Variance Reduction for Matrix Games

Yair Carmon, Yujia Jin, Aaron Sidford and Kevin Tian
Stanford University
{yairc,yujiajin,sidford,kjtian}@stanford.edu

Abstract

We present a randomized primal-dual algorithm that solves the problem $\min_x \max_y y^\top Ax$ to additive error ϵ in time $\mathrm{nnz}(A) + \sqrt{\mathrm{nnz}(A)n}/\epsilon$, for matrix A with larger dimension n and $\mathrm{nnz}(A)$ nonzero entries. This improves the best known exact gradient methods by a factor of $\sqrt{\mathrm{nnz}(A)/n}$ and is faster than fully stochastic gradient methods in the accurate and/or sparse regime $\epsilon \leq \sqrt{n/\mathrm{nnz}(A)}$. Our results hold for x, y in the simplex (matrix games, linear programming) and for x in an ℓ_2 ball and y in the simplex (perceptron / SVM, minimum enclosing ball). Our algorithm combines the Nemirovski's "conceptual prox-method" and a novel reduced-variance gradient estimator based on "sampling from the difference" between the current iterate and a reference point.

1 Introduction

Minimax problems—or games—of the form $\min_x \max_y f(x, y)$ are ubiquitous in economics, statistics, optimization and machine learning. In recent years, minimax formulations for neural network training rose to prominence [15, 23], leading to intense interest in algorithms for solving large scale minimax games [10, 14, 20, 9, 18, 24]. However, the algorithmic toolbox for minimax optimization is not as complete as the one for minimization. Variance reduction, a technique for improving stochastic gradient estimators by introducing control variates, stands as a case in point. A multitude of variance reduction schemes exist for finite-sum minimization [cf. 19, 34, 1, 4, 12], and their impact on complexity is well-understood [43]. In contrast, only a few works apply variance reduction to finite-sum minimax problems [3, 39, 5, 26], and the potential gains from variance reduction are not well-understood.

We take a step towards closing this gap by designing variance-reduced minimax game solvers that offer strict runtime improvements over non-stochastic gradient methods, similar to that of optimal variance reduction methods for finite-sum minimization. To achieve this, we focus on the fundamental class of bilinear minimax games,

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} y^\top Ax, \text{ where } A \in \mathbb{R}^{m \times n}.$$

In particular, we study the complexity of finding an ϵ-approximate saddle point (Nash equilibrium), namely x, y with

$$\max_{y' \in \mathcal{Y}} (y')^\top Ax - \min_{x' \in \mathcal{X}} y^\top Ax' \leq \epsilon.$$

In the setting where $\mathcal{X}$ and $\mathcal{Y}$ are both probability simplices, the problem corresponds to finding an approximate (mixed) equilibrium in a matrix game, a central object in game theory and economics. Matrix games are also fundamental to algorithm design due in part to their equivalence to linear programming [8]. Alternatively, when $\mathcal{X}$ is an ℓ_2 ball and $\mathcal{Y}$ is a simplex, solving the corresponding problem finds a maximum-margin linear classifier (hard-margin SVM), a fundamental task in machine learning and statistics [25]. We refer to the former as an ℓ_1-ℓ_1 game and the latter as an ℓ_2-ℓ_1 game; our primary focus is to give improved algorithms for these domains.

1.1 Our Approach

Our starting point is Nemirovski's "conceptual prox-method" [28] for solving $\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y)$, where $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is convex in x and concave in y. The method solves a sequence of subproblems parameterized by $\alpha > 0$, each of the form

$$\text{find } x, y \text{ s.t. } \forall x', y' \ \langle \nabla_x f(x, y), x - x' \rangle - \langle \nabla_y f(x, y), y - y' \rangle \leq \alpha V_{x_0}(x') + \alpha V_{y_0}(y') \quad (1)$$

for some $(x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$, where $V_a(b)$ is a norm-suitable Bregman divergence from a to b: squared Euclidean distance for ℓ_2 and KL divergence for ℓ_1. Combining each subproblem solution with an extragradient step, the prox-method solves the original problem to ϵ accuracy by solving $\widetilde{O}(\alpha/\epsilon)$ subproblems.[1] (Solving (1) with $\alpha = 0$ is equivalent to to solving $\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y)$.)

Our first contribution is showing that if a stochastic unbiased gradient estimator $\tilde{g}$ satisfies the "variance" bound

$$\mathbb{E} \|\tilde{g}(x, y) - \nabla f(x_0, y_0)\|_*^2 \leq L^2 \|x - x_0\|^2 + L^2 \|y - y_0\|^2 \quad (2)$$

for some $L > 0$, then $O(L^2/\alpha^2)$ regularized stochastic mirror descent steps using $\tilde{g}$ solve (1) in a suitable probabilistic sense. We call unbiased gradient estimators that satisfy (2) "centered."

Our second contribution is the construction of "centered" gradient estimators for ℓ_1-ℓ_1 and ℓ_2-ℓ_1 bilinear games, where $f(x, y) = y^\top A x$. Our ℓ_1 estimator has the following form. Suppose we wish to estimate $g^{\times} = A^\top y$ (the gradient of f w.r.t. x), and we already have $g_0^{\times} = A^\top y_0$. Let $p \in \Delta^m$ be some distribution over $\{1, \ldots, m\}$, draw $i \sim p$ and set

$$\tilde{g}^{\times} = g_0^{\times} + A_{i:} \frac{[y]_i - [y_0]_i}{p_i},$$

where $A_{i:}$ is the ith column of $A^\top$. This form is familiar from variance reduction techniques [19, 44, 1], that typically use a fixed distribution p. In our setting, however, a fixed p will not produce sufficiently low variance. Departing from prior variance-reduction work and building on [16, 6], we choose p *based on* y according to

$$p_i(y) = \frac{|[y]_i - [y_0]_i|}{\|y - y_0\|_1},$$

yielding exactly the variance bound we require. We call this technique "sampling from the difference."

For our ℓ_2 gradient estimator, we sample from the *squared* difference, drawing $\mathcal{X}$-block coordinate $j \sim q$, where $q_j(x) = ([x]_j - [x_0]_j)^2 / \|x - x_0\|_2^2$. To strengthen our results for ℓ_2-ℓ_1 games, we consider a refined version of the "centered" criterion (2) which allows regret analysis using local norms [37, 6]. To further facilitate this analysis we follow [6] and introduce gradient clipping. We extend our proofs to show that stochastic regularized mirror descent can solve (1) despite the (distance-bounded) bias caused by gradient clipping.

Our gradient estimators attain the bound (2) with L equal to the Lipschitz constant of ∇f. Specifically,

$$L = \begin{cases} \max_{ij} |A_{ij}| & \text{in the } \ell_1\text{-}\ell_1 \text{ setup} \\ \max_i \|A_{i:}\|_2 & \text{in the } \ell_2\text{-}\ell_1 \text{ setup.} \end{cases} \quad (3)$$

1.2 Method complexity compared with prior art

As per the discussion above, to achieve accuracy ϵ our algorithm solves $\widetilde{O}(\alpha/\epsilon)$ subproblems. Each subproblem takes $O(\text{nnz}(A))$ time for computing two exact gradients (one for variance reduction and one for an extragradient step), plus an additional $(m + n)L^2/\alpha^2$ time for the inner mirror descent iterations, with L as in (3). The total runtime is therefore

$$\widetilde{O} \left(\left(\text{nnz}(A) + \frac{(m + n)L^2}{\alpha^2} \right) \frac{\alpha}{\epsilon} \right).$$

---

[1] More precisely, the required number of subproblem solutions is at most $\Theta \cdot \frac{\alpha}{\epsilon}$, where Θ is a "domain size" parameter that depends on $\mathcal{X}$, $\mathcal{Y}$, and the Bregman divergence V (see Section 2). In the ℓ_1 and ℓ_2 settings considered in this paper, we have the bound $\Theta \leq \log(nm)$ and we use the $\widetilde{O}$ notation to suppress terms logarithmic in n and m. However, in other settings—e.g., ℓ_∞-ℓ_1 games [cf. 38, 40]—making the parameter Θ scale logarithmically with the problem dimension is far more difficult.

By setting α optimally to be $\max\{\epsilon, L\sqrt{(m+n)/\text{nnz}(A)}\}$, we obtain the runtime

$$\widetilde{O}(\text{nnz}(A) + \sqrt{\text{nnz}(A) \cdot (m+n)} \cdot L \cdot \epsilon^{-1}). \tag{4}$$

Comparison with mirror-prox and dual extrapolation. Nemirovski [28] instantiates his conceptual prox-method by solving the relaxed proximal problem (1) with $\alpha = L$ in time $O(\text{nnz}(A))$, where L is the Lipschitz constant of ∇f, as given in (3). The total complexity of the resulting method is therefore

$$\widetilde{O}(\text{nnz}(A) \cdot L \cdot \epsilon^{-1}). \tag{5}$$

The closely related dual extrapolation method of Nesterov [31] attains the same rate of convergence. We refer to the running time (5) as *linear* since it scales linearly with the problem description size $\text{nnz}(A)$. Our running time guarantee (4) is never worse than (5) by more than a constant factor, and improves on (5) when $\text{nnz}(A) = \omega(n+m)$, i.e. whenever A is not extremely sparse. In that regime, our method uses $\alpha \ll L$, hence solving a harder version of (1) than possible for mirror-prox.

Comparison with sublinear-time methods Using a randomized algorithm, Grigoriadis and Khachiyan [16] solve ℓ_1-ℓ_1 bilinear games in time

$$\widetilde{O}((m+n) \cdot L^2 \cdot \epsilon^{-2}), \tag{6}$$

and Clarkson et al. [6] extend this result to ℓ_2-ℓ_1 bilinear games, with the values of L as in (3). Since these runtimes scale with $n+m \le \text{nnz}(A)$, we refer to them as *sublinear*. Our guarantee improves on the guarantee (6) when $(m+n) \cdot L^2 \cdot \epsilon^{-2} \gg \text{nnz}(A)$, i.e. whenever (6) is not truly sublinear.

Our method carefully balances linear-time extragradient steps with cheap sublinear-time stochastic gradient steps. Consequently, our runtime guarantee (4) inherits strengths from both the linear and sublinear runtimes. First, our runtime scales linearly with L/ϵ rather than quadratically, as does the linear runtime (5). Second, while our runtime is not strictly sublinear, its component proportional to L/ϵ is $\sqrt{\text{nnz}(A)(n+m)}$, which is sublinear in $\text{nnz}(A)$.

Overall, our method offers the best runtime guarantee in the literature in the regime

$$\frac{\sqrt{\text{nnz}(A)(n+m)}}{\min\{n,m\}^\omega} \ll \frac{\epsilon}{L} \ll \sqrt{\frac{n+m}{\text{nnz}(A)}},$$

where the lower bound on ϵ is due to the best known theoretical runtimes of interior point methods: $\widetilde{O}(\max\{n,m\}^\omega \log(L/\epsilon))$ [7] and $\widetilde{O}(\text{nnz}(A) + \min\{n,m\}^2)\sqrt{\min\{n,m\}} \log(L/\epsilon))$ [21], where ω is the (current) matrix multiplication exponent.

In the square dense case (i.e. $\text{nnz}(A) \approx n^2 = m^2$), we improve on the accelerated runtime (5) by a factor of $\sqrt{n}$, the same improvement that optimal variance-reduced finite-sum minimization methods achieve over the fast gradient method [44, 1].

1.3 Related work

Matrix games, the canonical form of discrete zero-sum games, have long been studied in economics [32]. The classical mirror descent (i.e. no-regret) method yields an algorithm with running time $\widetilde{O}(\text{nnz}(A)L^2\epsilon^{-2})$ [30]. Subsequent work [16, 28, 31, 6] improve this runtime as described above. Our work builds on the extragradient scheme of Nemirovski [28] as well as the gradient estimation and clipping technique of Clarkson et al. [6].

Balamurugan and Bach [3] apply standard variance reduction [19] to bilinear ℓ_2-ℓ_2 games by sampling elements proportional to squared matrix entries. Using proximal-point acceleration they obtain a runtime of $\widetilde{O}(\text{nnz}(A) + \|A\|_F\sqrt{\text{nnz}(A) \max\{m,n\}}\epsilon^{-1} \log \frac{1}{\epsilon})$, a rate we recover using our algorithm (Appendix E). However, in this setting the mirror-prox method has runtime $\widetilde{O}(\|A\|_{\text{op}} \text{nnz}(A)\epsilon^{-1})$, which may be better than the result of [3] by a factor of $\sqrt{mn/\text{nnz}(A)}$ due to the discrepancy in the norm of A. Naive application of [3] to ℓ_1 domains results in even greater potential losses. Shi et al. [39] extend the method of [3] to smooth functions using general Bregman divergences, but their extension is unaccelerated and appears limited to a ϵ^{-2} rate.

Chavdarova et al. [5] propose a variance-reduced extragradient method with applications to generative adversarial training. In contrast to our algorithm, which performs extragadient steps in the outer loop,

the method of [5] performs stochastic extragradient steps in the inner loop, using finite-sum variance reduction as in [19]. Chavdarova et al. [5] analyze their method in the convex-concave setting, showing improved stability over direct application of the extragradient method to noisy gradients. However, their complexity guarantees are worse than those of linear-time methods. Following up on [5], Mishchenko et al. [26] propose to reduce the variance of the stochastic extragradient method by using the same stochastic sample for both the gradient and extragradient steps. In the Euclidean strongly convex case, they show a convergence guarantee with a relaxed variance assumption, and in the noiseless full-rank bilinear case they recover the guarantees of [27]. In the general convex case, however, they only show an ϵ^{-2} rate of convergence.

1.4 Paper outline and additional contributions

We define our notation in Section 2. In Section 3.1, we review Nemirovski's conceptual prox-method and introduce the notion of a relaxed proximal oracle; we implement such oracle using variance-reduced gradient estimators in Section 3.2. In Section 4, we construct these gradient estimators for the ℓ_1-ℓ_1 and ℓ_2-ℓ_1 domain settings, and complete the analyses of the corresponding algorithms; in Appendix E we provide analogous treatment for the ℓ_2-ℓ_2 setting, recovering the results of [3].

In Appendix F we provide three additional contributions: variance-reduction-based computation of proximal points for arbitrary convex-concave functions (Appendix F.1); extension of our results to "composite" saddle point problems of the form $\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \{f(x, y) + \phi(x) - \psi(y)\}$, where f admits a centered gradient estimator and ϕ, ψ are "simple" convex functions (Appendix F.2); and a number of alternative centered gradient estimators for the ℓ_2-ℓ_1 and ℓ_2-ℓ_2 settings (Appendix F.3).

2 Notation

Problem setup. A *setup* is the triplet $(\mathcal{Z}, \|\cdot\|, r)$ where: (i) $\mathcal{Z}$ is a compact and convex subset of $\mathbb{R}^n \times \mathbb{R}^m$, (ii) $\|\cdot\|$ is a norm on $\mathcal{Z}$ and (iii) r is 1-strongly-convex w.r.t. $\mathcal{Z}$ and $\|\cdot\|$, i.e. such that $r(z') \geq r(z) + \langle \nabla r(z), z - z' \rangle + \frac{1}{2} \|z' - z\|^2$ for all $z, z' \in \mathcal{Z}$.[2] We call r the *distance generating function* and denote the *Bregman divergence* associated with it by

$$V_z(z') := r(z') - r(z) - \langle \nabla r(z), z' - z \rangle \geq \frac{1}{2} \|z' - z\|^2.$$

We also denote $\Theta := \max_{z'} r(z') - \min_z r(z)$ and assume it is finite.

Norms and dual norms. We write $\mathcal{S}^*$ for the set of linear functions on $\mathcal{S}$. For $\zeta \in \mathcal{Z}^*$ we define the dual norm of $\|\cdot\|$ as $\|\zeta\|_* := \max_{\|z\| \leq 1} \langle \zeta, z \rangle$. For $p \geq 1$ we write the ℓ_p norm $\|z\|_p = (\sum_i z_i^p)^{1/p}$ with $\|z\|_\infty = \max_i |z_i|$. The dual norm of ℓ_p is ℓ_q with $q^{-1} = 1 - p^{-1}$.

Domain components. We assume $\mathcal{Z}$ is of the form $\mathcal{X} \times \mathcal{Y}$ for convex and compact sets $\mathcal{X} \subset \mathbb{R}^n$ and $\mathcal{Y} \subset \mathbb{R}^m$. Particular sets of interest are the simplex $\Delta^d = \{v \in \mathbb{R}^d \mid \|v\|_1 = 1, v \geq 0\}$ and the Euclidean ball $\mathbb{B}^d = \{v \in \mathbb{R}^d \mid \|v\|_2 \leq 1\}$. For any vector in $z \in \mathbb{R}^n \times \mathbb{R}^m$,

we write z^x and z^y for the first n and last m coordinates of z, respectively.

When totally clear from context, we sometimes refer to the $\mathcal{X}$ and $\mathcal{Y}$ components of z directly as x and y. We write the ith coordinate of vector v as $[v]_i$.

Matrices. We consider a matrix $A \in \mathbb{R}^{m \times n}$ and write $\mathrm{nnz}(A)$ for the number of its nonzero entries. For $i \in [n]$ and $j \in [m]$ we write $A_{i:}$, $A_{:j}$ and A_{ij} for the corresponding row, column and entry, respectively.[3] We consider the matrix norms $\|A\|_{\max} := \max_{ij} |A_{ij}|$, $\|A\|_{p \to q} := \max_{\|x\|_p \leq 1} \|Ax\|_q$ and $\|A\|_F := (\sum_{i,j} A_{ij}^2)^{1/2}$.

---

[2] For non-differentiable r, let $\langle \nabla r(z), w \rangle := \sup_{\gamma \in \partial r(z)} \langle \gamma, w \rangle$, where $\partial r(z)$ is the subdifferential of r at z.
[3] For $k \in \mathbb{N}$, we let $[k] := \{1, \ldots, k\}$.

