s}.$$

(The expectation $\mathbb{E}_v$ is common to all j, which is what allows the trace identity to be used inside it.) Consequently $2^{-d}\sum_v \mathbb{E}_v\,\mathrm{Ham}(\hat{v}, v) \ge \frac{1}{2}\big(d - a\sqrt{2dT}/s\big)$.

Choice of a. If $T \ge d/4$, take $a := s\sqrt{d/(8T)} = M/(4\sqrt{T})$; then $a\sqrt{d} = M\sqrt{d/T}/4 \le M/2$ and $a\sqrt{2dT}/s = d/2$, so the average Hamming distance is at least $d/4$ and the average gap at least $\frac{aR_x}{2\sqrt{d}}\cdot\frac{d}{4} = \frac{aR_x\sqrt{d}}{8} = \frac{MR_x}{32}\sqrt{\frac{d}{T}}$. If $T < d/4$, take $a := M/(2\sqrt{d})$; then $a\sqrt{2dT}/s = \sqrt{dT} < d/2$, the average Hamming distance exceeds $d/4$, and the average gap exceeds $\frac{aR_x\sqrt{d}}{8} = \frac{MR_x}{16} \ge \frac{MR_x}{32}$. In both cases some v attains at least the average, which proves the theorem. For the consequence, note that $\frac{MR_x}{32}\min\{1, \sqrt{d/T}\} \le \varepsilon < \frac{MR_x}{32}$ forces $\sqrt{d/T} \le 32\varepsilon/(MR_x)$.

E.2 Proof of Theorem 9.2

Let $\varphi_v(x) := \langle\theta_v, x\rangle$ with θ_v as above, and consider the problem $\min\{f(z) = Mt : \ g(z) = \varphi_v(x) - Mt \le 0,\ z = (x,t) \in Q'\}$. Its optimal value is $f^* = \min_{x\in R_x\mathbb{B}}\varphi_v(x) = \varphi_v^* = -\|\theta_v\|R_x \in [-MR_x/2, 0]$, attained at $t^* = \varphi_v^*/M \in [-R_x/2, 0]$, which lies inside the interval $[-2R_x, 2R_x]$ allowed for t. The stated constants are verified as follows. The diameter of $Q' = R_x\mathbb{B}\times[-2R_x, 2R_x]$ is $\sqrt{(2R_x)^2 + (4R_x)^2} = 2\sqrt{5}R_x$. The objective $f(z) = Mt$ is deterministic and M-Lipschitz, so $M_0 = M$. $G(\cdot,\xi)$ is $\|(\xi,-M)\|$-Lipschitz in z, with $\mathbb{E}\,\|(\xi,-M)\|^2 = \mathbb{E}\,\|\xi\|^2 + M^2 \le \frac{3}{4}M^2 + M^2 = \frac{7}{4}M^2$, so $M_g = \frac{\sqrt{7}}{2}M$; $G(z,\xi) - g(z) = \langle\xi - \theta_v, x\rangle \sim \mathcal{N}(0, s^2\|x\|^2)$ with $\|x\| \le R_x + \bar{\gamma}$ on $Q'_{\bar{\gamma}}$, so $\sigma_g^2 = s^2(R_x + \bar{\gamma})^2$; and $g(0, R_x) = -MR_x$ gives Slater's condition with $\rho = MR_x$. The Lagrangian is $Mt + y(\varphi_v(x) - Mt) = M(1-y)t + y\varphi_v(x)$; for $y < 1$ its minimum over Q' is attained at $t = -2R_x$ and equals $-2MR_x(1-y) + y\varphi_v^* < \varphi_v^*$, for $y > 1$ at $t = 2R_x$ with value $2MR_x(1-y) + y\varphi_v^* < \varphi_v^*$, and for $y = 1$ it equals $\varphi_v^* = f^*$; hence the dual function is maximized at $y^* = 1$, the active multiplier equals 1, and $R = 2$ is admissible in (14). Finally $\frac{MR_x}{32} = \frac{2M_g}{\sqrt{7}}\cdot\frac{D}{2\sqrt{5}}\cdot\frac{1}{32} = \frac{M_g D}{32\sqrt{35}}$ and $dM^2R_x^2/4096 = dM_g^2D^2/(4096\cdot 35) = dM_g^2D^2/143360$.

