Relative Lipschitzness in Extragradient Methods and a Direct Recipe for Acceleration

Michael B. Cohen
Massachusetts Institute of Technoolgy, Cambridge, MA, USA
micohen@mit.edu

Aaron Sidford
Stanford University, CA, USA
sidford@stanford.edu

Kevin Tian
Stanford University, CA, USA
kjtian@stanford.edu

Abstract

We show that standard extragradient methods (i.e. mirror prox [26] and dual extrapolation [28]) recover optimal accelerated rates for first-order minimization of smooth convex functions. To obtain this result we provide fine-grained characterization of the convergence rates of extragradient methods for solving monotone variational inequalities in terms of a natural condition we call *relative Lipschitzness*. We further generalize this framework to handle local and randomized notions of relative Lipschitzness and thereby recover rates for box-constrained ℓ_∞ regression based on area convexity [34] and complexity bounds achieved by accelerated (randomized) coordinate descent [5, 29] for smooth convex function minimization.

2012 ACM Subject Classification Mathematics of computing → Convex optimization

Keywords and phrases Variational inequalities, minimax optimization, acceleration, ℓ_∞ regression

Digital Object Identifier 10.4230/LIPIcs.ITCS.2021.62

Related Version A full version of the paper is available at `https://arxiv.org/abs/2011.06572`.

Funding Researchers supported in part by Microsoft Research Faculty Fellowship, NSF CAREER Award CCF-1844855, NSF Grant CCF-1955039, a PayPal research gift, and a Sloan Research Fellowship.

Acknowledgements The existence of an extragradient algorithm in the primal-dual formulation of smooth minimization directly achieving accelerated rates is due to discussions with the first author, Michael B. Cohen. The second and third authors are indebted to him, and this work is dedicated in his memory. We also thank our collaborators in concurrent works, Yair Carmon and Yujia Jin, for many helpful and encouraging conversations throughout the duration of this project, Jelena Diakonikolas, Jonathan Kelner, and Jonah Sherman for helpful conversations, and anonymous reviewers for multiple helpful comments on earlier versions of the paper.

1 Introduction

We study the classic extragradient algorithms of mirror prox [26] and dual extrapolation [28] for solving variational inequalities (VIs) in monotone operators. This family of problems includes convex optimization and finding the saddle point of a convex-concave game. Due to applications of the latter to adversarial and robust training, extragradient methods have received significant recent attention in the machine learning community, see e.g. [12, 24, 16]. Further, extragradient methods have been the subject of increasing study by the theoretical computer science and optimization communities due to recent extragradient-based runtime improvements for problems including maximum flow [34] and zero-sum games [10, 11].

12th Innovations in Theoretical Computer Science Conference (ITCS 2021).
Editor: James R. Lee; Article No. 62; pp. 62:1–62:18
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Given a Lipschitz monotone operator and a bounded strongly-convex regularizer, mirror prox [26] and dual extrapolation [28] achieve $O(T^{-1})$ regret for solving the associated VI after T iterations. This rate is worst-case optimal when the Lipschitzness of the operator and strong convexity of the regularizer are with respect to the Euclidean norm [30]. However, in certain structured problems related to VIs, alternative analyses and algorithms can yield improved rates. For instance, when minimizing a smooth convex function (i.e. one with a Lipschitz gradient), it is known that accelerated rates of $O(T^{-2})$ are attainable, improving upon the standard $O(T^{-1})$ extragradient rate for the naive associated VI. Further, algorithms inspired by extragradient methods have been developed recovering the $O(T^{-2})$ rate [13, 37].

Additionally, alternative analyses of extragradient methods, such as optimism [32] and area convexity [34] have shown that under assumptions beyond a Lipschitz operator and a strongly convex regularizer, improved rates can be achieved. These works leveraged modified algorithms which run efficiently under such non-standard assumptions. Further, the area convexity-based methods of [34] have had a number of implications, including faster algorithms for ℓ_∞ regression, maximum flow and multicommodity flow [34] as well as improved parallel algorithms for work-efficient positive linear programming [8] and optimal transport [17].