3 Primal-dual variance reduction framework

In this section, we establish a framework for solving the saddle point problem

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y),$$

where f is convex in x and concave y, and admits a (variance-reduced) stochastic estimator for the continuous and monotone[4] gradient mapping

$$g(z) = g(x, y) := (\nabla_x f(x, y), -\nabla_y f(x, y)) .$$

Our goal is to find an ϵ-approximate saddle point (Nash equilibrium), i.e. $z \in \mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ such that

$$\mathrm{Gap}(z) := \max_{y' \in \mathcal{Y}} f(z^{\mathsf{x}}, y') - \min_{x' \in \mathcal{X}} f(x', z^{\mathsf{y}}) \le \epsilon. \tag{7}$$

We achieve this by generating a sequence $z_1, z_2, \ldots, z_k$ such that $\frac{1}{K} \sum_{k=1}^{K} \langle g(z_k), z_k - u \rangle \le \epsilon$ for every $u \in \mathcal{Z}$ and using the fact that

$$\mathrm{Gap}\left(\frac{1}{K} \sum_{k=1}^{K} z_k\right) \le \max_{u \in \mathcal{Z}} \frac{1}{K} \sum_{k=1}^{K} \langle g(z_k), z_k - u \rangle \tag{8}$$

due to convexity-concavity of f (see proof in Appendix A.1).

In Section 3.1 we define the notion of a (randomized) *relaxed proximal oracle*, and describe how Nemirovski's mirror-prox method leverages it to solve the problem (3). In Section 3.2 we define a class of *centered* gradient estimators, whose variance is proportional to the squared distance from a reference point. Given such a centered gradient estimator, we show that a regularized stochastic mirror descent scheme constitutes a relaxed proximal oracle. For a technical reason, we limit our oracle guarantee in Section 3.2 to the bilinear case $f(x, y) = y^\top A x$, which suffices for the applications in Section 4. We lift this limitation in Appendix F.1, where we show a different oracle implementation that is valid for general convex-concave f, with only a logarithmic increase in complexity.

3.1 The mirror-prox method with a randomized oracle

Recall that we assume the space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ is equipped with a norm $\|\cdot\|$ and distance generating function $r : \mathcal{Z} \to \mathbb{R}$ that is 1-strongly-convex w.r.t. $\|\cdot\|$ and has range Θ. We write the induced Bregman divergence as $V_z(z') = r(z') - r(z) - \langle \nabla r(z), z' - z \rangle$. We use the following fact throughout the paper: by definition, the Bregman divergence satisfies, for any $z, z', u \in \mathcal{Z}$,

$$- \langle \nabla V_z(z'), z' - u \rangle = V_z(u) - V_{z'}(u) - V_z(z'). \tag{9}$$

For any $\alpha > 0$ we define the α-proximal mapping $\mathrm{Prox}_z^\alpha(g)$ to be the solution of the variational inequality corresponding to the strongly monotone operator $g + \alpha \nabla V_z$, i.e. the unique $z_\alpha \in \mathcal{Z}$ such that $\langle g(z_\alpha) + \alpha \nabla V_z(z_\alpha), z_\alpha - u \rangle \le 0$ for all $u \in \mathcal{Z}$ [cf. 11]. Equivalently (by (9)),

$$\mathrm{Prox}_z^\alpha(g) := \text{the unique } z_\alpha \in \mathcal{Z} \text{ s.t. } \langle g(z_\alpha), z_\alpha - u \rangle \le \alpha V_z(u) - \alpha V_{z_\alpha}(u) - \alpha V_z(z_\alpha) \ \forall u \in \mathcal{Z}. \tag{10}$$

When $V_z(z') = V_x^{\mathsf{x}}(x') + V_y^{\mathsf{y}}(y')$, $\mathrm{Prox}_z^\alpha(g)$ is also the unique solution of the saddle point problem

$$\min_{x' \in \mathcal{X}} \max_{y' \in \mathcal{Y}} \left\{ f(x', y') + \alpha V_x^{\mathsf{x}}(x') - \alpha V_y^{\mathsf{y}}(y') \right\} .$$

Consider iterations of the form $z_k = \mathrm{Prox}_{z_{k-1}}^\alpha(g)$, with $z_0 = \arg\min_z r(z)$. Averaging the definition (10) over k, using the bound (8) and the nonnegativity of Bregman divergences gives

$$\mathrm{Gap}\left(\frac{1}{K} \sum_{k=1}^{K} z_k\right) \le \max_{u \in \mathcal{Z}} \frac{1}{K} \sum_{k=1}^{K} \langle g(z_k), z_k - u \rangle \le \max_{u \in \mathcal{Z}} \frac{\alpha (V_{z_0}(u) - V_{z_K}(u))}{K} \le \frac{\alpha \Theta}{K}.$$

Thus, we can find an ϵ-suboptimal point in $K = \alpha \Theta / \epsilon$ exact proximal steps. However, computing $\mathrm{Prox}_z^\alpha(g)$ exactly may be as difficult as solving the original problem. Nemirovski [28] proposes a relaxation of the exact proximal mapping, which we slightly extend to include the possibility of randomization, and formalize in the following.

---

[4] A mapping $q : \mathcal{Z} \to \mathcal{Z}^*$ is monotone if and only if $\langle q(z') - q(z), z' - z \rangle \ge 0$ for all $z, z' \in \mathcal{Z}$; g is monotone due to convexity-concavity of f.

Definition 1 ((α, ε)-relaxed proximal oracle). *Let g be a monotone operator and $\alpha, \varepsilon > 0$. An (α, ε)-relaxed proximal oracle for g is a (possibly randomized) mapping $\mathcal{O} : \mathcal{Z} \to \mathcal{Z}$ such that $z' = \mathcal{O}(z)$ satisfies*

$$\mathbb{E} \left[\max_{u \in \mathcal{Z}} \left\{ \langle g(z'), z' - u \rangle - \alpha V_z(u) \right\} \right] \leq \varepsilon.$$

Note that $\mathcal{O}(z) = \mathrm{Prox}_z^{\alpha}(g)$ is an ($\alpha, 0$)-relaxed proximal oracle. Algorithm 1 describes the "conceptual prox-method" of Nemirovski [28], which recovers the error guarantee of exact proximal iterations. The kth iteration consists of (i) a relaxed proximal oracle call producing $z_{k-1/2} = \mathcal{O}(z_{k-1})$, and (ii) a *linearized* proximal (mirror) step where we replace $z \mapsto g(z)$ with the constant function $z \mapsto g(z_{k-1/2})$, producing $z_k = \mathrm{Prox}_{z_{k-1}}^{\alpha}(g(z_{k-1/2}))$. We now state the convergence guarantee for the mirror-prox method, first shown in [28] (see Appendix B.1 for a simple proof).

Algorithm 1: `OuterLoop`($\mathcal{O}$) (Nemirovski [28])

Input: (α, ε)-relaxed proximal oracle $\mathcal{O}(z)$ for gradient mapping g, distance-generating r
Parameters: Number of iterations K
Output: Point $\bar{z}_K$ with $\mathbb{E} \mathrm{Gap}(\bar{z}) \leq \frac{\alpha \Theta}{K} + \varepsilon$

1   $z_0 \leftarrow \arg\min_{z \in \mathcal{Z}} r(z)$
2   **for** $k = 1, \ldots, K$ **do**
3     $z_{k-1/2} \leftarrow \mathcal{O}(z_{k-1})$ $\triangleright$ We implement $\mathcal{O}(z_{k-1})$ by calling `InnerLoop`$(z_{k-1}, \tilde{g}_{z_{k-1}}, \alpha)$
4     $z_k \leftarrow \mathrm{Prox}_{z_{k-1}}^{\alpha}(g(z_{k-1/2})) = \arg\min_{z \in \mathcal{Z}} \left\{ \langle g(z_{k-1/2}), z \rangle + \alpha V_{z_{k-1}}(z) \right\}$
5   **return** $\bar{z}_K = \frac{1}{K} \sum_{k=1}^{K} z_{k-1/2}$

Proposition 1 (Mirror prox convergence via oracles). *Let $\mathcal{O}$ be an (α, ε)-relaxed proximal oracle with respect to gradient mapping g and distance-generating function r with range at most Θ. Let $z_{1/2}, z_{3/2}, \ldots, z_{K-1/2}$ be the iterates of Algorithm 1 and let $\bar{z}_K$ be its output. Then*

$$\mathbb{E} \mathrm{Gap}(\bar{z}_K) \leq \mathbb{E} \max_{u \in \mathcal{Z}} \frac{1}{K} \sum_{k=1}^{K} \langle g(z_{k-1/2}), z_{k-1/2} - u \rangle \leq \frac{\alpha \Theta}{K} + \varepsilon.$$

3.2 Implementation of an ($\alpha, 0$)-relaxed proximal oracle

We now explain how to use stochastic variance-reduced gradient estimators to design an efficient ($\alpha, 0$)-relaxed proximal oracle. We begin by introducing the bias and variance properties of the estimators we require.

Definition 2. *Let $z_0 \in \mathcal{Z}$ and $L > 0$. A stochastic gradient estimator $\tilde{g}_{z_0} : \mathcal{Z} \to \mathcal{Z}^*$ is called (z_0, L)-centered for g if for all $z \in \mathcal{Z}$*

> *1. $\mathbb{E} \left[\tilde{g}_{z_0}(z) \right] = g(z)$,*
>
> *2. $\mathbb{E} \left\| \tilde{g}_{z_0}(z) - g(z_0) \right\|_*^2 \leq L^2 \left\| z - z_0 \right\|^2$.*

Lemma 1. *A (z_0, L)-centered estimator for g satisfies $\mathbb{E} \left\| \tilde{g}_{z_0}(z) - g(z) \right\|_*^2 \leq (2L)^2 \left\| z - z_0 \right\|^2$.*

Proof. Writing $\tilde{\delta} = \tilde{g}_{z_0}(z) - g(z_0)$, we have $\mathbb{E}\tilde{\delta} = g(z) - g(z_0)$ by the first centered estimator property. Therefore,

$$\mathbb{E} \left\| \tilde{g}_{z_0}(z) - g(z) \right\|_*^2 = \mathbb{E} \| \tilde{\delta} - \mathbb{E}\tilde{\delta} \|_*^2 \overset{(i)}{\leq} 2\mathbb{E} \| \tilde{\delta} \|_*^2 + 2 \| \mathbb{E}\tilde{\delta} \|_*^2 \overset{(ii)}{\leq} 4\mathbb{E} \| \tilde{\delta} \|_*^2 \overset{(iii)}{\leq} (2L)^2 \left\| z - z_0 \right\|^2,$$

where the bounds follow from (i) the triangle inequality, (ii) Jensen's inequality and (iii) the second centered estimator property. $\qquad \square$

Remark 1. A gradient mapping that admits a (z, L)-centered gradient estimator for every $z \in \mathcal{Z}$ is $2L$-Lipschitz, since by Jensen's inequality and Lemma 1 we have for all $w \in \mathcal{Z}$

$$\| g(w) - g(z) \|_* = \| \mathbb{E}\tilde{g}_z(w) - g(z) \|_* \leq (\mathbb{E} \| \tilde{g}_z(w) - g(z) \|_*^2)^{1/2} \leq 2L \| w - z \|.$$

<div align="center">6</div>

Remark 2. Definition 2 bounds the gradient variance using the distance to the reference point. Similar bounds are used in variance reduction for bilinear saddle-point problems with Euclidean norm [3], as well as for finding stationary points in smooth nonconvex finite-sum problems [2, 33, 12, 45]. However, known variance reduction methods for smooth convex finite-sum minimization require stronger bounds [cf. 1, Section 2.1].

With the variance bounds defined, we describe Algorithm 2 which (for the bilinear case) implements a relaxed proximal oracle. The algorithm is stochastic mirror descent with an additional regularization term around the initial point w_0. Note that we do not perform extragradient steps in this stochastic method. When combined with a centered gradient estimator, the iterates of Algorithm 2 provide the following guarantee, which is one of our key technical contributions.

Algorithm 2: $\texttt{InnerLoop}(w_0, \tilde{g}_{w_0}, \alpha)$

Input: Initial $w_0 \in \mathcal{Z}$, gradient estimator $\tilde{g}_{w_0}$, oracle quality $\alpha > 0$
Parameters: Step size η, number of iterations T
Output: Point $\bar{w}_T$ satisfying Definition 1 (for appropriate $\tilde{g}_{w_0}, \eta, T$)

1 **for** $t = 1, \ldots, T$ **do**
2 $\quad w_t \leftarrow \arg\min_{w \in \mathcal{Z}} \left\{ \langle \tilde{g}_{w_0}(w_{t-1}), w \rangle + \frac{\alpha}{2} V_{w_0}(w) + \frac{1}{\eta} V_{w_{t-1}}(w) \right\}$
3 **return** $\bar{w}_T = \frac{1}{T} \sum_{t=1}^{T} w_t$

Proposition 2. *Let $\alpha, L > 0$, let $w_0 \in \mathcal{Z}$ and let $\tilde{g}_{w_0}$ be (w_0, L)-centered for monotone g. Then, for $\eta = \frac{\alpha}{10L^2}$ and $T \geq \frac{4}{\eta\alpha} = \frac{40L^2}{\alpha^2}$, the iterates of Algorithm 2 satisfy*

$$\mathbb{E} \max_{u \in \mathcal{Z}} \left[\frac{1}{T} \sum_{t \in [T]} \langle g(w_t), w_t - u \rangle - \alpha V_{w_0}(u) \right] \leq 0. \tag{11}$$

Before discussing the proof of Proposition 2, we state how it implies the relaxed proximal oracle property for the bilinear case.

Corollary 1. *Let $A \in \mathbb{R}^{m \times n}$ and let $g(z) = (A^\top z^{\mathsf{y}}, -A z^{\mathsf{x}})$. Then, in the setting of Proposition 2, $\mathcal{O}(w_0) = \texttt{InnerLoop}(w_0, \tilde{g}_{w_0}, \alpha)$ is an $(\alpha, 0)$-relaxed proximal oracle.*

Proof. Note that $\langle g(z), w \rangle = -\langle g(w), z \rangle$ for any $z, w \in \mathcal{Z}$ and consequently $\langle g(z), z \rangle = 0$. Therefore, the iterates $w_1, \ldots, w_T$ of Algorithm 2 and its output $\bar{w}_T = \frac{1}{T} \sum_{t=1}^{T} w_t$ satisfy for every $u \in \mathcal{Z}$,

$$\frac{1}{T} \sum_{t \in [T]} \langle g(w_t), w_t - u \rangle = \frac{1}{T} \sum_{t \in [T]} \langle g(u), w_t \rangle = \langle g(u), \bar{w}_T \rangle = \langle g(\bar{w}_T), \bar{w}_T - u \rangle.$$

Substituting into the bound (11) yields the $(\alpha, 0)$-relaxed proximal oracle property in Definition 1. □

More generally, the proof of Corollary 1 shows that Algorithm 2 implements a relaxed proximal oracle whenever $z \mapsto \langle g(z), z - u \rangle$ is convex for every u. In Appendix F.1 we implement an (α, ε)-relaxed proximal oracle without such an assumption.

The proof of Proposition 2 is a somewhat lengthy application of existing techniques for stochastic mirror descent analysis in conjunction with Definition 2. We give it in full in Appendix B.2 and sketch it briefly here. We view Algorithm 2 as mirror descent with stochastic gradients $\tilde{\delta}_t = \tilde{g}_{w_0}(w_t) - g(w_0)$ and composite term $\langle g(w_0), z \rangle + \frac{\alpha}{2} V_{w_0}(z)$. For any $u \in \mathcal{Z}$, the standard mirror descent analysis (see Lemma 4 in Appendix A.2) bounds the regret $\sum_{t \in [T]} \langle \tilde{g}_{w_0}(w_t) + \frac{\alpha}{2} \nabla V_{w_0}(w_t), w_t - u \rangle$ in terms of the distance to initialization $V_{w_0}(u)$ and the stochastic gradient norms $\|\tilde{\delta}_t\|_*^2$ for $t \in [T]$. Bounding these norms via Definition 2 and rearranging the $\langle \nabla V_{w_0}(w_t), w_t - u \rangle$ terms, we show that $\mathbb{E}\left[\frac{1}{T} \sum_{t \in [T]} \langle g(w_t), w_t - u \rangle - \alpha V_{w_0}(u) \right] \leq 0$ for all $u \in \mathcal{Z}$. To reach our desired result we must swap the order of the expectation and "for all." We do so using the "ghost iterate" technique due to Nemirovski et al. [29].

4 Application to bilinear saddle point problems

We now construct centered gradient estimators (as per Definition 2) for the linear gradient mapping

$$g(z) = (A^\top z^{\mathsf{y}}, -Az^{\mathsf{x}}) \text{ corresponding to the bilinear saddle point problem } \min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} y^\top Ax.$$

Sections 4.1 and 4.2 consider the ℓ_1-ℓ_1 and ℓ_2-ℓ_1 settings, respectively; in Appendix E we show how our approach naturally extends to the ℓ_2-ℓ_2 setting as well. Throughout, we let w_0 denote the "center" (i.e. reference point) of our stochastic gradient estimator and consider a general query point $w \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. We also recall the notation $[v]_i$ for the ith entry of vector v.