A paired constraint query at (z^+, z^-) returns $\langle\xi, x^\pm\rangle - Mt^\pm$; since $t^\pm$ are known to the algorithm, this is equivalent to observing $X^\top\xi$ with $X = [x^+\ x^-]$, exactly as in the proof of Theorem 9.1; objective queries return the known value Mt and carry no information. Therefore any algorithm for the constrained problem with output $\hat{z} = (\hat{x}, \hat{t})$ induces an algorithm for minimizing φ_v over $R_x\mathbb{B}$ with output $\hat{x}$ and the same transcript laws, and Theorem 9.1 (whose proof only used the transcript laws and the output) yields $\max_v \mathbb{E}_v[\varphi_v(\hat{x}) - \varphi_v^*] \ge \frac{MR_x}{32}\min\{1, \sqrt{d/T}\}$. Finally, $\varphi_v(\hat{x}) = M\hat{t} + g(\hat{z}) \le M\hat{t} + [g(\hat{z})]_+$, so $\varphi_v(\hat{x}) - \varphi_v^* \le (M\hat{t} - f^*) + [g(\hat{z})]_+ = (f(\hat{z}) - f^*) + [g(\hat{z})]_+$, and taking expectations gives the claim. For the consequence, (3) gives $\mathbb{E}[(f(\hat{z}) - f^*) + [g(\hat{z})]_+] \le 2\varepsilon$, and $2\varepsilon < MR_x/32$ forces $\sqrt{d/T} \le 64\varepsilon/(MR_x)$.

E.3 Proof of Theorem 9.3

The class. Since $a \le 1/2$, both feasible sets below are nonempty subsets of $[-1, 1]$, and $g_i(-1) = -1 - \beta_i \le -1 + a \le -1/2$ for all i, which is the Slater condition at $x^{\mathrm{s}} = -1$ with

$\rho = 1/2$. The samples $G_i(\cdot, \xi)$ are 1-Lipschitz and $\|\xi - \beta\|_\infty \le \sigma + a \le 2\sigma$, so $\sigma_g = 2\sigma$.

Reduction to a test. Under $\beta^{(0)}$ the feasible set is $\{x \le a\}$, $x^* = a$ and $f^* = -a$; under $\beta^{(j)}$ it is $\{x \le -a\}$, $x^* = -a$ and $f^* = a$. Let P_0 and P_j be the transcript laws, and suppose the algorithm satisfies the guarantee. Under P_0: $f(\hat{x}) - f^* = a - \hat{x}$ and $\mathsf{v}(\hat{x}) \ge [\hat{x} - a]_+$; since $a - \hat{x} \ge a\mathbb{1}_{\{\hat{x}<0\}} - [\hat{x} - a]_+$ pointwise, $a\,P_0(\hat{x} < 0) \le \mathbb{E}_0[(a - \hat{x}) + [\hat{x} - a]_+] \le 2\varepsilon$, i.e. $P_0(\hat{x} < 0) \le 2\varepsilon/a = 1/4$. Under P_j: $\mathsf{v}(\hat{x}) \ge [\hat{x} + a]_+ \ge a\mathbb{1}_{\{\hat{x}\ge 0\}}$, so $P_j(\hat{x} \ge 0) \le \varepsilon/a = 1/8$. Hence, for the mixture $\bar{P} := \frac{1}{m}\sum_{j=1}^m P_j$,

$$\mathrm{TV}(P_0, \bar{P}) \ge P_0(\hat{x} \ge 0) - \bar{P}(\hat{x} \ge 0) \ge \frac{3}{4} - \frac{1}{8} = \frac{5}{8}, \qquad \text{so} \qquad \chi^2(\bar{P}\|P_0) \ge 4\,\mathrm{TV}(P_0, \bar{P})^2 \ge \frac{25}{16}$$

by Theorem A.4.

The χ^2-divergence of the mixture. A vector call at x returns $x - \xi$ for a fresh ξ, which is equivalent to observing $\xi \in \{\pm\sigma\}^m$ itself; a paired call reveals the same ξ twice. The law of ξ does not depend on the query point, so (Theorem A.5) the likelihood ratio of the transcript is the product of the likelihood ratios of the T observed vectors $\xi^{(1)}, \dots, \xi^{(T)}$, and under $\beta^{(j)}$ only the j-th coordinates have a different law: $dP_j/dP_0 = \prod_{t=1}^T r(\xi_j^{(t)})$ with $r := d\nu_-/d\nu_+$, where $\nu_\pm$ is the law of a $\pm\sigma$-valued variable with mean $\pm a$. Hence, using the independence of the coordinates under P_0 and $\mathbb{E}_{\nu_+} r = 1$,

$$\begin{aligned}\chi^2(\bar{P}\|P_0) = \mathbb{E}_0\Big[\Big(\frac{1}{m}\sum_j \prod_t r(\xi_j^{(t)})\Big)^2\Big] - 1 &= \frac{1}{m^2}\sum_{j,l}\prod_{t=1}^T \mathbb{E}_0\big[r(\xi_j^{(t)}) r(\xi_l^{(t)})\big] - 1 \\ &= \frac{m(m-1) + m(1+\eta)^T}{m^2} - 1 = \frac{(1+\eta)^T - 1}{m},\end{aligned}$$

where $\eta := \mathbb{E}_{\nu_+}[r^2] - 1 = \chi^2(\nu_-\|\nu_+)$. With $p := \nu_+(\{\sigma\}) = (1 + a/\sigma)/2$ we have $\nu_-(\{\sigma\}) = 1 - p$ and $\eta = (2p-1)^2\big(\frac{1}{p} + \frac{1}{1-p}\big) = \frac{(a/\sigma)^2}{p(1-p)} = \frac{4a^2}{\sigma^2 - a^2} \le \frac{16a^2}{3\sigma^2}$ for $a \le \sigma/2$, which holds because $a = 8\varepsilon \le \sigma/2$.

Conclusion. Combining, $(1+\eta)^T \ge 1 + \frac{25}{16}m \ge 1 + m$, so $T \ge \log(1+m)/\log(1+\eta) \ge \log(1+m)/\eta \ge \frac{3\sigma^2\log(1+m)}{16a^2} = \frac{3\sigma^2\log(1+m)}{1024\,\varepsilon^2}$. The second form uses $\sigma_g = 2\sigma$.

E.4 Proof of Remark 9.5

Only the reduction to a test changes. Let $f(x) = -x + \frac{\mu}{2}x^2$ with $\mu \in (0, 1]$ and $a = 16\varepsilon \le 1/2$; then $f'(x) = -1 + \mu x \le 0$ on $[-1, 1]$, so f is decreasing, f is $(1+\mu)$-Lipschitz on $[-1, 1]$, and the minimizers are again $x^* = a$ under $\beta^{(0)}$ and $x^* = -a$ under $\beta^{(j)}$. Under $\beta^{(0)}$: if $\hat{x} < 0$ then $f(\hat{x}) - f^* \ge f(0) - f(a) = a - \frac{\mu}{2}a^2 \ge \frac{a}{2}$ because $\mu a \le 1$; if $\hat{x} > a$ then $f(\hat{x}) - f(a) = -(\hat{x} - a) + \frac{\mu}{2}(\hat{x}^2 - a^2) \ge -[\hat{x} - a]_+$; hence $f(\hat{x}) - f^* \ge \frac{a}{2}\mathbb{1}_{\{\hat{x}<0\}} - [\hat{x} - a]_+$ pointwise and, as $\mathsf{v}(\hat{x}) \ge [\hat{x} - a]_+$, $\frac{a}{2}P_0(\hat{x} < 0) \le \mathbb{E}_0[(f(\hat{x}) - f^*) + \mathsf{v}(\hat{x})] \le 2\varepsilon$, i.e. $P_0(\hat{x} < 0) \le 4\varepsilon/a = 1/4$. Under $\beta^{(j)}$, $\mathsf{v}(\hat{x}) \ge [\hat{x} + a]_+ \ge a\mathbb{1}_{\{\hat{x}\ge 0\}}$ gives $P_j(\hat{x} \ge 0) \le \varepsilon/a = 1/16$. Hence $\mathrm{TV}(P_0, \bar{P}) \ge \frac{3}{4} - \frac{1}{16} \ge \frac{5}{8}$ and $\mathrm{TV}(P_0, P_j) \ge \frac{5}{8}$, exactly the bounds used above and in the proof of Theorem 9.4; the condition $a \le \sigma/2$ holds because $\varepsilon \le \sigma/32$. The remaining steps are unchanged and give $T \ge 3\sigma^2\log(1+m)/(16a^2) = 3\sigma^2\log(1+m)/(4096\,\varepsilon^2)$ for the vector oracle and $\mathbb{E}_0[T] \ge 25m\sigma^2/(128a^2) = 25m\sigma^2/(32768\,\varepsilon^2)$ for the scalar oracle.