4.1 ℓ_1-ℓ_1 games

Setup. Denoting the d-dimensional simplex by Δ^d, we let $\mathcal{X} = \Delta^n$, $\mathcal{Y} = \Delta^m$ and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. We take $\|\cdot\|$ to be the ℓ_1 norm with conjugate norm $\|\cdot\|_* = \|\cdot\|_\infty$. We take the distance generating function r to be the negative entropy, i.e. $r(z) = \sum_i [z]_i \log[z]_i$. We note that both $\|\cdot\|_1$ and r are separable and in particular separate over the $\mathcal{X}$ and $\mathcal{Y}$ blocks of $\mathcal{Z}$. Finally we set

$$\|A\|_{\max} := \max_{i,j} |A_{ij}|$$

and note that this is the Lipschitz constant of the gradient mapping g under the chosen norm.

Gradient estimator. Given $w_0 = (w_0^{\mathsf{x}}, w_0^{\mathsf{y}})$ and $g(w_0) = (A^\top w_0^{\mathsf{y}}, -Aw_0^{\mathsf{x}})$, we describe the reduced-variance gradient estimator $\tilde{g}_{w_0}(w)$. First, we define the probabilities $p(w) \in \Delta^m$ and $q(w) \in \Delta^n$ according to,

$$p_i(w) := \frac{|[w^{\mathsf{y}}]_i - [w_0^{\mathsf{y}}]_i|}{\|w^{\mathsf{y}} - w_0^{\mathsf{y}}\|_1} \quad \text{and} \quad q_j(w) := \frac{|[w^{\mathsf{x}}]_j - [w_0^{\mathsf{x}}]_j|}{\|w^{\mathsf{x}} - w_0^{\mathsf{x}}\|_1}. \tag{12}$$

To compute $\tilde{g}_{w_0}$ we sample $i \sim p(w)$ and $j \sim q(w)$ independently, and set

$$\tilde{g}_{w_0}(w) := \left(A^\top w_0^{\mathsf{y}} + A_{i:} \frac{[w^{\mathsf{y}}]_i - [w_0^{\mathsf{y}}]_i}{p_i(w)}, -Aw_0^{\mathsf{x}} - A_{:j} \frac{[w^{\mathsf{x}}]_j - [w_0^{\mathsf{x}}]_j}{q_j(w)} \right), \tag{13}$$

where $A_{i:}$ and $A_{:j}$ are the ith row and jth column of A, respectively. Since the sampling distributions $p(w), q(w)$ are proportional to the absolute value of the difference between blocks of w and w_0, we call strategy (12) "sampling from the difference." Substituting (12) into (13) gives the explicit form

$$\tilde{g}_{w_0}(w) = g(w_0) + (A_{i:}\|w^{\mathsf{y}} - w_0^{\mathsf{y}}\|_1 \mathrm{sign}([w^{\mathsf{y}} - w_0^{\mathsf{y}}]_i), -A_{:j}\|w^{\mathsf{x}} - w_0^{\mathsf{x}}\|_1 \mathrm{sign}([w^{\mathsf{x}} - w_0^{\mathsf{x}}]_j)). \tag{14}$$

A straightforward calculation shows that this construction satisfies Definition 2.

Lemma 2. *In the ℓ_1-ℓ_1 setup, the estimator* (14) *is* (w_0, L)-*centered with* $L = \|A\|_{\max}$.

Proof. The first property ($\mathbb{E}\tilde{g}_{w_0}(w) = g(w)$) follows immediately by inspection of (13). The second property follows from (14) by noting that

$$\|\tilde{g}_{w_0}(w) - g(w_0)\|_\infty = \max\left\{ \|A_{i:}\|_\infty \|w^{\mathsf{y}} - w_0^{\mathsf{y}}\|_1, \|A_{:j}\|_\infty \|w^{\mathsf{x}} - w_0^{\mathsf{x}}\|_1 \right\} \le \|A\|_{\max} \|w - w_0\|_1$$

for all i, j, and therefore $\mathbb{E} \|\tilde{g}_{w_0}(w) - g(w_0)\|_\infty^2 \le \|A\|_{\max}^2 \|w - w_0\|_1^2$. $\qquad\square$

The proof of Lemma 2 reveals that the proposed estimator satisfies a stronger version of Definition 2: the last property and also Lemma 1 hold with probability 1 rather than in expectation.

Runtime bound. Combining the centered gradient estimator (13), the relaxed oracle implementation (Algorithm 2) and the extragradient outer loop (Algorithm 1), we obtain our main result for ℓ_1-ℓ_1 games: an accelerated stochastic variance reduction algorithm. We write the resulting complete method explicitly as Algorithm 3 in Appendix C.1. The algorithm enjoys the following runtime guarantee (see proof in Appendix C.2).

8

Theorem 1. *Let $A \in \mathbb{R}^{m \times n}$, $\epsilon > 0$, and $\alpha \geq \epsilon/\log(nm)$. Algorithm 3 outputs a point $z = (z^{\mathsf{x}}, z^{\mathsf{y}})$ such that $\mathbb{E}\left[\max_{y \in \Delta^m} y^\top A z^{\mathsf{x}} - \min_{x \in \Delta^n}(z^{\mathsf{y}})^\top A x\right] = \mathbb{E}\left[\max_i [Az^{\mathsf{x}}]_i - \min_j [A^\top z^{\mathsf{y}}]_j\right] \leq \epsilon$, and runs in time*

$$O\left(\left(\mathrm{nnz}(A) + \frac{(m+n)\|A\|_{\max}^2}{\alpha^2}\right)\frac{\alpha \log(mn)}{\epsilon}\right). \tag{15}$$

Setting α optimally, the running time is

$$O\left(\mathrm{nnz}(A) + \frac{\sqrt{\mathrm{nnz}(A)(m+n)}\,\|A\|_{\max}\log(mn)}{\epsilon}\right). \tag{16}$$

4.2 ℓ_2-ℓ_1 games

Setup.   We set $\mathcal{X} = \mathbb{B}^n$ to be the n-dimensional Euclidean ball of radius 1, while $\mathcal{Y} = \Delta^m$ remains the simplex. For $z = (z^{\mathsf{x}}, z^{\mathsf{y}}) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ we define a norm by

$$\|z\|^2 = \|z^{\mathsf{x}}\|_2^2 + \|z^{\mathsf{y}}\|_1^2 \quad \text{with dual norm} \quad \|g\|_*^2 = \|g^{\mathsf{x}}\|_2^2 + \|g^{\mathsf{y}}\|_\infty^2.$$

For distance generating function we take $r(z) = r^{\mathsf{x}}(z^{\mathsf{x}}) + r^{\mathsf{y}}(z^{\mathsf{y}})$ with $r^{\mathsf{x}}(x) = \frac{1}{2}\|x\|_2^2$ and $r^{\mathsf{y}}(y) = \sum_i y_i \log y_i$; r is 1-strongly convex w.r.t. to $\|\cdot\|$ and has range $\frac{1}{2} + \log m \leq \log(2m)$. Finally, we denote

$$\|A\|_{2\to\infty} = \max_{i \in [m]} \|A_{i:}\|_2,$$

and note that this is the Lipschitz constant of g under $\|\cdot\|$.

Gradient estimator.   To account for the fact that $\mathcal{X}$ is now the ℓ_2 unit ball, we modify the sampling distribution q in (12) to $q_j(w) = \frac{([w^{\mathsf{x}}]_j - [w_0^{\mathsf{x}}]_j)^2}{\|w^{\mathsf{x}} - w_0^{\mathsf{x}}\|_2^2}$, and keep p the same. As we explain in detail in Appendix D.1.1, substituting these probabilities into the expression (13) yields a centered gradient estimator with a constant $(\sum_{j \in [n]} \|A_{:j}\|_\infty^2)^{1/2}$ that is larger than $\|A\|_{2\to\infty}$ by a factor of up to $\sqrt{n}$. Using local norms analysis allows us to tighten these bounds whenever the stochastic steps have bounded infinity norm. Following Clarkson et al. [6], we enforce such bound on the step norms via gradient clipping. The final gradient estimator is

$$\tilde{g}_{w_0}(w) := \left(A^\top w_0^{\mathsf{y}} + A_{i:}\frac{\|w^{\mathsf{y}} - w_0^{\mathsf{y}}\|_1}{\mathrm{sign}([w^{\mathsf{y}} - w_0^{\mathsf{y}}]_i)}, -Aw_0^{\mathsf{x}} - \mathsf{T}_\tau\left(A_{:j}\frac{\|w^{\mathsf{x}} - w_0^{\mathsf{x}}\|_2^2}{[w^{\mathsf{x}}]_j - [w_0^{\mathsf{x}}]_j}\right)\right),$$

$$\text{where } [\mathsf{T}_\tau(v)]_i = \begin{cases} -\tau & [v]_i < -\tau \\ [v]_i & -\tau \leq [v]_i \leq \tau \\ \tau & [v]_i > \tau, \end{cases}$$

The clipping operation T_τ introduces bias to the gradient estimator, which we account for by carefully choosing a value of τ for which the bias is on the same order as the variance, and yet the resulting steps are appropriately bounded; see Appendix D.1.2. In Appendix F.3.1 we describe an alternative gradient estimator for which the distribution q does not depend on the current iterate w.

Runtime bound.   Algorithm 4 in Appendix D.5 combines our clipped gradient estimator with our general variance reduction framework. The analysis in Appendix D gives the following guarantee.

Theorem 2. *Let $A \in \mathbb{R}^{m \times n}$, $\epsilon > 0$, and any $\alpha \geq \epsilon/\log(2m)$. Algorithm 4 outputs a point $z = (z^{\mathsf{x}}, z^{\mathsf{y}})$ such that $\mathbb{E}\left[\max_{y \in \Delta^m} y^\top A z^{\mathsf{x}} - \min_{x \in \mathbb{B}^n}(z^{\mathsf{y}})^\top A x\right] = \mathbb{E}\left[\max_i [Az^{\mathsf{x}}]_i + \|A^\top z^{\mathsf{y}}\|_2\right] \leq \epsilon$, and runs in time*

$$O\left(\left(\mathrm{nnz}(A) + \frac{(m+n)\|A\|_{2\to\infty}^2}{\alpha^2}\right)\frac{\alpha \log(2m)}{\epsilon}\right). \tag{17}$$

Setting α optimally, the running time is

$$O\left(\mathrm{nnz}(A) + \frac{\sqrt{\mathrm{nnz}(A)(m+n)}\,\|A\|_{2\to\infty}\log(2m)}{\epsilon}\right). \tag{18}$$

References

[1] Z. Allen-Zhu. Katyusha: the first direct acceleration of stochastic gradient methods. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1200–1205, 2017.

[2] Z. Allen-Zhu and E. Hazan. Variance reduction for faster non-convex optimization. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 699–707, 2016.

[3] P. Balamurugan and F. R. Bach. Stochastic variance reduction methods for saddle-point problems. In *Advances in Neural Information Processing Systems*, 2016.

[4] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[5] T. Chavdarova, G. Gidel, F. Fleuret, and S. Lacoste-Julien. Reducing noise in GAN training with variance reduced extragradient. In *Advances in Neural Information Processing Systems*, 2019.

[6] K. L. Clarkson, E. Hazan, and D. P. Woodruff. Sublinear optimization for machine learning. In *51th Annual IEEE Symposium on Foundations of Computer Science*, pages 449–457, 2010.

[7] M. B. Cohen, Y. T. Lee, and Z. Song. Solving linear programs in the current matrix multiplication time. *arXiv preprint arXiv:1810.07896*, 2018.

[8] G. B. Dantzig. *Linear Programming and Extensions*. Princeton University Press, Princeton, NJ, 1953.

[9] C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng. Training GANs with optimism. In *International Conference on Learning Representations*, 2019.

[10] Y. Drori, S. Sabach, and M. Teboulle. A simple algorithm for a class of nonsmooth convex-concave saddle-point problems. *Operations Research Letters*, 43(2):209–214, 2015.

[11] J. Eckstein. Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming. *Mathematics of Operations Research*, 18(1):202–226, 1993.

[12] C. Fang, C. J. Li, Z. Lin, and T. Zhang. SPIDER: near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, 2018.

[13] R. Frostig, R. Ge, S. Kakade, and A. Sidford. Un-regularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2540–2548, 2015.

[14] G. Gidel, T. Jebara, and S. Lacoste-Julien. Frank-Wolfe algorithms for saddle point problems. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017.

[15] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

[16] M. D. Grigoriadis and L. G. Khachiyan. A sublinear-time randomized approximation algorithm for matrix games. *Operation Research Letters*, 18(2):53–58, 1995.

[17] J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I*. Springer, New York, 1993.

[18] C. Jin, P. Netrapalli, and M. I. Jordan. Minmax optimization: Stable limit points of gradient descent ascent are locally optimal. *arXiv preprint arXiv:1902.00618*, 2019.

[19] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, 2013.

[20] O. Kolossoski and R. D. Monteiro. An accelerated non-Euclidean hybrid proximal extragradient-type algorithm for convex-concave saddle-point problems. *Optimization Methods and Software*, 32(6):1244–1272, 2017.

[21] Y. T. Lee and A. Sidford. Efficient inverse maintenance and faster algorithms for linear programming. In *IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 230–249, 2015.

[22] H. Lin, J. Mairal, and Z. Harchaoui. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems*, 2015.

[23] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[24] P. Mertikopoulos, H. Zenati, B. Lecouat, C.-S. Foo, V. Chandrasekhar, and G. Piliouras. Mirror descent in saddle-point problems: Going the extra (gradient) mile. In *International Conference on Learning Representations*, 2019.

[25] M. Minsky and S. Papert. *Perceptrons—an introduction to computational geometry*. MIT Press, 1987.

[26] K. Mishchenko, D. Kovalev, E. Shulgin, P. Richtárik, and Y. Malitsky. Revisiting stochastic extragradient. *arXiv preprint arXiv:1905.11373*, 2019.

[27] A. Mokhtari, A. Ozdaglar, and S. Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. *arXiv preprint arXiv:1901.08511*, 2019.

[28] 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.

[29] 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.

[30] A. Nemirovsky and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. J. Wiley & Sons, New York, NY, 1983.

[31] Y. Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programing*, 109(2-3):319–344, 2007.

[32] J. V. Neumann. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100:295–320, 1928.

[33] S. J. Reddi, A. Hefny, S. Sra, B. Poczos, and A. Smola. Stochastic variance reduction for nonconvex optimization. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 314–323, 2016.

[34] M. W. Schmidt, N. L. Roux, and F. R. Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programing*, 162(1-2):83–112, 2017.

[35] S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss. *Journal of Machine Learning Research*, 14(1):567–599, 2013.

[36] S. Shalev-Shwartz and T. Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *Mathematical Programing*, 155(1-2):105–145, 2016.

[37] S. Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012.

[38] J. Sherman. Area-convexity, ℓ_∞ regularization, and undirected multicommodity flow. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 452–460. ACM, 2017.

[39] Z. Shi, X. Zhang, and Y. Yu. Bregman divergence for stochastic variance reduction: Saddle-point and adversarial prediction. In *Advances in Neural Information Processing Systems*, 2017.

[40] A. Sidford and K. Tian. Coordinate methods for accelerating ℓ_∞ regression and faster approximate maximum flow. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science*, pages 922–933. IEEE, 2018.

[41] T. Strohmer and R. Vershynin. A randomized Kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262, 2009.

[42] M. D. Vose. A linear algorithm for generating random numbers with a given distribution. *IEEE Transactions on software engineering*, 17(9):972–975, 1991.

[43] B. E. Woodworth and N. Srebro. Tight complexity bounds for optimizing composite objectives. In *Advances in Neural Information Processing Systems*, 2016.

[44] L. Xiao and T. Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

[45] D. Zhou, P. Xu, and Q. Gu. Stochastic nested variance reduced gradient descent for nonconvex optimization. In *Advances in Neural Information Processing Systems*, 2018.

Appendix

A   Standard results

Below we give two standard results in convex optimization: bounding suboptimality via regret (Section A.1) and the mirror descent regret bound (Section A.2).

A.1   Duality gap bound

Let $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be convex in $\mathcal{X}$, concave in $\mathcal{Y}$ and differentiable, and let $g(z) = g(x, y) = (\nabla_x f(x, y), -\nabla_y f(x, y))$. For $z, u \in \mathcal{Z}$ define

$$\mathrm{gap}(z; u) := f(z^{\mathsf{x}}, u^{\mathsf{y}}) - f(u^{\mathsf{x}}, z^{\mathsf{y}}) \quad \text{and} \quad \mathrm{Gap}(z) := \max_{u \in \mathcal{Z}} \mathrm{gap}(z; u).$$

Lemma 3. *For every $z_1, \ldots, z_K \in \mathcal{Z}$,*

$$\mathrm{Gap}\left(\frac{1}{K} \sum_{k=1}^{K} z_k\right) \leq \max_{u \in \mathcal{Z}} \frac{1}{K} \sum_{k=1}^{K} \langle g(z_k), z_k - u \rangle.$$

Proof. Note that $\mathrm{gap}(z; u)$ is concave in u for every z, and that $\mathrm{gap}(z; z) = 0$, therefore

$$\mathrm{gap}(z; u) \leq \langle \nabla_u \mathrm{gap}(z; z), u - z \rangle = \langle g(z), z - u \rangle.$$

Moreover, $\mathrm{gap}(z; u)$ is convex in z for every u. Therefore, for a sequence $z_1, \ldots, z_K$ and any $u \in \mathcal{Z}$

$$\mathrm{gap}\left(\frac{1}{K} \sum_{k=1}^{K} z_k; u\right) \leq \frac{1}{K} \sum_{k=1}^{K} \mathrm{gap}(z_k; u) \leq \frac{1}{K} \sum_{k=1}^{K} \langle g(z_k), z_k - u \rangle.$$

Maximizing the inequality over u yields the lemma. $\qquad\square$

A.2   The mirror descent regret bound

Recall that $V_z(z') = r(z') - r(z) - \langle \nabla r(z), z' - z \rangle$ is the Bregman divergence induced by a 1-strongly-convex distance generating function r.