E.5 Proof of Theorem 9.4

With a scalar oracle, the t-th observation is $x_t - \xi_{i_t}^{(t)}$ for an index i_t chosen by the algorithm, equivalently $\xi_{i_t}^{(t)}$. Let $T_j := \#\{t \le T : i_t = j\}$. Under P_0 every observation has law ν_+; under P_j the observations with $i_t = j$ have law ν_- and the others ν_+. Conditionally on the past, the divergence of the t-th observation is $\mathrm{KL}(\nu_+\|\nu_-)\mathbb{1}_{\{i_t=j\}}$, so the chain rule (Theorem A.5, which

applies to a random number of calls by padding with uninformative calls) gives $\mathrm{KL}(P_0\|P_j) = \mathbb{E}_0[T_j]\,\mathrm{KL}(\nu_+\|\nu_-)$. Now

$$\mathrm{KL}(\nu_+\|\nu_-) = p\log\frac{p}{1-p} + (1-p)\log\frac{1-p}{p} = (2p-1)\log\frac{p}{1-p}$$
$$= \frac{a}{\sigma}\log\frac{1+a/\sigma}{1-a/\sigma} \le \frac{a}{\sigma}\cdot\frac{2a/\sigma}{1-a/\sigma} \le \frac{4a^2}{\sigma^2},$$

using $\log\frac{1+x}{1-x} \le x + \frac{x}{1-x} \le \frac{2x}{1-x}$ for $x \in [0,1)$ and $a/\sigma \le \frac{1}{2}$. On the other hand, as in the previous proof, $\mathrm{TV}(P_0,P_j) \ge P_0(\hat{x}\ge 0) - P_j(\hat{x}\ge 0) \ge \frac{5}{8}$, so by Pinsker's inequality $\mathrm{KL}(P_0\|P_j) \ge 2\,\mathrm{TV}^2 \ge \frac{25}{32}$. Hence $\mathbb{E}_0[T_j] \ge \frac{25}{32}\cdot\frac{\sigma^2}{4a^2} = \frac{25\sigma^2}{128a^2} = \frac{25\sigma^2}{8192\,\varepsilon^2}$ for every j, and summing over j gives $\mathbb{E}_0[T] = \sum_j \mathbb{E}_0[T_j] \ge \frac{25m\sigma^2}{8192\,\varepsilon^2}$.

F Experimental details

Instances. The random data $(A_i,\, q_i)$ are generated once from a fixed seed and stored with the results (`data/instance_*.npz`). The constraint rows A_i $(i = 2,\dots,m-1)$ are standard Gaussian vectors with the first coordinate set to zero and normalized to unit length, so that $g_i(x^*) = -0.3 < 0$; the noise directions $q_0,\dots,q_m$ are independent random unit vectors. With $Q = \mathbb{B}$ and $x_0 = a$, the primal radius is $\mathcal{A} = \frac{1}{2}\|a - x^*\|^2 = \frac{1}{2}(0.55)^2$ for both instances and the dual radius is $\mathcal{B} = R^2\log(m+1) = 4\log 9$.