Lemma 4. *Let $Q : \mathcal{Z} \to \mathbb{R}$ be convex, let $T \in \mathbb{N}$ and let $w_0 \in \mathcal{Z}$, $\gamma_0, \gamma_1, \ldots, \gamma_T \in \mathcal{Z}^*$. The sequence $w_1, \ldots, w_T$ defined by*

$$w_t = \arg\min_{w \in \mathcal{Z}} \left\{ \langle \gamma_{t-1}, w \rangle + Q(w) + V_{w_{t-1}}(w) \right\}$$

satisfies for all $u \in \mathcal{Z}$ (denoting $w_{T+1} := u$),

$$\sum_{t=1}^{T} \langle \gamma_t + \nabla Q(w_t), w_t - u \rangle \leq V_{w_0}(u) + \sum_{t=0}^{T} \left\{ \langle \gamma_t, w_t - w_{t+1} \rangle - V_{w_t}(w_{t+1}) \right\}$$

$$\leq V_{w_0}(u) + \frac{1}{2} \sum_{t=0}^{T} \|\gamma_t\|_*^2.$$

Proof. Fix $u \equiv w_{T+1} \in \mathcal{Z}$. We note that by definition w_t is the solution of a convex optimization problem with (sub)gradient $\gamma_{t-1} + \nabla Q(\cdot) + \nabla V_{w_{t-1}}(\cdot)$, and therefore by by the first-order optimality condition [cf. 17, Chapter VII] satisfies

$$\langle \gamma_{t-1} + \nabla Q(w_t) + \nabla V_{w_{t-1}}(w_t), w_t - w_{T+1} \rangle \leq 0.$$

By the equality (9) we have $-\langle \nabla V_{w_{t-1}}(w_t), w_t - w_{T+1} \rangle = V_{w_{t-1}}(w_{T+1}) - V_{w_t}(w_{T+1}) - V_{w_{t-1}}(w_t)$. Substituting and summing over $t \in [T]$ gives

$$\sum_{t=1}^{T} \langle \gamma_{t-1} + \nabla Q(w_t), w_t - w_{T+1} \rangle \leq V_{w_0}(w_{T+1}) - \sum_{t=0}^{T} V_{w_t}(w_{t+1}).$$

Rearranging the LHS and adding $\langle \gamma_T, w_T - w_{T+1} \rangle$ to both sides of the inequality gives

$$\sum_{t=1}^{T} \langle \gamma_t + \nabla Q(w_t), w_t - w_{T+1} \rangle \leq V_{w_0}(w_{T+1}) + \sum_{t=0}^{T} \left\{ \langle \gamma_t, w_t - w_{t+1} \rangle - V_{w_t}(w_{t+1}) \right\},$$

which is the first bound stated in the lemma. The second bound follows since for every t we have

$$\langle \gamma_t, w_t - w_{t+1} \rangle \overset{(i)}{\leq} \|\gamma_t\|_* \|w_t - w_{t+1}\| \overset{(ii)}{\leq} \frac{1}{2} \|\gamma_t\|_*^2 + \frac{1}{2} \|w_t - w_{t+1}\|^2 \overset{(iii)}{\leq} \frac{1}{2} \|\gamma_t\|_*^2 + V_{w_t}(w_{t+1})$$

(19)

due to (i) Hölder's inquality, (ii) Young's inequality and (iii) strong convexity of r. $\qquad \square$

B Proofs from Section 3

B.1 Derivation of the Nemirovski's conceptual prox-method

Proposition 1 (Mirror prox convergence via oracles)**.** *Let $\mathcal{O}$ be an (α,ε)-relaxed proximal oracle with respect to gradient mapping g and distance-generating function r with range at most Θ. Let $z_{1/2}, z_{3/2}, \ldots, z_{K-1/2}$ be the iterates of Algorithm 1 and let $\bar{z}_K$ be its output. Then*

$$\mathbb{E} \, \mathrm{Gap}(\bar{z}_K) \leq \mathbb{E} \max_{u \in \mathcal{Z}} \frac{1}{K} \sum_{k=1}^{K} \left\langle g(z_{k-1/2}), z_{k-1/2} - u \right\rangle \leq \frac{\alpha \Theta}{K} + \varepsilon.$$

Proof. Fix iteration k, and note that by the definition (10), $z_k = \mathrm{Prox}_{z_{k-1}}^{\alpha}(g(z_{k-1/2}))$ satisfies

$$\left\langle g(z_{k-1/2}), z_k - u \right\rangle \leq \alpha \left(V_{z_{k-1}}(u) - V_{z_k}(u) - V_{z_{k-1}}(z_k) \right) \quad \forall u \in \mathcal{Z}.$$

Summing over k, writing

$$\left\langle g(z_{k-1/2}), z_k - u \right\rangle = \left\langle g(z_{k-1/2}), z_{k-1/2} - u \right\rangle - \left\langle g(z_{k-1/2}), z_{k-1/2} - z_k \right\rangle$$

and rearranging yields

$$\sum_{k=1}^{K} \left\langle g(z_{k-1/2}), z_{k-1/2} - u \right\rangle \leq \alpha V_{z_0}(u) + \sum_{k=1}^{K} \left[\left\langle g(z_{k-1/2}), z_{k-1/2} - z_k \right\rangle - \alpha V_{z_{k-1}}(z_k) \right]$$

for all $u \in \mathcal{Z}$. Note that since z_0 minimizes r, $V_{z_0}(u) = r(u) - r(z_0) \leq \Theta$ for all u. Therefore, maximizing the above display over u and afterwards taking expectation gives

$$\mathbb{E} \max_{u \in \mathcal{Z}} \sum_{k=1}^{K} \left\langle g(z_{k-1/2}), z_{k-1/2} - u \right\rangle \leq \alpha \Theta + \sum_{k=1}^{K} \mathbb{E} \left[\left\langle g(z_{k-1/2}), z_{k-1/2} - z_k \right\rangle - \alpha V_{z_{k-1}}(z_k) \right].$$

Finally, by Definition 1, $\mathbb{E} \left[\left\langle g(z_{k-1/2}), z_{k-1/2} - z_k \right\rangle - \alpha V_{z_{k-1}}(z_k) \right] \leq \varepsilon$ for every k, and and the result follows by dividing by K and using the bound (8). $\qquad \square$

B.2 Proof of Proposition 2

Proposition 2. *Let $\alpha, L > 0$, let $w_0 \in \mathcal{Z}$ and let $\tilde{g}_{w_0}$ be (w_0, L)-centered for monotone g. Then, for $\eta = \frac{\alpha}{10L^2}$ and $T \geq \frac{4}{\eta \alpha} = \frac{40L^2}{\alpha^2}$, the iterates of Algorithm 2 satisfy*

$$\mathbb{E} \max_{u \in \mathcal{Z}} \left[\frac{1}{T} \sum_{t \in [T]} \langle g(w_t), w_t - u \rangle - \alpha V_{w_0}(u) \right] \leq 0.$$

(11)

Proof. Recall the expression $w_t = \arg\min_{w \in \mathcal{Z}} \left\{ \langle \eta \tilde{g}_{w_0}(w_{t-1}), w \rangle + \frac{\eta \alpha}{2} V_{w_0}(w) + V_{w_{t-1}}(w) \right\}$ for the iterates of Algorithm 2. We apply Lemma 4 with $Q(z) = \langle g(w_0), z \rangle + \frac{\alpha}{2} V_{w_0}(z)$ and $\gamma_t = \eta \tilde{\delta}_t$, where

$$\tilde{\delta}_t = \tilde{g}_{w_0}(w_t) - g(w_0).$$

14

Dividing through by η, the resulting regret bound reads

$$\sum_{t \in [T]} \left\langle \tilde{g}_{w_0}(w_t) + \tfrac{\alpha}{2} \nabla V_{w_0}(w_t), w_t - u \right\rangle \leq \frac{V_{w_0}(u)}{\eta} + \frac{\eta}{2} \sum_{t \in [T]} \|\tilde{\delta}_t\|_*^2, \qquad (20)$$

where we used the fact that $\tilde{\delta}_0 = 0$ to drop the summation over $t = 0$ in the RHS. Now, let

$$\tilde{\Delta}_t = g(w_t) - \tilde{g}_{w_0}(w_t).$$

Rearranging the inequality (20), we may write it as

$$\sum_{t \in [T]} \left\langle g(w_t) + \tfrac{\alpha}{2} \nabla V_{w_0}(w_t), w_t - u \right\rangle \leq \frac{V_{w_0}(u)}{\eta} + \frac{\eta}{2} \sum_{t \in [T]} \|\tilde{\delta}_t\|_*^2 + \sum_{t \in [T]} \left\langle \tilde{\Delta}_t, w_t - u \right\rangle. \qquad (21)$$

Define the "ghost iterate" sequence $s_1, s_2, \ldots, s_T$ according to

$$s_t = \arg\min_{s \in \mathcal{Z}} \left\{ \left\langle \eta \tilde{\Delta}_{t-1}, s \right\rangle + V_{s_{t-1}}(s) \right\} \quad \text{with} \quad s_0 = w_0.$$

Applying Lemma 4 with $Q = 0$ and $\gamma_t = \eta \tilde{\Delta}_t$, we have

$$\sum_{t \in [T]} \left\langle \tilde{\Delta}_t, s_t - u \right\rangle \leq \frac{V_{w_0}(u)}{\eta} + \frac{\eta}{2} \sum_{t \in [T]} \|\tilde{\Delta}_t\|_*^2, \qquad (22)$$

where here too we used $\tilde{\Delta}_0 = 0$. Writing $\left\langle \tilde{\Delta}_t, w_t - u \right\rangle = \left\langle \tilde{\Delta}_t, w_t - s_t \right\rangle + \left\langle \tilde{\Delta}_t, s_t - u \right\rangle$ and substituting (22) into (21) we have

$$\sum_{t \in [T]} \left\langle g(w_t) + \tfrac{\alpha}{2} \nabla V_{w_0}(w_t), w_t - u \right\rangle \leq \frac{2V_{w_0}(u)}{\eta} + \frac{\eta}{2} \sum_{t \in [T]} \left[\|\tilde{\delta}_t\|_*^2 + \|\tilde{\Delta}_t\|_*^2 \right] + \sum_{t \in [T]} \left\langle \tilde{\Delta}_t, w_t - s_t \right\rangle.$$

Substituting

$$-\tfrac{\alpha}{2} \left\langle \nabla V_{w_0}(w_t), w_t - u \right\rangle = \tfrac{\alpha}{2} V_{w_0}(u) - \tfrac{\alpha}{2} V_{w_t}(u) - \tfrac{\alpha}{2} V_{w_0}(w_t) \leq \tfrac{\alpha}{2} V_{w_0}(u) - \tfrac{\alpha}{2} V_{w_0}(w_t)$$

and dividing by T, we have

$$\frac{1}{T} \sum_{t \in [T]} \left\langle g(w_t), w_t - u \right\rangle \leq \left(\tfrac{2}{\eta T} + \tfrac{\alpha}{2} \right) V_{w_0}(u) + \frac{1}{T} \sum_{t \in [T]} \left[\tfrac{\eta}{2} \|\tilde{\delta}_t\|_*^2 + \tfrac{\eta}{2} \|\tilde{\Delta}_t\|_*^2 - \tfrac{\alpha}{2} V_{w_0}(w_t) + \left\langle \tilde{\Delta}_t, w_t - s_t \right\rangle \right].$$

Subtracting $\alpha V_{w_0}(u)$ from both sides and using $\tfrac{2}{\eta T} - \tfrac{\alpha}{2} \leq 0$ due to $T \geq \tfrac{4}{\eta \alpha}$, we obtain

$$\frac{1}{T} \sum_{t \in [T]} \left\langle g(w_t), w_t - u \right\rangle - \alpha V_{w_0}(u) \leq \frac{1}{T} \sum_{t \in [T]} \left[\tfrac{\eta}{2} \|\tilde{\delta}_t\|_*^2 + \tfrac{\eta}{2} \|\tilde{\Delta}_t\|_*^2 - \tfrac{\alpha}{2} V_{w_0}(w_t) + \left\langle \tilde{\Delta}_t, w_t - s_t \right\rangle \right].$$

Note that this inequality holds with probability 1 for all u. We may therefore maximize over u and then take expectation, obtaining

$$\mathbb{E} \max_{u \in \mathcal{Z}} \left\{ \frac{1}{T} \sum_{t \in [T]} \left\langle g(w_t), w_t - u \right\rangle - \alpha V_{w_0}(u) \right\}$$

$$\leq \frac{1}{T} \sum_{t \in [T]} \mathbb{E} \left[\tfrac{\eta}{2} \|\tilde{\delta}_t\|_*^2 + \tfrac{\eta}{2} \|\tilde{\Delta}_t\|_*^2 - \tfrac{\alpha}{2} V_{w_0}(w_t) + \left\langle \tilde{\Delta}_t, w_t - s_t \right\rangle \right]. \qquad (23)$$

It remains to argue the the RHS is nonpositive. By the first centered estimator property, we have

$$\mathbb{E} \left[\tilde{\Delta}_t \mid w_t, s_t \right] = \mathbb{E} \left[g(w_t) - \tilde{g}_{w_0}(w_t) \mid w_t, s_t \right] = 0$$

and therefore $\mathbb{E} \left\langle \tilde{\Delta}_t, w_t - s_t \right\rangle = 0$ for all t. By the second property

$$\mathbb{E} \|\tilde{\delta}_t\|_*^2 = \mathbb{E} \|\tilde{g}_{w_0}(w_t) - g(w_0)\|_*^2 \leq L^2 \|w_t - w_0\|^2 \leq 2L^2 V_{w_0}(w_t),$$

where the last transition used the strong convexity of r. Similarly, by Lemma 1 we have

$$\mathbb{E} \|\tilde{\Delta}_t\|_*^2 = \mathbb{E} \|\tilde{g}_{w_0}(w_t) - g(w)\|_*^2 \leq 4L^2 \|w_t - w_0\|^2 \leq 8L^2 V_{w_0}(w_t).$$

Therefore

$$\mathbb{E} \left[\tfrac{\eta}{2} \|\tilde{\delta}_t\|_*^2 + \tfrac{\eta}{2} \|\tilde{\Delta}_t\|_*^2 - \tfrac{\alpha}{2} V_{w_0}(w_t) \right] \leq (5\eta L^2 - \tfrac{\alpha}{2}) \mathbb{E} V_{w_0}(w_t) = 0,$$

using $\eta = \tfrac{\alpha}{10L^2}$. $\qquad \square$

C The ℓ_1-ℓ_1 setup

C.1 Complete pseudo-code

Algorithm 3: Variance reduction for ℓ_1-ℓ_1 games

Input: Matrix $A \in \mathbb{R}^{m \times n}$ with ith row $A_{i:}$ and jth column $A_{:j}$, target accuracy ϵ
Output: A point with expected duality gap below ϵ

1 $L \leftarrow \max_{ij} |A_{ij}|$, $\alpha \leftarrow L\sqrt{\frac{n+m}{\text{nnz}(A)}}$, $K \leftarrow \left\lceil \frac{\log(nm)\alpha}{\epsilon} \right\rceil$, $\eta \leftarrow \frac{\alpha}{10L^2}$, $T \leftarrow \left\lceil \frac{4}{\eta\alpha} \right\rceil$, $z_0 \leftarrow (\frac{1}{n}\mathbf{1}_n, \frac{1}{m}\mathbf{1}_m)$

2 **for** $k = 1, \ldots, K$ **do**

    ▷ *Relaxed oracle query:*

3      $(x_0, y_0) \leftarrow (z^{\mathsf{x}}_{k-1}, z^{\mathsf{y}}_{k-1})$, $(g^{\mathsf{x}}_0, g^{\mathsf{y}}_0) \leftarrow (A^\top y_0, -Ax_0)$

4   **for** $t = 1, \ldots, T$ **do**

        ▷ *Gradient estimation:*

5   Sample $i \sim p$ where $p_i = \frac{|[y_{t-1}]_i - [y_0]_i|}{\|y_{t-1} - y_0\|_1}$, sample $j \sim q$ where $q_j = \frac{|[x_{t-1}]_j - [x_0]_j|}{\|x_{t-1} - x_0\|_1}$

6   Set $\tilde{g}_{t-1} = g_0 + \left(A_{i:} \frac{[y_{t-1}]_i - [y_0]_i}{p_i}, -A_{:j} \frac{[x_{t-1}]_j - [x_0]_j}{q_j} \right)$

        ▷ *Mirror descent step:*

7          $x_t \leftarrow \Pi_{\mathcal{X}} \left(\frac{1}{1 + \eta\alpha/2} \left(\log x_{t-1} + \frac{\eta\alpha}{2} \log x_0 - \eta\tilde{g}^{\mathsf{x}}_{t-1} \right) \right)$ ▷ $\Pi_{\mathcal{X}}(v) = \frac{e^v}{\|e^v\|_1}$

8          $y_t \leftarrow \Pi_{\mathcal{Y}} \left(\frac{1}{1 + \eta\alpha/2} \left(\log y_{t-1} + \frac{\eta\alpha}{2} \log y_0 - \eta\tilde{g}^{\mathsf{y}}_{t-1} \right) \right)$ ▷ $\Pi_{\mathcal{Y}}(v) = \frac{e^v}{\|e^v\|_1}$

9      $z_{k-1/2} \leftarrow \frac{1}{T} \sum_{t=1}^{T} (x_t, y_t)$

    ▷ *Extragradient step:*

10      $z^{\mathsf{x}}_k \leftarrow \Pi_{\mathcal{X}} \left(\log z^{\mathsf{x}}_{k-1} - \frac{1}{\alpha} A^\top z^{\mathsf{y}}_{k-1/2} \right)$

11      $z^{\mathsf{y}}_k \leftarrow \Pi_{\mathcal{Y}} \left(\log z^{\mathsf{y}}_{k-1} + \frac{1}{\alpha} A z^{\mathsf{x}}_{k-1/2} \right)$

12 **return** $\frac{1}{K} \sum_{k=1}^{K} z_{k-1/2}$

C.2 Proof of runtime bound

Theorem 1. *Let $A \in \mathbb{R}^{m \times n}$, $\epsilon > 0$, and $\alpha \geq \epsilon/\log(nm)$. Algorithm 3 outputs a point $z = (z^{\mathsf{x}}, z^{\mathsf{y}})$ such that $\mathbb{E}\left[\max_{y \in \Delta^m} y^\top A z^{\mathsf{x}} - \min_{x \in \Delta^n} (z^{\mathsf{y}})^\top A x \right] = \mathbb{E}\left[\max_i [Az^{\mathsf{x}}]_i - \min_j [A^\top z^{\mathsf{y}}]_j \right] \leq \epsilon$, and runs in time*

$$O\left(\left(\text{nnz}(A) + \frac{(m+n)\|A\|_{\max}^2}{\alpha^2} \right) \frac{\alpha \log(mn)}{\epsilon} \right). \tag{15}$$

Setting α optimally, the running time is

$$O\left(\text{nnz}(A) + \frac{\sqrt{\text{nnz}(A)(m+n)}\, \|A\|_{\max} \log(mn)}{\epsilon} \right). \tag{16}$$

Proof. First, we prove the expected duality gap bound. By Lemma 2 and Corollary 1 (with $L = \|A\|_{\max}$), $\texttt{InnerLoop}$ is an $(\alpha, 0)$-relaxed proximal oracle. On Δ^d, negative entropy has minimum value $-\log d$ and is non-positive, therefore for the ℓ_1-ℓ_1 domain we have $\Theta = \max_{z'} r(z') - \min_z r(z) = \log(nm)$. By Proposition 1, running $K \geq \alpha \log(nm)/\epsilon$ iterations guarantees an ϵ-approximate saddle point in expectation.

Now, we prove the runtime bound. Lines 3, 10 and 11 of Algorithm 3 each take time $O(\text{nnz}(A))$, as they involve matrix-vector products with A and $A^\top$. All other lines run in time $O(n+m)$, as they consist of sampling and vector arithmetic (the time to compute sampling probabilities dominates the runtime of sampling). Therefore, the total runtime is $O((\text{nnz}(A) + (n+m)T)K)$. Substituting $T \le 1 + \frac{40L^2}{\alpha^2}$ and $K \le 1 + \frac{\log(nm)\alpha}{\epsilon}$ gives the bound (15). Setting

$$\alpha = \max\left\{ \frac{\epsilon}{\log nm}\,,\, \|A\|_{\max}\sqrt{\frac{n+m}{\text{nnz}(A)}} \right\}$$

gives the optimized bound (16). $\qquad\square$

Remark 3. We can improve the $\log(mn)$ factor in (15) and (16) to $\sqrt{\log m \log n}$ by the transformation $\mathcal{X} \to \mathcal{X}\sqrt{\frac{\log m}{\log n}}$ and $\mathcal{Y} \to \mathcal{Y}\sqrt{\frac{\log n}{\log m}}$. This transformation leaves the problem unchanged and reduces Θ from $\log(mn)$ to $2\sqrt{\log m \log n}$. It is also equivalent to proportionally using slightly different step-sizes for the $\mathcal{X}$ and $\mathcal{Y}$ block.

D The ℓ_2-ℓ_1 setup

D.1 Derivation of gradient clipping

D.1.1 Basic gradient estimator

We first present a straightforward adaptation of the ℓ_1-ℓ_1 gradient estimator, which we subsequently improve to obtain the optimal Lipschitz constant dependence. Following the "sampling from the difference" strategy, consider a gradient estimator $\tilde{g}_{w_0}$ computed as in (13), but with the following different choice of $q(w)$:

$$p_i(w) = \frac{|[w^y]_i - [w_0^y]_i|}{\|w^y - w_0^y\|_1} \quad \text{and} \quad q_j(w) = \frac{([w^x]_j - [w_0^x]_j)^2}{\|w^x - w_0^x\|_2^2}. \tag{24}$$

The resulting gradient estimator has the explicit form

$$\tilde{g}_{w_0}(w) = g(w_0) + \left(A_{i:}\frac{\|w^y - w_0^y\|_1}{\text{sign}([w^y - w_0^y]_i)}, -A_{:j}\frac{\|w^x - w_0^x\|_2^2}{[w^x - w_0^x]_j} \right). \tag{25}$$

(Note that $\tilde{g}_{w_0}$ of the form (13) is finite with probability 1.) Direct calculation shows it is centered.

Lemma 5. *In the ℓ_2-ℓ_1 setup, the estimator (25) is (w_0, L)-centered with $L = \sqrt{\sum_{j\in[n]}\|A_{:j}\|_\infty^2}$.*

Proof. The estimator is unbiased since it is of the form (13). To show the variance bound, first consider the $\mathcal{X}$-block. We have

$$\left\|\tilde{g}_{w_0}^x(w) - g^x(w_0)\right\|_2^2 = \|A_{i:}\|_2^2 \|w^y - w_0^y\|_1^2 \le \|A\|_{2\to\infty}^2 \|w^y - w_0^y\|_1^2 \le L^2\|w^y - w_0^y\|_1^2, \tag{26}$$

where we used $\|A\|_{2\to\infty}^2 = \max_{i\in[n]}\|A_{i:}\|_2^2 \le \sum_{j\in[m]}\|A_{:j}\|_\infty^2 = L^2$. Second, for the $\mathcal{Y}$-block,

$$\mathbb{E}\left\|\tilde{g}_{w_0}^y(w) - g^y(w_0)\right\|_\infty^2 = \sum_{j\in[n]}\frac{\|A_{:j}\|_\infty^2 [w^x - w_0^x]_j^2}{q_j(w)} = L^2\|w^x - w_0^x\|_2^2. \tag{27}$$

Combining (26) and (27), we have the second property $\mathbb{E}\|\tilde{g}_{w_0}(w) - g(w_0)\|_*^2 \le L^2\|w - w_0\|^2$. $\quad\square$

D.1.2 Improved gradient estimator

The constant L in Lemma 5 is larger than the Lipschitz constant of g (i.e. $\|A\|_{2\to\infty}$) by a factor of up to $\sqrt{n}$. Consequently, a variance reduction scheme based on the estimator (25) will not always improve on the linear-time mirror prox method.

Inspecting the proof of Lemma 5, we see that the cause for the inflated value of L is the bound (27) on $\mathbb{E}\left\|\tilde{g}_{w_0}^{\mathsf{y}}(w) - g^{\mathsf{y}}(w_0)\right\|_\infty^2$. We observe that swapping the order of expectation and maximization would solve the problem, as

$$\max_{k \in [m]} \mathbb{E}\left[\tilde{g}_{w_0}^{\mathsf{y}}(w) - g^{\mathsf{y}}(w_0)\right]_k^2 = \max_{k \in [m]} \sum_{j \in [n]} \frac{A_{kj}^2 [w^{\mathsf{x}} - w_0^{\mathsf{x}}]_j^2}{q_j(w)} = \|A\|_{2 \to \infty}^2 \|w^{\mathsf{x}} - w_0^{\mathsf{x}}\|_2^2. \qquad (28)$$

Moreover, inspecting the proof of Proposition 2 reveals that instead of bounding terms of the form $\mathbb{E}\left\|\tilde{g}_{w_0}^{\mathsf{y}}(w_t) - g^{\mathsf{y}}(w_0)\right\|_\infty^2$ we may directly bound $\mathbb{E}\left[\eta\langle \tilde{g}_{w_0}^{\mathsf{y}}(w_t) - g^{\mathsf{y}}(w_0), y_t - y_{t+1}\rangle - V_{y_t}(y_{t+1})\right]$, where we write $w_t = (x_t, y_t)$ and recall that η is the step-size in Algorithm 2. Suppose that $\eta\left\|\tilde{g}_{w_0}^{\mathsf{y}}(w_t) - g^{\mathsf{y}}(w_0)\right\|_\infty \leq 1$ holds. In this case we may use a "local norms" bound (Lemma **??** in Appendix D.2) to write

$$\eta\langle \tilde{g}_{w_0}^{\mathsf{y}}(w_t) - g^{\mathsf{y}}(w_0), y_t - y_{t+1}\rangle - V_{y_t}(y_{t+1}) \leq \eta^2 \sum_{k \in [m]} [y_t]_k [\tilde{g}_{w_0}^{\mathsf{y}}(w_t) - g^{\mathsf{y}}(w_0)]_k^2$$

and bound the expectation of the RHS using (28) conditional on w_t.

Unfortunately, the gradient estimator (25) does not always satisfy $\eta\left\|\tilde{g}_{w_0}^{\mathsf{y}}(w_t) - g^{\mathsf{y}}(w_0)\right\|_\infty \leq 1$. Following Clarkson et al. [6], we enforce this bound by clipping the gradient estimates, yielding the estimator

$$\tilde{g}_{w_0}(w) := \left(A^\top w_0^{\mathsf{y}} + A_{i:} \frac{[w^{\mathsf{y}}]_i - [w_0^{\mathsf{y}}]_i}{p_i(w)}, -A w_0^{\mathsf{x}} - \mathsf{T}_\tau\left(A_{:j} \frac{[w^{\mathsf{x}}]_j - [w_0^{\mathsf{x}}]_j}{q_j(w)}\right)\right),$$

$$\text{where } [\mathsf{T}_\tau(v)]_i = \begin{cases} -\tau & [v]_i < -\tau \\ [v]_i & -\tau \leq [v]_i \leq \tau \\ \tau & [v]_i > \tau, \end{cases} \qquad (29)$$

where $i \sim p(w)$ and $j \sim q(w)$ with p, q as defined in (24). The clipping in (29) does not significantly change the variance of the estimator, but it introduces some bias for which we must account. We summarize the relevant properties of the clipped gradient estimator in the following.

Definition 3. *Let $w_0 = (w_0^{\mathsf{x}}, w_0^{\mathsf{y}}) \in \mathcal{Z}$ and $\tau, L > 0$. A stochastic gradient estimator $\tilde{g}_{w_0} : \mathcal{Z} \to \mathcal{Z}^*$ is called (w_0, L, τ)-centered-bounded-biased (CBB) if it satisfies for all $w = (w^{\mathsf{x}}, w^{\mathsf{y}}) \in \mathcal{Z}$,*

1. *$\mathbb{E}\tilde{g}_{w_0}^{\mathsf{x}}(w) = g^{\mathsf{x}}(w)$ and $\left\|\mathbb{E}\tilde{g}_{w_0}^{\mathsf{y}}(w) - g^{\mathsf{y}}(w)\right\|_* \leq \frac{L^2}{\tau}\|w - w_0\|^2$,*

2. *$\left\|\tilde{g}_{w_0}^{\mathsf{y}}(w) - g^{\mathsf{y}}(w_0)\right\|_* \leq \tau$ and $\left\|\tilde{g}_{w_0}^{\mathsf{y}}(w) - g^{\mathsf{y}}(w)\right\|_* \leq 2L + \tau$,*

3. *$\mathbb{E}\left\|\tilde{g}_{w_0}^{\mathsf{x}}(w) - g^{\mathsf{x}}(w_0)\right\|_*^2 + \max_{i \in [m]} \mathbb{E}\left[\tilde{g}_{w_0}^{\mathsf{y}}(w) - g^{\mathsf{y}}(w_0)\right]_i^2 \leq L^2\|w - w_0\|^2.$*

Lemma 6. *In the ℓ_2-ℓ_1 setup, the estimator (29) is (w_0, L, τ)-CBB with $L = \|A\|_{2 \to \infty}$.*

Proof. The $\mathcal{X}$ component for the gradient estimator is unbiased. We bound the bias in the $\mathcal{Y}$ block as follows. Fixing an index $i \in [m]$, we have

$$\left|\mathbb{E}\left[\tilde{g}_{w_0}^{\mathsf{y}}(w) - g^{\mathsf{y}}(w)\right]_i\right| = \left|\mathbb{E}_j\left[A_{ij}\frac{[w^{\mathsf{x}}]_j - [w_0^{\mathsf{x}}]_j}{q_j} - \mathsf{T}_\tau\left(A_{ij}\frac{[w^{\mathsf{x}}]_j - [w_0^{\mathsf{x}}]_j}{q_j}\right)\right]\right|$$

$$\leq \sum_{j \in \mathcal{J}_\tau(i)} q_j\left|A_{ij}\frac{[w^{\mathsf{x}}]_j - [w_0^{\mathsf{x}}]_j}{q_j} - \mathsf{T}_\tau\left(A_{ij}\frac{[w^{\mathsf{x}}]_j - [w_0^{\mathsf{x}}]_j}{q_j}\right)\right|$$

$$\leq \sum_{j \in \mathcal{J}_\tau(i)} |A_{ij}|\,|[w^{\mathsf{x}}]_j - [w_0^{\mathsf{x}}]_j|$$

where the last transition used $|a - \mathsf{T}_\tau(a)| \leq |a|$ for all a, and

$$\mathcal{J}_\tau(i) = \left\{j \in [n] \mid \mathsf{T}_\tau\left(A_{ij}\frac{[w^{\mathsf{x}}]_j - [w_0^{\mathsf{x}}]_j}{q_j}\right) \neq A_{ij}\frac{[w^{\mathsf{x}}]_j - [w_0^{\mathsf{x}}]_j}{q_j}\right\}.$$

Note that $j \in \mathcal{J}_\tau(i)$ if and only if

$$\left| A_{ij} \frac{[w^\times]_j - [w_0^\times]_j}{q_j} \right| = \frac{\|w^\times - w_0^\times\|_2^2 |A_{ij}|}{|[w^\times]_j - [w_0^\times]_j|} > \tau \Rightarrow |[w^\times]_j - [w_0^\times]_j| \leq \frac{1}{\tau} \|w^\times - w_0^\times\|_2^2 |A_{ij}|.$$

Therefore,

$$\sum_{j \in \mathcal{J}_\tau(i)} |A_{ij}| \, |[w^\times]_j - [w_0^\times]_j| \leq \frac{1}{\tau} \|w^\times - w_0^\times\|_2^2 \sum_{j \in \mathcal{J}_\tau(i)} |A_{ij}|^2 \leq \frac{1}{\tau} \|w^\times - w_0^\times\|_2^2 \|A_{i:}\|_2^2$$

and $\left\| \mathbb{E} \tilde{g}_{w_0}^\mathsf{y}(w) - g^\mathsf{y}(w) \right\|_\infty \leq \frac{L^2}{\tau} \|w^\times - w_0^\times\|_2^2$ follows by taking the maximum over $i \in [m]$.

By definition of T_τ we have $\left\| \tilde{g}_{w_0}^\mathsf{y}(w) - g^\mathsf{y}(w_0) \right\|_\infty \leq \tau$ and by the triangle inequality and L-Lipschitz continuity of g we have

$$\left\| \tilde{g}_{w_0}^\mathsf{y}(w) - g^\mathsf{y}(w) \right\|_\infty \leq \left\| g^\mathsf{y}(w) - g^\mathsf{y}(w_0) \right\|_\infty + \left\| \tilde{g}_{w_0}^\mathsf{y}(w) - g^\mathsf{y}(w_0) \right\|_\infty \leq L \|w^\times - w_0^\times\|_2 + \tau \leq 2L + \tau,$$

since we assume $\mathcal{X}$ is the unit Euclidean ball.

Finally, we note that for all k, the addition of T_τ never increases $[\tilde{g}_{w_0}^\mathsf{y}(w) - g^\mathsf{y}(w_0)]_k^2$, and so the third property follows from (28) and (26). $\qquad \square$

To guarantee $\eta \left\| \tilde{g}_{w_0}^\mathsf{y}(w_t) - g^\mathsf{y}(w_0) \right\|_\infty \leq 1$, we set the threshold τ to be $1/\eta$. By the first property in Definition 3, the bias caused by this choice of τ is of the order of the variance of the estimator, and we may therefore cancel it with the regularizer by choosing η slightly smaller than in Proposition 2. In Appendix D we prove (using the observations from the preceding discussion) that Algorithm 2 with a CBB gradient estimator implements a relaxed proximal oracle.

Proposition 3. *In the ℓ_2-ℓ_1 setup, let $\alpha, L > 0$, let $w_0 \in \mathcal{Z}$ and let $\tilde{g}_{w_0}$ be $(w_0, L, \frac{20L^2}{\alpha})$-CBB for monotone g. Then, for $\alpha \leq L$, $\eta = \frac{\alpha}{20L^2}$ and $T \geq \frac{4}{\eta \alpha} = \frac{80L^2}{\alpha^2}$, the iterates of Algorithm 2 satisfy the bound* (11). *Moreover, for $g(z) = (A^\top z^\mathsf{y}, -Az^\times)$, $\mathcal{O}(w_0) = \mathtt{InnerLoop}(w_0, \tilde{g}_{w_0}, \alpha)$ is an $(\alpha, 0)$-relaxed proximal oracle.*

We remark that the proof of Proposition 3 relies on the structure of the simplex with negative entropy as the distance generating function. For this reason, we state the proposition for the ℓ_2-ℓ_1 setup. However, Proposition 3 would also hold for other setups where $\mathcal{Y}$ is the simplex and r^y is the negative entropy, provided a CBB gradient estimator is available.

With Proposition 3 in hand, the proof of Theorem 2 follows identically to that of Theorem 1, except Proposition 3 replaces Corollary 1, L is now $\|A\|_{2 \to \infty}$ instead of $\|A\|_{\max}$, and $\Theta = \max_{z'} r(z') - \min_z r(z) = \frac{1}{2} + \log m \leq \log(2m)$ rather than $\log(mn)$.

Before giving the proof of Proposition 3 is Section D.4, we first collect some properties of the KL divergence (Section D.2) and of centered-bounded-biased (CBB) gradient estimators (Section D.3).