Oracle. Each paired query draws one direction $u \sim U(\mathbb{S})$ (a normalized Gaussian vector) and one sample $\xi = (s_0,\dots,s_m,o_1,\dots,o_m)$ of independent Rademacher variables, evaluates the noisy values at $c \pm \gamma u$ with the *same* ξ, and returns the two vectors; the objective and the constraint estimates of (9) are then formed by the algorithm. Level estimates use one point $c + \gamma v$ with $v \sim U(\mathbb{B})$ (a normalized Gaussian scaled by $U^{1/d}$, U uniform on $[0,1]$) and a fresh ξ. The implementation guards against any reuse of a mini-batch: every batch receives an identifier when it is drawn and raises an error if it is consumed twice. The oracle counts every point evaluation, i.e. the number N_{orc} of vector calls of Theorem 2.8; the identities $N_{\mathrm{orc}} = (2b_p + 3b_y)K_N + 2b_p(N-1)$ of (29) for ZO-Sliding (with $b_p = b_y = 8$ in all runs) and $N_{\mathrm{orc}} = 6bN$ for B-SMP are checked at the end of every run, as is the schedule condition $a_t M_g \varrho \le L$ in every phase, and the columns N_f of Table 6 are $2b_pK_N$ and $4bN$.

Local constants of the quadratic instance. The quadratic objective $\frac{1}{2}\|x-a\|^2$ and the constraint g_m are convex and 1-smooth on $\mathbb{R}^d$ (Theorem 2.4), the objective is 1-strongly convex (Theorem 2.6), and both are Lipschitz on every bounded set; Theorem 2.2 only requires the Lipschitz property on $Q_{\bar{\gamma}}$ (radius $1+\gamma$), where the objective is $(1.75+\gamma)$-Lipschitz and g_m is $(1+\gamma)$-Lipschitz, and these are the constants used below (Table 7). No global Lipschitz property is assumed and none is needed: by Theorem 2.7(i), a global Lipschitz bound would be incompatible with the strong convexity of the objective, and the theory uses only the local constants. In the smooth regime the smoothed Lagrangian is L-smooth on $\mathbb{R}^d$ with $L = L_0 + RL_g = 3$ directly from Theorem 2.4, which is the property required by Theorem 6.3.

Smoothness constant of the cusp instance. For the cusp instance the value $L = 2 + \sqrt{d}/(4\gamma)$ of Table 7 uses the additive structure of the data rather than the generic bound $H \le \sqrt{d}M_R/\gamma \approx 240$: the linear part $-x_1$ and the affine constraints are unchanged by smoothing and contribute nothing to H; the nonsmooth part $\frac{1}{4}\|x_{2:d}\|$ is $\frac{1}{4}$-Lipschitz, so by Theorem 3.1(c) its smoothed version is $\sqrt{d}/(4\gamma)$-smooth; and g_m is 1-smooth, contributing $RL_g = 2$. Hence $H \le 2 + \sqrt{d}/(4\gamma)$, which is all that Theorem 6.5 requires $(L \ge H)$. The slope noise is linear in x and does not affect the smoothness of the expected functions.

Parameters. Table 7 lists the parameters. For ZO-Sliding the constants entering the theoretical stabilizer are the upper bounds implied by the instance: $M_g = 1+\gamma+0.02$, $M_0 = 1.75+\gamma+0.02$ (smooth) or $M_0 = \sqrt{1+1/16}+0.02$ (cusp), $\sigma_h^2 = (2\gamma(1+\gamma)+0.02(1+\gamma)+0.02)^2$ (a deterministic bound on $\left\|\hat{h}-h\right\|_\infty$), $\nu_m = 4d$, $\varrho = 1$, $D = 2$; the stabilizer of Theorem 6.6 is then $\lambda = \Sigma_b\sqrt{A_N/2}/D_A = \Sigma_A\sqrt{A_N/(2b)}/D_A$ (with $b_p = b_y = b = 8$) and $D_A^2 = 3\mathcal{A}+\mathcal{B}$. The inner schedule is exactly the one of (28), $S_t = \max\{1, \lceil M_g\varrho t/L\rceil\}$ with $\varrho = 1$ and the same $M_g = 1+\gamma+0.02$ that enters the stabilizer (a bound that covers the affine constraints, g_m and the slope noise); it gives $K_N = 678$ (smooth) and $K_N = 147$ (cusp), and the condition $a_t M_g \varrho \le L$ of Theorem 6.1(i) is asserted by the code in every phase. For B-SMP the constant L_α is computed from (24) with $H = L$ and the mean-square Lipschitz constant $M_g = 1+\gamma$ of the smoothed constraints.