D.2 Local norms bounds

For this subsection, let $\mathcal{Y}$ be the m dimensional simplex Δ^m, and let $r(y) = \sum_{i=1}^m y_i \log y_i$ be the negative entropy distance generating function. The corresponding Bregman divergence is the KL divergence, which is well-defined for any $y, y' \in \mathbb{R}_{\geq 0}^m$ and has the form

$$V_y(y') = \sum_{i \in [m]} \left[y_i' \log \frac{y_i'}{y_i} + y_i - y_i' \right] = \int_0^1 dt \int_0^t \sum_{i \in [m]} \frac{(y_i - y_i')^2}{(1 - \tau) y_i + \tau y_i'} d\tau. \qquad (30)$$

In the literature, "local norms" regret analysis [37, Section 2.8] relies on the fact that $r^*(\gamma) = \log(\sum_i e^{\gamma_i})$ (the conjugate of negative entropy in the simplex) is locally smooth with respect to a Euclidean norm weighted by $\nabla r^*(\gamma) = \frac{e^\gamma}{\|e^\gamma\|_1}$. More precisely, the Bregman divergence $V_\gamma^*(\gamma') = r^*(\gamma') - r^*(\gamma) - \langle \nabla r^*(\gamma), \gamma' - \gamma \rangle$ satisfies

$$V_\gamma^*(\gamma + \delta) \leq \|\delta\|_{\nabla r^*(\gamma)}^2 := \sum_i [\nabla r^*(\gamma)]_i \cdot \delta_i^2 \quad \text{whenever } \delta_i \leq 1.79 \; \forall i. \qquad (31)$$

Below, we state this bound in a form that is directly applicable to our analysis.

Lemma 7. *Let $y, y' \in \Delta^m$ and $\delta \in \mathbb{R}^m$. If δ satisfies $\delta_i \le 1.79$ for all $i \in [m]$ then the KL divergence $V_y(y')$ satisfies*

$$\langle \delta, y' - y \rangle - V_y(y') \le \|\delta\|_y^2 := \sum_{i \in [m]} y_i \delta_i^2$$

Proof. It suffices to consider y in the relative interior of the simplex where r is differentiable; the final result will hold for any y in the simplex by continuity. Recall the following general facts about convex conjugates: $\langle \gamma', y' \rangle - r(y') \le r^*(\gamma')$ for any $\gamma' \in \mathbb{R}^m$, $y = \nabla r^*(\nabla r(y))$ and $r^*(\nabla r(y)) = \langle \nabla r(y), y \rangle - r(y)$. Therefore, we have for all $y' \in \Delta^m$,

$$\langle \delta, y' - y \rangle - V_y(y') = \langle \nabla r(y) + \delta, y' \rangle - r(y') - [\langle \nabla r(y), y \rangle - r(y)] - \langle y, \delta \rangle$$
$$\le r^*(\nabla r(y) + \delta) - r^*(\nabla r(y)) - \langle \nabla r^*(\nabla r(y)), \delta \rangle = V_{\nabla r(y)}^*(\nabla r(y) + \delta).$$

The result follows from (31) with $\gamma = \nabla r(y)$, recalling again that $y = \nabla r^*(\nabla r(y))$. For completeness we prove (31) below, following [37]. We have

$$r^*(\gamma + \delta) - r^*(\gamma) = \log \left(\frac{\sum_{i \in [m]} e^{\gamma_i + \delta_i}}{\sum_{i \in [m]} e^{\gamma_i}} \right) \overset{(i)}{\le} \log \left(1 + \frac{\sum_{i \in [m]} e^{\gamma_i}(\delta_i + \delta_i^2)}{\sum_{i \in [m]} e^{\gamma_i}} \right)$$
$$= \log(1 + \langle \nabla r^*(\gamma), \delta + \delta^2 \rangle) \overset{(ii)}{\le} \langle \nabla r^*(\gamma), \delta \rangle + \langle \nabla r^*(\gamma), \delta^2 \rangle,$$

where (i) follows from $e^x \le 1 + x + x^2$ for all $x \le 1.79$ and (ii) follows from $\log(1 + x) \le x$ for all x. Therefore,

$$V_\gamma^*(\gamma + \delta) = r^*(\gamma + \delta) - r^*(\gamma) - \langle \nabla r^*(\gamma), \delta \rangle \le \langle \nabla r^*(\gamma), \delta^2 \rangle = \|\delta\|_{\nabla r^*(\gamma)}^2,$$

completing the proof. $\qquad \square$

D.3 Properties of CBB gradient estimators

We recall the definition of a centered-bounded-biased gradient estimator.

Definition 3. *Let $w_0 = (w_0^{\mathsf{x}}, w_0^{\mathsf{y}}) \in \mathcal{Z}$ and $\tau, L > 0$. A stochastic gradient estimator $\tilde{g}_{w_0} : \mathcal{Z} \to \mathcal{Z}^*$ is called (w_0, L, τ)-centered-bounded-biased (CBB) if it satisfies for all $w = (w^{\mathsf{x}}, w^{\mathsf{y}}) \in \mathcal{Z}$,*

1. $\mathbb{E}\tilde{g}_{w_0}^{\mathsf{x}}(w) = g^{\mathsf{x}}(w)$ *and* $\left\| \mathbb{E}\tilde{g}_{w_0}^{\mathsf{y}}(w) - g^{\mathsf{y}}(w) \right\|_* \le \frac{L^2}{\tau} \|w - w_0\|^2$,

2. $\left\| \tilde{g}_{w_0}^{\mathsf{y}}(w) - g^{\mathsf{y}}(w_0) \right\|_* \le \tau$ *and* $\left\| \tilde{g}_{w_0}^{\mathsf{y}}(w) - g^{\mathsf{y}}(w) \right\|_* \le 2L + \tau$,

3. $\mathbb{E} \left\| \tilde{g}_{w_0}^{\mathsf{x}}(w) - g^{\mathsf{x}}(w_0) \right\|_*^2 + \max_{i \in [m]} \mathbb{E} [\tilde{g}_{w_0}^{\mathsf{y}}(w) - g^{\mathsf{y}}(w_0)]_i^2 \le L^2 \|w - w_0\|^2$.

CBB estimators have the following additional property, analogous to Lemma 1.

Lemma 8. *In the ℓ_2-ℓ_1 setup, a (w_0, L, τ)-CBB estimator with for g with $\tau \ge 2\sqrt{2}L$ also satisfies, for all $w \in \mathcal{Z}$,*

$$\mathbb{E} \left\| \tilde{g}_{w_0}^{\mathsf{x}}(w) - g^{\mathsf{x}}(w) \right\|_2^2 + \max_{i \in [m]} \mathbb{E} [\tilde{g}_{w_0}^{\mathsf{y}}(w) - g^{\mathsf{y}}(w)]_i^2 \le 2L^2 \|w - w_0\|^2.$$

Proof. We have $\mathbb{E} \left\| \tilde{g}_{w_0}^{\mathsf{x}}(w) - g^{\mathsf{x}}(w) \right\|_2^2 \le \mathbb{E} \left\| \tilde{g}_{w_0}^{\mathsf{x}}(w) - g^{\mathsf{x}}(w_0) \right\|_2^2$ since the $\mathcal{X}$ component is unbiased. For the $\mathcal{Y}$ component, fix $i \in [m]$ and write

$$\mathbb{E} [\tilde{g}_{w_0}^{\mathsf{y}}(w) - g^{\mathsf{y}}(w)]_i^2 = \mathbb{E} [\tilde{g}_{w_0}^{\mathsf{y}}(w) - \mathbb{E}\tilde{g}_{w_0}^{\mathsf{y}}(w)]_i^2 + [\mathbb{E}\tilde{g}_{w_0}^{\mathsf{y}}(w) - g^{\mathsf{y}}(w)]_i^2$$
$$\le \mathbb{E} [\tilde{g}_{w_0}^{\mathsf{y}}(w) - g(w_0)]_i^2 + \left(\frac{L^2}{\tau} \|w - w_0\|^2 \right)^2,$$

where the last inequality follows from the first CBB property and the fact that $[v]_i^2 \le \|v\|_\infty^2$. Using $\tau \ge 2\sqrt{2}L$ and $\|w - w_0\| \le 2\sqrt{2}$ for every $w, w_0 \in \mathbb{B}^n \times \Delta^m$, we obtain the result. $\qquad \square$

D.4 Proof of Proposition 3

Proposition 3. *In the ℓ_2-ℓ_1 setup, let $\alpha, L > 0$, let $w_0 \in \mathcal{Z}$ and let $\tilde{g}_{w_0}$ be $(w_0, L, \frac{20L^2}{\alpha})$-CBB for monotone g. Then, for $\alpha \leq L$, $\eta = \frac{\alpha}{20L^2}$ and $T \geq \frac{4}{\eta\alpha} = \frac{80L^2}{\alpha^2}$, the iterates of Algorithm 2 satisfy the bound* (11). *Moreover, for $g(z) = (A^\top z^y, -Az^x)$, $\mathcal{O}(w_0) = \texttt{InnerLoop}(w_0, \tilde{g}_{w_0}, \alpha)$ is an $(\alpha, 0)$-relaxed proximal oracle.*

Proof. Let $w_1, ..., w_T$ denote the iterates of Algorithm 2 and let $w_{T+1} \equiv u$. We recall the following notation from the proof of Proposition 2: $\tilde{\delta}_t = \tilde{g}_{w_0}(w_t) - g(w_0)$, $\tilde{\Delta}_t = g(w_t) - \tilde{g}_{w_0}(w_t)$ and $s_t = \arg\min_{s \in \mathcal{Z}} \left\{ \langle \eta\tilde{\Delta}_{t-1}, s \rangle + V_{s_{t-1}}(s) \right\}$. Retracing the steps of the proof of Proposition 2 leading up to the bound (23), we observe that by using the first inequality in Lemma 4 rather than the second, the bound (23) becomes

$$\mathbb{E} \max_{u \in \mathcal{Z}} \left\{ \frac{1}{T} \sum_{t \in [T]} \langle g(w_t), w_t - u \rangle - \alpha V_{w_0}(u) \right\} \leq \frac{1}{T} \sum_{t \in [T]} \mathbb{E} \left[-\frac{\alpha}{2} V_{w_0}(w_t) + \langle \tilde{\Delta}_t, w_t - s_t \rangle \right]$$

$$+ \frac{1}{\eta T} \sum_{t \in [T]} \mathbb{E} \left[\left\langle \eta\tilde{\delta}_t, w_t - w_{t+1} \right\rangle - V_{w_t}(w_{t+1}) + \left\langle \eta\tilde{\Delta}_t, s_t - s_{t+1} \right\rangle - V_{s_t}(s_{t+1}) \right]. \quad (32)$$

Let us bound the various expectations in the RHS of (32) one by one. By the first CBB property, $E\left[\tilde{\Delta}_t^x \mid w_t, s_t\right] = 0$ and also $\left\| E\left[\tilde{\Delta}_t^y \mid w_t, s_t\right] \right\|_* \leq \frac{L^2}{\tau} \|w_t - w_0\|^2$. Consequently,

$$\mathbb{E}\langle \tilde{\Delta}_t, w_t - s_t \rangle \leq \frac{L^2}{\tau} \mathbb{E} \|w_t - w_0\|^2 \|w_t^y - s_t^y\|_1.$$

Using $\|y - y'\|_1 \leq 2$ for every $y, y' \in \mathcal{Y} = \Delta^m$ as well as $\tau = \frac{1}{\eta}$, we obtain

$$\mathbb{E}\langle \tilde{\Delta}_t, w_t - s_t \rangle \leq 2\eta L^2 \mathbb{E} \|w_t - w_0\|^2 \leq 4\eta L^2 \mathbb{E} V_{w_0}(w_t). \quad (33)$$

To bound the expectation of $\left\langle \eta\tilde{\delta}_t, w_t - w_{t+1} \right\rangle - V_{w_t}(w_{t+1})$, we write $w_t = (x_t, y_t)$, and note that for the ℓ_2-ℓ_1 setup the Bregman divergence is separable, i.e. $V_{w_t}(w_{t+1}) = V_{x_t}(x_{t+1}) + V_{y_t}(y_{t+1})$. For the $\mathcal{X}$ component, we proceed as in Lemma 4, and write

$$\left\langle \eta\tilde{\delta}_t^x, x_t - x_{t+1} \right\rangle - V_{x_t}(x_{t+1}) \leq \frac{\eta^2}{2} \|\tilde{\delta}_t^x\|_2^2.$$

For the $\mathcal{Y}$ component, we observe that

$$\|\eta\tilde{\delta}_t^y\|_\infty = \eta\|\tilde{g}_{w_0}^y(w_t) - g^y(w_0)\|_\infty \leq \eta\tau = 1$$

by the second CBB property and $\tau = \frac{1}{\eta}$. Therefore, we may apply Lemma 7 with $\delta = -\eta\tilde{\delta}_t^y$ and obtain

$$\left\langle \eta\tilde{\delta}_t^y, y_t - y_{t+1} \right\rangle - V_{y_t}(y_{t+1}) \leq \eta^2 \sum_{i \in [m]} [y_t]_i [\tilde{\delta}_t^y]_i^2.$$

Taking expectation and using the fact that y_t is in the simplex gives

$$\mathbb{E} \left[\left\langle \eta\tilde{\delta}_t^y, y_t - y_{t+1} \right\rangle - V_{y_t}(y_{t+1}) \right] \leq \eta^2 \mathbb{E} \max_{i \in [m]} \mathbb{E} \left[[\tilde{\delta}_t^y]_i^2 \mid w_t \right].$$

The third CBB property reads $\mathbb{E} \left[\|\tilde{\delta}_t^x\|_2^2 \mid w_t \right] + \max_{i \in [m]} \mathbb{E} \left[[\tilde{\delta}_t^y]_i^2 \mid w_t \right] \leq L^2 \|w_t - w_0\|^2$. Therefore, for $t < T$, the above discussion yields

$$\mathbb{E} \left[\left\langle \eta\tilde{\delta}_t, w_t - w_{t+1} \right\rangle - V_{w_t}(w_{t+1}) \right] \leq \eta^2 \mathbb{E} \left[\frac{1}{2} \|\tilde{\delta}_t^x\|_2^2 + \max_{i \in [m]} \mathbb{E} \left[[\tilde{\delta}_t^y]_i^2 \mid w_t \right] \right]$$

$$\leq \eta^2 L^2 \mathbb{E} \|w_t - w_0\|^2 \leq 2\eta^2 L^2 \mathbb{E} V_{w_0}(w_t). \quad (34)$$

To bound the expectation of $\left\langle \eta\tilde{\Delta}_t, s_t - s_{t+1} \right\rangle - V_{s_t}(s_{t+1})$ we proceed just as we had with $\tilde{\delta}_t$. By the second CBB property,

$$\|\eta\tilde{\Delta}_t^y\|_\infty = \eta \left\| \tilde{g}_{w_0}^y(w_t) - g^y(w_t) \right\|_\infty \leq 2\eta L + \eta\tau = \frac{2\alpha}{20L} + 1 \leq 1.79,$$

21

where we used $\eta = \frac{\alpha}{20L^2}$, $\tau = \frac{1}{\eta}$, and $\alpha \le L$. Therefore, Lemma 7 with $\delta = -\eta\tilde{\Delta}_t^{\mathsf{y}}$ gives

$$\mathbb{E}\left[\langle \eta\tilde{\Delta}_t, s_t - s_{t+1}\rangle - V_{s_t}(s_{t+1})\right] \le \eta^2 \mathbb{E}\sum_{i\in[m]}[s_t^{\mathsf{y}}]_i[\tilde{\Delta}_t^{\mathsf{y}}]_i^2 \le \eta^2\,\mathbb{E}\max_{i\in[m]}\mathbb{E}\left[[\tilde{\Delta}_t^{\mathsf{y}}]_i^2 \mid w_t\right],$$

where in the final transition we used the fact that $\tilde{\Delta}_t$ conditioned on w_t is independent of s_t. Since $\alpha \le L$, we have $\tau = \frac{20L^2}{\alpha} \ge 20L \ge 2\sqrt{2}L$. Therefore, by Lemma 8, $\mathbb{E}\left[\|\tilde{\Delta}_t^{\mathsf{x}}\|_2^2 \mid w_t\right] + \max_{i\in[m]}\mathbb{E}\left[[\tilde{\Delta}_t^{\mathsf{y}}]_i^2 \mid w_t\right] \le 2L^2\|w_t - w_0\|^2$. Substituting back, this gives

$$\mathbb{E}\left[\langle \eta\tilde{\Delta}_t, s_t - s_{t+1}\rangle - V_{s_t}(s_{t+1})\right] \le \eta^2\,\mathbb{E}\left[\tfrac{1}{2}\|\tilde{\Delta}_t^{\mathsf{x}}\|_2^2 + \max_{i\in[m]}\mathbb{E}\left[[\tilde{\Delta}_t^{\mathsf{y}}]_i^2 \mid w_t\right]\right]$$
$$\le 2\eta^2 L^2\mathbb{E}\|w_t - w_0\|^2 \le 4\eta^2 L^2\mathbb{E}V_{w_0}(w_t). \qquad (35)$$

Substituting (33), (34) and (35) back into (32), we have

$$\mathbb{E}\max_{u\in\mathcal{Z}}\left\{\frac{1}{T}\sum_{t\in[T]}\langle g(w_t), w_t - u\rangle - \alpha V_{w_0}(u)\right\} \le \frac{1}{T}\sum_{t\in[T]}\left[10\eta L^2 - \tfrac{\alpha}{2}\right]\mathbb{E}V_{w_0}(w_t) = 0$$

where the last transition follows from $\eta = \frac{\alpha}{20L^2}$; this establishes the bound (11) for the iterates of Algorithm 2 with a CBB gradient estimators. By the argument in the proof of Corollary 1, for $g(z) = (A^\top z^{\mathsf{y}}, -Az^{\mathsf{x}})$, the average of those iterates constitutes an $(\alpha, 0)$-relaxed proximal oracle. $\qquad\square$