Table 7: Parameters of the experiments.

	smooth instance	cus
γ, shift of g_m	0.04, $\gamma^2 \cdot 10/52$	0.0
L (upper bound of H)	$L_0 + RL_g = 3$	2 +
R	2	2
ZO-Sliding: N, $b_p = b_y$, ϱ, S_t	60, 8, 1, $\lceil 1.06\,t/3\rceil$	60,
ZO-Sliding: λ (practical / theoretical)	0.1 / 865.1	0.1
ZO-Sliding: K_N, N_{orc}, N_f	678, 28,064, 10,848	147
B-SMP: α, η	$\sqrt{20}$, $0.45/L_\alpha$	$\sqrt{2}$
B-SMP: iterations for $b = 1$ / $b = 8$	4,677 / 584	1,1
restarts (Section 11.3)	7 stages of 10 rounds, $b_k = 2^{k-1}$, $\varrho_k = 2^{(k-1)/2}$; plain runs $N = 60$, $b \in \{8, 16, 23\}$	—
seeds	$10000, \ldots, 10029$	100

Level-testing diagnostic. In Figure 6b the null hypothesis is $\beta^{(0)}$ and the alternative is the uniform mixture of $\beta^{(1)}, \ldots, \beta^{(m)}$ of the proof of Theorem 9.3, with $\sigma = 0.2$, $a = 0.04$ and T vector calls, each returning the m independent $\pm\sigma$-valued coordinates. The test statistic is the exact log-likelihood ratio of the mixture against the null, $\log \frac{1}{m}\sum_{j=1}^m r_j^T$ with $r_j^T = \exp\big((2c_j - T)\log\frac{\sigma-a}{\sigma+a}\big)$ and c_j the number of $+\sigma$ observations in coordinate j, thresholded at 0; the reported error is the average of the two error probabilities (the Bayes risk with equal priors), estimated from 4 000 independent trials under each hypothesis, so that its standard error is at most 0.01.

Metrics. After every round (resp. every Mirror–Prox iteration) the current output $\bar{x}$ is evaluated exactly (the exact f and g are used only for reporting, never by the algorithms): the objective gap $f(\bar{x}) - f^*$, the violation $\mathsf{v}(\bar{x}) = \max_i [g_i(\bar{x})]_+$ and the joint error $J(\bar{x}) = \max\{0, f(\bar{x}) - f^*, \mathsf{v}(\bar{x})\}$. The means in Table 6 are over the 30 seeds; the confidence intervals are percentile bootstrap intervals (10 000 resamples) for the mean of the final J. The bands in the figures are pointwise interquartile ranges over seeds.

Reproducibility. The archive contains the code in the directory `code/`: the file `zo_constrained.py` (problem, oracle and the two methods), `run_all.py` (all experiments) and `make_figures.py` (figures and table); the raw records in `data/`: `trajectories.csv`, `restarts.csv`, `level_testing.csv` and `noise_diagnostic.csv`; the configurations `config.json`, `restart_config.json` and `environment.json`; and the figures.

Running `python3 run_all.py` followed by `python3 make_figures.py` in the `code` directory regenerates all data, figures, the table `sections/results_table.tex` and the numerical macros `sections/results_macros.tex` used in the text. The total running time is about five minutes on a single CPU core (Python 3 with NumPy, SciPy, pandas and Matplotlib).