D.5 Complete pseudo-code

Algorithm 4: Variance reduction for ℓ_2-ℓ_1 games

Input: Matrix $A \in \mathbb{R}^{m\times n}$ with ith row $A_{i:}$ and jth column $A_{:j}$, target accuracy ϵ
Output: A point with expected duality gap below ϵ

1   $L \leftarrow \|A\|_{2\to\infty}$, $\alpha \leftarrow L\sqrt{\frac{n+m}{\mathrm{nnz}(A)}}$, $K \leftarrow \left\lceil\frac{\log(2m)\alpha}{\epsilon}\right\rceil$, $\eta \leftarrow \frac{\alpha}{20L^2}$, $\tau \leftarrow \frac{1}{\eta}$, $T \leftarrow \left\lceil\frac{4}{\eta\alpha}\right\rceil$, $(x_0, y_0) \leftarrow (\mathbf{0}_n, \frac{1}{m}\mathbf{1}_m)$

2   **for** $k = 1, \ldots, K$ **do**

     ▷ *Relaxed oracle query:*

3      $(x_0, y_0) \leftarrow (z_{k-1}^{\mathsf{x}}, z_{k-1}^{\mathsf{y}})$, $(g_0^{\mathsf{x}}, g_0^{\mathsf{y}}) \leftarrow (A^\top y_0, -Ax_0)$

4   **for** $t = 1, \ldots, T$ **do**

         ▷ *Gradient estimation:*

5   Sample $i \sim p$ where $p_i = \frac{|[y_{t-1}]_i - [y_0]_i|}{\|y_{t-1} - y_0\|_1}$, sample $j \sim q$ where $q_j = \frac{([x_{t-1}]_j - [x_0]_j)^2}{\|x_{t-1} - x_0\|_2^2}$

6   Set $\tilde{g}_{t-1} = g_0 + \left(A_{i:}\frac{[y_{t-1}]_i - [y_0]_i}{p_i}, -\mathsf{T}_\tau\left(A_{:j}\frac{[x_{t-1}]_j - [x_0]_j}{q_j}\right)\right)$

                        ▷ $[\mathsf{T}_\tau(v)]_k := \min\{\tau, \max\{-\tau, [v]_k\}\}$

         ▷ *Mirror descent step:*

7          $x_t \leftarrow \Pi_{\mathcal{X}}\left(\frac{1}{1+\eta\alpha/2}\left(x_{t-1} + \frac{\eta\alpha}{2}x_0 - \eta\tilde{g}_{t-1}^{\mathsf{x}}\right)\right)$ ▷ $\Pi_{\mathcal{X}}(v) = \frac{v}{\max\{1, \|v\|_2\}}$

8          $y_t \leftarrow \Pi_{\mathcal{Y}}\left(\frac{1}{1+\eta\alpha/2}\left(\log y_{t-1} + \frac{\eta\alpha}{2}\log y_0 - \eta\tilde{g}_{t-1}^{\mathsf{y}}\right)\right)$ ▷ $\Pi_{\mathcal{Y}}(v) = \frac{e^v}{\|e^v\|_1}$

9      $z_{k-1/2} \leftarrow \frac{1}{T}\sum_{t=1}^{T}(x_t, y_t)$

     ▷ *Extragradient step:*

10      $z_k^{\mathsf{x}} \leftarrow \Pi_{\mathcal{X}}\left(z_{k-1}^{\mathsf{x}} - \frac{1}{\alpha}A^\top z_{k-1/2}^{\mathsf{y}}\right)$, $z_k^{\mathsf{y}} \leftarrow \Pi_{\mathcal{Y}}\left(\log z_{k-1}^{\mathsf{y}} + \frac{1}{\alpha}Az_{k-1/2}^{\mathsf{x}}\right)$

11 **return** $\frac{1}{K}\sum_{k=1}^{K}z_{k-1/2}$

E The ℓ_2-ℓ_2 setup

Setup. In the ℓ_2-ℓ_2 setup, both $\mathcal{X} = \mathbb{B}^n$ and $\mathcal{Y} = \mathbb{B}^m$ are Euclidean unit balls, the norm over $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ is the Euclidean norm (which is dual to itself), and the distance generating function is $r(z) = \frac{1}{2} \|z\|_2^2$. Under the Euclidean norm, the Lipschitz constant of g is $\|A\|_{2 \to 2}$ (the largest singular value of A), and we also consider the Frobenius norm $\|A\|_{\mathrm{F}} = (\sum_{i,j} A_{ij}^2)^{1/2}$, i.e. the Euclidean norm of the singular values of A.

Remark 4. In the ℓ_2-ℓ_2 setup, problems of the form $\min_{x \in \mathbb{B}^n} \max_{y \in \mathbb{B}^m} y^\top A x$ are trivial, since the saddle point is always the origin. However, as we explain in Section F.2, our results extend to problems of the form $\min_{x \in \mathbb{B}^n} \max_{y \in \mathbb{B}^m} \{y^\top A x + \phi(x) - \psi(y)\}$ for convex functions ϕ, ψ, e.g. $\min_{x \in \mathbb{B}^n} \max_{y \in \mathbb{B}^m} \{y^\top A x + b^\top x + c^\top y\}$, which are nontrivial.

Our centered gradient estimator for the ℓ_2-ℓ_2 setup is of the form (13), where we sample from

$$p_i(w) = \frac{([w^{\mathsf{y}}]_i - [w_0^{\mathsf{y}}]_i)^2}{\|w^{\mathsf{y}} - w_0^{\mathsf{y}}\|_2^2} \quad \text{and} \quad q_j(w) = \frac{([w^{\mathsf{x}}]_j - [w_0^{\mathsf{x}}]_j)^2}{\|w^{\mathsf{x}} - w_0^{\mathsf{x}}\|_2^2}. \tag{36}$$

The resulting gradient estimator has the explicit form

$$\tilde{g}_{w_0}(w) = g(w_0) + \left(A_{i:} \frac{\|w^{\mathsf{y}} - w_0^{\mathsf{y}}\|_2^2}{[w^{\mathsf{y}} - w_0^{\mathsf{y}}]_i}, -A_{:j} \frac{\|w^{\mathsf{x}} - w_0^{\mathsf{x}}\|_2^2}{[w^{\mathsf{x}} - w_0^{\mathsf{x}}]_j} \right). \tag{37}$$

Lemma 9. *In the ℓ_2-ℓ_2 setup, the estimator* (37) *is* (w_0, L)*-centered with* $L = \|A\|_{\mathrm{F}}$.

Proof. Unbiasedness follows from the estimator definition. The second property follows from

$$\mathbb{E} \|\tilde{g}_{w_0}(w) - g(w_0)\|_2^2 = \sum_{i \in [m]} \frac{\|A_{i:}\|_2^2}{p_i} ([w^{\mathsf{y}}]_i - [w_0^{\mathsf{y}}]_i)^2 + \sum_{j \in [n]} \frac{\|A_{:j}\|_2^2}{q_j} ([w^{\mathsf{x}}]_j - [w_0^{\mathsf{x}}]_j)^2$$

$$= \|A\|_{\mathrm{F}}^2 \|w - w_0\|_2^2.$$

$\square$

In Appendix F.3.2 we provide two additional sampling distribution that yield estimators with the same guarantee. We may use these gradient estimator to build an algorithm with a convergence guarantee similar to Theorem 2, except with $\|A\|_{\mathrm{F}}$ instead of $\|A\|_{2 \to \infty}$ and 1 instead of $\log(2m)$. This result improves the runtime of Balamurugan and Bach [3] by a $\log(1/\epsilon)$ factor. However, as we discuss in Section 1.3, unlike our ℓ_1-ℓ_1 and ℓ_2-ℓ_1 results, it is not a strict improvement over the linear-time mirror-prox method, which in the ℓ_2-ℓ_2 setting achieves running time $O(\|A\|_{2 \to 2} \operatorname{nnz}(A) \epsilon^{-1})$. The regime in which our variance-reduced method has a stronger guarantee than mirror-prox is

$$\operatorname{srank}(A) := \frac{\|A\|_{\mathrm{F}}^2}{\|A\|_{2 \to 2}^2} \ll \frac{\operatorname{nnz}(A)}{n + m},$$

i.e. when the spectral sparsity of A is significantly greater than its spatial sparsity.

We remark that ℓ_2-ℓ_2 games are closely related to linear regression, as

$$\min_{x \in \mathbb{B}^n} \|A x - b\|_2^2 = \left(\min_{x \in \mathbb{B}^n} \max_{y \in \mathbb{B}^m} \{y^\top A x - y^\top b\} \right)^2.$$

The smoothness of the objective $\|A x - b\|_2^2$ is $\|A\|_{2 \to 2}^2$, but runtimes of stochastic linear regression solvers typically depend on $\|A\|_{\mathrm{F}}^2$ instead [41, 19, 35, 13, 22, 36, 34, 1]. Viewed in this context, it is not surprising that our ℓ_2-ℓ_2 runtime scales as it does.

F Extensions

In this section we collect a number of results that extend our framework and its applications. In Appendix F.1 we show how to use variance reduction to solve the proximal subproblem to high accuracy. This allows us to implement a relaxed gradient oracle for any monotone operator that admits an appropriate gradient estimator, overcoming a technical limitation in the analysis of Algorithm 2 (see discussion following Corollary 1). In Section F.2 we explain how to extend our results to composite saddle point problems of the form $\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \{f(x, y) + \phi(x) - \psi(y)\}$, where f admits a centered gradient estimator and ϕ, ψ are convex functions. Finally, in Section F.3 we return to the bilinear case and provide a number of alternative gradient estimators for the ℓ_2-ℓ_1 and ℓ_2-ℓ_2 settings.

F.1 High precision proximal mappings via variance reduction

Here we describe how to use gradient estimators that satisfy Definition 2 to obtain high precision approximations to the exact proximal mapping, as well as a relaxed proximal oracle valid beyond the bilinear case. Algorithm 5 is a modification of Algorithm 2, where we restart the mirror-descent iteration N times, with each restarting constituting a *phase*. In each phase, we re-center the gradient estimator g, but regularize towards the original initial point w_0. To analyze the performance of the algorithm, we require two properties of proximal mappings with general Bregman divergences (10).

Lemma 10. *Let g by a monotone operator, let $z \in \mathcal{Z}$ and let $\alpha > 0$. Then, for every $w \in \mathcal{Z}$, $z_\alpha = \mathrm{Prox}_z^\alpha(g)$ satisfies*

$$\langle g(w) + \alpha \nabla V_z(w), w - z_\alpha \rangle \geq \alpha V_{z_\alpha}(w) + \alpha V_w(z_\alpha).$$

Proof. By definition of z_α, $\langle g(z_\alpha) + \alpha \nabla V_z(z_\alpha), z_\alpha - w \rangle \leq 0$ for all $w \in \mathcal{Z}$. Therefore

$$
\begin{aligned}
\langle g(w) + \alpha \nabla V_z(w), w - z_\alpha \rangle &\geq \langle g(w) + \alpha \nabla V_z(w), w - z_\alpha \rangle + \langle g(z_\alpha) + \alpha \nabla V_z(z_\alpha), z_\alpha - w \rangle \\
&= \langle g(w) - g(z_\alpha), w - z_\alpha \rangle + \alpha \langle \nabla V_z(w) - \nabla V_z(z_\alpha), w - z_\alpha \rangle \\
&\underset{(i)}{\geq} \alpha \langle \nabla V_z(w) - \nabla V_z(z_\alpha), w - z_\alpha \rangle \underset{(ii)}{=} \alpha V_{z_\alpha}(w) + \alpha V_w(z_\alpha),
\end{aligned}
$$

where (i) follows from monotonicity of g and (ii) holds by definition of the Bregman divergence. $\square$

Lemma 11. *Let g be a monotone operator and let $\alpha > 0$. Then, for every $z \in \mathcal{Z}$, $z_\alpha = \mathrm{Prox}_z^\alpha(g)$ satisfies*

$$V_{z_\alpha}(z) + V_z(z_\alpha) \leq \frac{\|g(z)\|_* \|z - z_\alpha\|}{\alpha} \leq \frac{\|g(z)\|_*^2}{\alpha^2}.$$

Proof. Using Lemma 10 with $w = z$ gives

$$\alpha V_{z_\alpha}(z) + \alpha V_z(z_\alpha) \leq \langle g(z) + \alpha \nabla V_z(z), z - z_\alpha \rangle \leq \langle g(z), z - z_\alpha \rangle,$$

where we used the fact that z minimizes the convex function $V_z(\cdot)$ and therefore $\langle \nabla V_z(z), z - u \rangle \leq 0$ for all $u \in \mathcal{Z}$. Writing $\langle g(z), z - z_\alpha \rangle \leq \|g(z)\|_* \|z - z_\alpha\|$ gives the first bound in the lemma. Next, strong convexity of r implies

$$\|z - z_\alpha\|^2 \leq V_{z_\alpha}(z) + V_z(z_\alpha) \leq \frac{\|g(z)\|_* \|z - z_\alpha\|}{\alpha},$$

and the second bound follows from dividing by $\|z - z_\alpha\|$. $\square$

We now state the main convergence result for Algorithm 5.

Proposition 4. *Let $\alpha, L > 0$, let $w_0 \in \mathcal{Z}$, let $\tilde{g}_z$ be (z, L)-centered for monotone g and every $z \in \mathcal{Z}$ and let $z_\alpha = \mathrm{Prox}_{w_0}^\alpha(g)$. Then, for $\eta = \frac{\alpha}{8L^2}$, $T \geq \frac{4}{\eta \alpha} = \frac{32L^2}{\alpha^2}$, and any $N \in \mathbb{N}$ the output $\hat{w}_N$ of Algorithm 5 satisfies*

$$\mathbb{E} V_{\hat{w}_N}(z_\alpha) \leq 2^{-N} V_{w_0}(z_\alpha). \tag{38}$$

Algorithm 5: `RestartedInnerLoop`$(w_0, z \mapsto \tilde{g}_z, \alpha)$

Input: Initial $w_0 \in \mathcal{Z}$, centered gradient estimator $\tilde{g}_z \; \forall z \in \mathcal{Z}$, oracle quality $\alpha > 0$
Parameters: Step size η, inner iteration count T, phase count N
Output: Point $\hat{w}_N$ satisfying $\mathbb{E} V_{\hat{w}_N}(z_\alpha) \leq 2^{-N} V_{w_0}(z_\alpha)$ where $z_\alpha = \text{Prox}_{w_0}^\alpha(g)$ (for appropriate $\tilde{g}$, η, T)

1 Set $\hat{w}_0 \leftarrow w_0$
2 **for** $n = 1, \ldots, N$ **do**
3 $\quad$ Prepare centered gradient estimator $\tilde{g}_{\hat{w}_{n-1}}$ $\qquad\qquad$ ▷ e.g. by computing $g(\hat{w}_{n-1})$
4 $\quad$ Draw $\hat{T}$ uniformly from $[T]$
5 $\quad$ $w_0^{(n)} \leftarrow \hat{w}_{n-1}$
6 $\quad$ **for** $t = 1, \ldots, \hat{T}$ **do**
7 $\quad\quad$ $w_t^{(n)} \leftarrow \arg\min_{w \in \mathcal{Z}} \left\{ \left\langle \tilde{g}_{\hat{w}_{n-1}}\big(w_{t-1}^{(n)}\big), w \right\rangle + \alpha V_{w_0}(w) + \frac{1}{\eta} V_{w_{t-1}^{(n)}}(w) \right\}$
8 $\quad$ $\hat{w}_n \leftarrow w_{\hat{T}}^{(n)}$
9 **return** $\hat{w}_N$

Proof. Fix a phase $n \in [N]$. For every $u \in \mathcal{Z}$ we have the mirror descent regret bound

$$\sum_{t \in [T]} \left\langle \tilde{g}_{\hat{w}_{n-1}}\big(w_t^{(n)}\big) + \alpha \nabla V_{w_0}\big(w_t^{(n)}\big), w_t^{(n)} - u \right\rangle \leq \frac{V_{\hat{w}_{n-1}}(u)}{\eta} + \frac{\eta}{2} \sum_{t \in [T]} \left\| \tilde{g}_{\hat{w}_{n-1}}\big(w_t^{(n)}\big) - g(\hat{w}_{n-1}) \right\|_*^2 ;$$

see Lemma 4 in Appendix A.2, with $Q(z) = \eta \langle g(\hat{w}_{n-1}), z \rangle + \eta \alpha V_{w_0}(z)$. Choosing $u = z_\alpha$, taking expectation and using Definition 2 gives

$$\mathbb{E} \sum_{t \in [T]} \left\langle g\big(w_t^{(n)}\big) + \alpha \nabla V_{w_0}\big(w_t^{(n)}\big), w_t^{(n)} - z_\alpha \right\rangle \leq \frac{\mathbb{E} V_{\hat{w}_{n-1}}(z_\alpha)}{\eta} + \frac{\eta L^2}{2} \sum_{t \in [T]} \mathbb{E} \left\| w_t^{(n)} - \hat{w}_{n-1} \right\|^2. \tag{39}$$

(Note that z_α is a function of w_0 and hence independent of stochastic gradient estimates.) By the triangle inequality and strong convexity of r,

$$\|w_t^{(n)} - \hat{w}_{n-1}\|^2 \leq 2\|z_\alpha - \hat{w}_{n-1}\|^2 + 2\|w_t^{(n)} - z_\alpha\|^2 \leq 4 V_{\hat{w}_{n-1}}(z_\alpha) + 4 V_{z_\alpha}\big(w_t^{(n)}\big). \tag{40}$$

By Lemma 10 we have that for every $t \in [T]$

$$\left\langle g\big(w_t^{(n)}\big) + \alpha \nabla V_{w_0}\big(w_t^{(n)}\big), w_t^{(n)} - z_\alpha \right\rangle \geq \alpha V_{w_t^{(n)}}(z_\alpha) + \alpha V_{z_\alpha}\big(w_t^{(n)}\big). \tag{41}$$

Substituting the bounds (40) and (41) into the expected regret bound (39) and rearranging gives

$$\frac{1}{T} \sum_{t \in [T]} \mathbb{E} V_{w_t^{(n)}}(z_\alpha) \leq \left(\frac{1}{\eta \alpha T} + \frac{2\eta L^2}{\alpha} \right) \mathbb{E} V_{\hat{w}_{n-1}}(z_\alpha) + \frac{2\eta L^2 - \alpha}{\alpha T} \sum_{t \in [T]} \mathbb{E} V_{z_\alpha}\big(w_t^{(n)}\big) \leq \frac{1}{2} \mathbb{E} V_{w_t^{(n-1)}}(z_\alpha),$$

where in the last transition we substituted $\eta = \frac{\alpha}{8L^2}$ and $T \geq \frac{4}{\eta \alpha}$. Noting that $\frac{1}{T} \sum_{t \in [T]} \mathbb{E} V_{w_t^{(n)}}(z_\alpha) = \mathbb{E} V_{\hat{w}_n}(z_\alpha)$ and recursing on n completes the proof. $\qquad \square$

The linear convergence bound (38) combined with Lemma 11 implies that Algorithm 5 implements a relaxed proximal oracle.

Corollary 2. *Let $G, D > 0$ be such that $\|g(z)\|_* \leq G$ and $\|z - z'\| \leq D$ for every $z, z' \in \mathcal{Z}$ and let $\varepsilon > 0$. Then, in the setting of Proposition 4 with $N \geq 1 + 2\log_2\left(\frac{G(G + 2LD)}{\alpha \varepsilon}\right)$, we have that $\mathcal{O}(w_0) = \texttt{RestartedInnerLoop}(w_0, \tilde{g}, \alpha)$ is an (α, ε)-relaxed proximal oracle.*

Proof. Let $\hat{w} = \texttt{RestartedInnerLoop}(w_0, \tilde{g}, \alpha)$ and let $z_\alpha = \text{Prox}_{w_0}^\alpha(g)$. For every $u \in \mathcal{Z}$, we have

$$\langle g(\hat{w}), \hat{w} - u \rangle = \langle g(z_\alpha), z_\alpha - u \rangle + \langle g(z_\alpha), \hat{w} - z_\alpha \rangle + \langle g(\hat{w}) - g(z_\alpha), \hat{w} - u \rangle.$$

25

By the definition (10) of z_α we have $\langle g(z_\alpha), z_\alpha - u \rangle \leq \alpha V_{w_0}(u)$. By Hölder's inequality and the assumption that g is bounded, we have $\langle g(z_\alpha), \hat{w} - z_\alpha \rangle \leq G \| \hat{w} - z_\alpha \|$. Finally, since g is $2L$-Lipschitz (see Remark 1) and $\| \hat{w} - u \| \leq D$ by assumption, we have $\langle g(\hat{w}) - g(z_\alpha), \hat{w} - u \rangle \leq 2LD \| \hat{w} - z_\alpha \|$. Substituting back these three bounds and rearranging yields

$$\langle g(\hat{w}), \hat{w} - u \rangle - \alpha V_{w_0}(u) \leq (G + 2LD) \| \hat{w} - z_\alpha \| \leq (G + 2LD) \sqrt{2 V_{\hat{w}}(z_\alpha)},$$

where the last bound is due to strong convexity of r. Maximizing over u and taking expectation, we have by Jensen's inequality and Proposition 4,

$$\mathbb{E} \max_{u \in \mathcal{Z}} \{ \langle g(\hat{w}), \hat{w} - u \rangle - \alpha V_{w_0}(u) \} \leq (G + 2LD) \sqrt{2 \mathbb{E} V_{\hat{w}}(z_\alpha)} \leq 2^{-(N-1)/2} (G + 2LD) \sqrt{V_{w_0}(z_\alpha)}.$$

Lemma 11 gives us $\sqrt{V_{w_0}(z_\alpha)} \leq \sqrt{\|g(w_0)\|_*^2 / \alpha^2} \leq G / \alpha$, and therefore $N \geq 1 + 2 \log_2 \left(\frac{G(G+2LD)}{\alpha \varepsilon} \right)$ establishes the oracle property $\mathbb{E} \max_{u \in \mathcal{Z}} \{ \langle g(\hat{w}), \hat{w} - u \rangle - \alpha V_{w_0}(u) \} \leq \varepsilon$. $\qquad \square$

Remark 5. In the ℓ_2-ℓ_1 setup of Section 4.2, Proposition 4 and Corollary 2 extend straightforwardly to centered-bounded-biased gradient estimators (Definition 3) using arguments from the proof of Proposition 3.

Since Algorithm 5 computes a highly accurate approximation of the proximal mapping, it is reasonable to expect that directly iterating $z_k = \texttt{RestartedInnerLoop}(z_{k-1}, \tilde{g}, \alpha)$ for $k \in [K]$ would yield an $O(\alpha \Theta / K)$ error bound, without requiring the extragradient step in Algorithm 1. However, we could not show such a bound without additionally requiring uniform smoothness of the distance generating function r, which does not hold for the negative entropy we use in the ℓ_1 setting.

F.2 Composite saddle point problems

Consider the "composite" saddle point problem of the form

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \{ f(x, y) + \phi(x) - \psi(y) \},$$

where ∇f admits a centered gradient estimator and ϕ, ψ are "simple" convex functions in the sense they have efficiently-computable proximal mappings. As usual in convex optimization, it is straightforward to extend our framework to this setting. Let $\Upsilon(z) := \phi(z^{\mathsf{x}}) + \psi(z^{\mathsf{y}})$ so that $g(z) + \nabla \Upsilon(z)$ denotes the (sub-)gradient mapping for the composite problem at point z. Algorithmically, the extension consists of changing Line 4 of Algorithm 1 to

$$z_k \leftarrow \arg\min_{z \in \mathcal{Z}} \left\{ \langle g(z_{k-1/2}) + \nabla \Upsilon(z_{k-1/2}), z \rangle + \alpha V_{z_{k-1}}(z) \right\},$$

changing line 2 of Algorithm 2 to

$$w_t \leftarrow \arg\min_{w \in \mathcal{Z}} \left\{ \langle \tilde{g}_{w_0}(w_{t-1}), w \rangle + \Upsilon(w) + \frac{\alpha}{2} V_{w_0}(w) + \frac{1}{\eta} V_{w_{t-1}}(w) \right\},$$

and similarly adding $\Upsilon(w)$ to the minimization in line 7 of Algorithm 5.

Analytically, we replace g with $g + \nabla \Upsilon$ in the duality gap bound (8), Definition 1 (relaxed proximal oracle), and Proposition 1 and its proof, which holds without further change. To implement the composite relaxed proximal oracle we still assume a centered gradient estimator for g only. However, with the algorithmic modifications described above, the guarantee (11) of Proposition 2 now has $g + \nabla \Upsilon$ instead of g; the only change to the proof is that we now invoke Lemma 4 (in Appendix A.2) with the composite term $\eta \left[\langle g(w_0), z \rangle + \Upsilon(z) + \frac{\alpha}{2} V_{w_0}(z) \right]$, and the bound (20) becomes

$$\sum_{t \in [T]} \langle \tilde{g}_{w_0}(w_t) + \nabla \Upsilon(w_t) + \frac{\alpha}{2} \nabla V_{w_0}(w_t), w_t - u \rangle \leq \frac{V_{w_0}(u)}{\eta} + \frac{\eta}{2} \sum_{t \in [T]} \| \tilde{\delta}_t \|_*^2.$$

Proposition 3, Proposition 4 and Corollary 2 similarly extend to the composite setup.

The only point in our development that does not immediately extend to the composite setting is Corollary 1 and its subsequent discussion. There, we argue that Algorithm 2 implements a relaxed proximal oracle only when $\langle g(z), z - u \rangle$ is convex in z for all u, which is the case for bilinear f. However, this condition might fail for $g + \nabla \Upsilon$ even when it holds for g. In this case, we may still use the oracle implementation guaranteed by Corollary 2 for any convex Υ.

F.3 Additional gradient estimators

We revisit the bilinear setting studied in Section 4 and provide additional gradient estimators that meet our variance requirements. In Section F.3.1 we consider ℓ_2-ℓ_1 games and construct an "oblivious" estimator for the $\mathcal{Y}$ component of the gradient that involves sampling from a distribution independent of the query point. In Section F.3.2 we describe two additional centered gradient estimators for ℓ_2-ℓ_2 games; one of them is the "factored splits" estimator proposed in [3].

F.3.1 ℓ_2-ℓ_1 games

Consider the ℓ_2-ℓ_1 setup described in the beginning of Section 4.2. We describe an alternative for the $\mathcal{Y}$ component of (29), that is "oblivious" in the sense that it involves sampling from distributions that do not depend on the current iterate. The estimator generates each coordinate of $\tilde{g}^{\mathsf{y}}_{w_0}$ independently in the following way: for every $i \in [m]$ we define the probability $q^{(i)} \in \Delta^n$ by

$$q^{(i)}_j = A^2_{ij} / \|A_{i:}\|^2_2, \ \ \forall j \in [n].$$

Then, independently for every $i \in [m]$, draw $j(i) \sim q^{(i)}$ and set

$$[\tilde{g}^{\mathsf{y}}_{w_0}(w)]_i = -[Aw^{\mathsf{x}}_0]_i - \mathsf{T}_\tau \left(A_{ij(i)} \frac{[w^{\mathsf{x}}]_{j(i)} - [w^{\mathsf{x}}_0]_{j(i)}}{q^{(i)}_{j(i)}} \right), \tag{42}$$

where T_τ is the clipping operator defined in (29). Note that despite requiring m independent samples from different distributions over n elements, $\tilde{g}^{\mathsf{y}}_{w_0}$ still admits efficient evaluation. This is because the distributions $q^{(i)}$ are fixed in advance, and we can pre-process them to perform each of the m samples in time $O(1)$ [42]. However, the oblivious gradient estimator produces fully dense estimates regardless of the sparsity of A, which limits its running time guarantees to terms proportional to m rather than the maximum number of nonzero elements in columns of A.

The oblivious estimator has the same "centered-bounded-biased" properties (Definition 3) as the "dynamic" estimator (29).

Lemma 12. *In the ℓ_2-ℓ_1 setup, a gradient estimator with $\mathcal{X}$ block as in (29) and $\mathcal{Y}$ block as in (42) is (w_0, L, τ)-CBB with $L = \|A\|_{2 \to \infty}$.*

Proof. We show the bias bound similarly to the proof of Lemma 6,

$$\left| \mathbb{E} \left[\tilde{g}^{\mathsf{y}}_{w_0}(w) - g^{\mathsf{y}}(w) \right]_i \right| \le \sum_{j \in \mathcal{J}_\tau(i)} |A_{ij}| \, |[w^{\mathsf{x}}]_j - [w^{\mathsf{x}}_0]_j|$$

for all $i \in [m]$, where

$$\mathcal{J}_\tau(i) = \left\{ j \in [n] \mid \mathsf{T}_\tau \left(\frac{A_{ij}}{q^{(i)}_j} ([w^{\mathsf{x}}]_j - [w^{\mathsf{x}}_0]_j) \right) \ne \frac{A_{ij}}{q^{(i)}_j} ([w^{\mathsf{x}}]_j - [w^{\mathsf{x}}_0]_j) \right\}.$$

Note that $j \in \mathcal{J}_\tau(i)$ if and only if

$$\left| \frac{A_{ij}}{q^{(i)}_j} ([w^{\mathsf{x}}]_j - [w^{\mathsf{x}}_0]_j) \right| = \frac{\|A_{i:}\|^2_2 \, |[w^{\mathsf{x}}]_j - [w^{\mathsf{x}}_0]_j|}{|A_{ij}|} > \tau \Rightarrow |A_{ij}| \le \frac{1}{\tau} \|A_{i:}\|^2_2 \, |[w^{\mathsf{x}}]_j - [w^{\mathsf{x}}_0]_j|.$$

Therefore,

$$\sum_{j \in \mathcal{J}_\tau(i)} |A_{ij}| \, |[w^{\mathsf{x}}]_j - [w^{\mathsf{x}}_0]_j| \le \frac{1}{\tau} \|A_{i:}\|^2_2 \sum_{j \in \mathcal{J}_\tau} |[w^{\mathsf{x}}]_j - [w^{\mathsf{x}}_0]_j|^2 = \frac{1}{\tau} \|A_{i:}\|^2_2 \, \|w^{\mathsf{x}} - w^{\mathsf{x}}_0\|^2_2$$

and $\left\| \mathbb{E} \tilde{g}^{\mathsf{y}}_{w_0}(w) - g^{\mathsf{y}}(w) \right\|_\infty \le \frac{L^2}{\tau} \|w^{\mathsf{x}} - w^{\mathsf{x}}_0\|^2_2$ follows by taking the maximum over $i \in [m]$.

The second property follows exactly as in the proof of Lemma 6. For the third property, note that the bound (26) on the $\mathcal{X}$ component still holds, and that for each $i \in [m]$ we have $q_j^{(i)} = A_{ij}^2 / \|A_{i:}\|_2^2$ and

$$
\mathbb{E}\left[\tilde{g}_{w_0}^{\mathsf{y}}(w) - g^{\mathsf{y}}(w)\right]_i^2 = \sum_{j \in [n]} q_j^{(i)} \left(\mathsf{T}_\tau \left(\frac{A_{ij}}{q_j^{(i)}} \left([w^{\mathsf{x}}]_j - [w_0^{\mathsf{x}}]_j\right)\right)\right)^2
$$

$$
\leq \sum_{j \in [n]} q_j^{(i)} \left(\frac{A_{ij}}{q_j^{(i)}} \left([w^{\mathsf{x}}]_j - [w_0^{\mathsf{x}}]_j\right)\right)^2 = \|A_{i:}\|_2^2 \|w^{\mathsf{x}} - w_0^{\mathsf{x}}\|_2^2 .
$$

$\square$

F.3.2 ℓ_2-ℓ_2 games

In the ℓ_2-ℓ_2 setup described in Section E it is possible to use a completely oblivious gradient estimator. It has the form (13) with the following sampling distributions that do not depend on w_0, w,

$$
p_i = \frac{\|A_{i:}\|_2^2}{\|A\|_{\mathrm{F}}^2} \quad \text{and} \quad q_j = \frac{\|A_{:j}\|_2^2}{\|A\|_{\mathrm{F}}^2} . \tag{43}
$$

Balamurugan and Bach [3] use these sampling distributions, referring to them as "factored splits." Another option is to use the dynamic sampling probabilities

$$
p_i(w) = \frac{\|A_{i:}\|_2 \, |[w^{\mathsf{y}}]_i - [w_0^{\mathsf{y}}]_i|}{\sum_{i' \in [m]} \|A_{i':}\|_2 \, |[w^{\mathsf{y}}]_{i'} - [w_0^{\mathsf{y}}]_{i'}|} \quad \text{and} \quad q_j(w) = \frac{\|A_{:j}\|_2 \, |[w^{\mathsf{x}}]_j - [w_0^{\mathsf{x}}]_j|}{\sum_{j' \in [n]} \|A_{:j'}\|_2 \, |[w^{\mathsf{x}}]_{j'} - [w_0^{\mathsf{x}}]_{j'}|} . \tag{44}
$$

Both the distributions above yield centered gradient estimators.

Lemma 13. *In the ℓ_2-ℓ_2 setup, the estimator* (13) *with either sampling probabilities* (43) *or* (44) *is* (w_0, L)*-centered for* $L = \|A\|_{\mathrm{F}}$.

Proof. Unbiasedness follows from the estimator definition. For the oblivious sampling strategy (43) the second property follows from

$$
\mathbb{E} \|\tilde{g}_{w_0}(w) - g(w_0)\|_2^2 = \sum_{i \in [m]} \frac{\|A_{i:}\|_2^2}{p_i} ([w^{\mathsf{y}}]_i - [w_0^{\mathsf{y}}]_i)^2 + \sum_{j \in [n]} \frac{\|A_{:j}\|_2^2}{q_j} ([w^{\mathsf{x}}]_j - [w_0^{\mathsf{x}}]_j)^2
$$

$$
= \|A\|_{\mathrm{F}}^2 \|w - w_0\|_2^2 .
$$

For the dynamic sampling strategy (44), we have

$$
\mathbb{E} \|\tilde{g}_{w_0}(w) - g(w_0)\|_2^2 = \left(\sum_{i' \in [m]} \|A_{i':}\|_2 \, |[w^{\mathsf{y}}]_{i'} - [w_0^{\mathsf{y}}]_{i'}|\right)^2 + \left(\sum_{j' \in [n]} \|A_{:j'}\|_2 \, |[w^{\mathsf{x}}]_{j'} - [w_0^{\mathsf{x}}]_{j'}|\right)^2
$$

$$
\leq \|A\|_{\mathrm{F}}^2 \|w - w_0\|_2^2 ,
$$

where the inequality is due to Cauchy–Schwarz. $\square$

We remark that out of the three sampling strategies (36), (43) and (44), only for (44) the bound $\mathbb{E} \|\tilde{g}_{w_0}(w) - g(w_0)\|_2^2 \leq \|A\|_{\mathrm{F}}^2 \|w - w_0\|_2^2$ is an inequality, whereas for the other two it holds with equality. Consequently, the dynamic sampling probabilities (44) might be preferable in certain cases.