In this work we seek a better understanding of these structured problems and the somewhat disparate-seeming analyses and algorithms for solving them. We ask, *are the algorithmic changes enabling these results necessary? Can standard mirror prox and dual extrapolation be leveraged to obtain these results? Can we unify analyses for these problems, and further clarify the relationship between acceleration, extragradient methods, and primal-dual methods?*

Towards addressing these questions, we provide a general condition, which we term *relative Lipschitzness* (cf. Definition 1), and analyze the convergence of mirror prox and dual extrapolation for a monotone relatively Lipschitz operator.[1] This condition is derived directly from the standard analysis of the methods and is stated in terms of a straightforward relationship between the operator g and the regularizer r which define the algorithm. Our condition is inspired by both area convexity as well as the "relative smoothness" condition in convex optimization [6, 23], and can be thought of as a generalization of the latter to variational inequalities (see Lemma 3). Further, through this analysis we show that standard extragradient methods directly yield accelerated rates for smooth minimization and recover the improved rates of [34] for box-constrained ℓ_∞ regression, making progress on the questions outlined above. We also show our methods recover certain randomized accelerated rates and have additional implications, outlined below.

Extragradient methods directly yield acceleration. In Section 4, we show that applying algorithms of [26, 28] to a minimax formulation of $\min_{x\in\mathbb{R}^d} f(x)$, when f is smooth and strongly convex, yields accelerated rates when analyzed via relative Lipschitzness. Specifically, we consider the following problem, termed the *Fenchel game* in [37]:

$$\min_{x\in\mathbb{R}^d} \max_{y\in\mathbb{R}^d} \langle y, x\rangle - f^*(y), \tag{1}$$

[1] A somewhat similarly-named property appeared in [22], which also studied mirror descent algorithms under relaxed conditions; their property $\|g(x)\|_*^2 \le \frac{MV_x(y)}{\|y-x\|^2}$ for all x, y, is different than the one we propose. Further, during the preparation of this work, the relative Lipschitzness condition we propose was also independently stated in [36] (unbeknownst to the authors of this paper until recently). However, the work [36] does not derive the various consequences of relative Lipschitzness contained in this work (e.g. recovery of acceleration and randomized acceleration, as well as applications of area convexity).

and show that when f is μ-strongly convex and L-smooth, $O(\sqrt{L/\mu})$ iterations of either mirror prox [26] or dual extrapolation [28] produces an average iterate which halves the function error of f. By repeated application, this yields an accelerated linear rate of convergence and the optimal $O(T^{-2})$ rates for non-strongly convex, smooth function minimization by a reduction [4]. Crucially, to attain this rate we give a sharpened bound on the relative Lipschitzness of the gradient operator of (1) with respect to a primal-dual regularizer.

Our result advances a recent line of research, [1, 2, 13], on applying primal-dual analyses to shed light on the mysterious nature of acceleration. Specifically, [1, 2] show that the classical algorithm of [27] can be rederived via applying primal-dual "optimistic" dynamics, inspired by the framework of [32]. Further, [13] showed that an appropriate discretization of dynamics inspired by extragradient algorithms yields an alternative accelerated algorithm. While these results clarify the primal-dual nature of acceleration, additional tuning is ultimately required to obtain their final algorithms and analysis. We obtain acceleration as a direct application of known frameworks, i.e. standard mirror prox and dual extrapolation, applied to the formulation (1), and hope this helps demystify acceleration.

In the full version of the paper, we further show that analyzing extragradient methods tailored to strongly monotone operators via relative Lipschitzness, and applying this more fine-grained analysis to a variant of the objective (1), also yields an accelerated linear rate of convergence. The resulting proof strategy extends readily to accelerated minimization of smooth and strongly convex functions in general norms, as we discuss at the end of Section 4, and we believe it may be of independent interest.

Finally, we remark that there has been documented difficulty in accelerating the minimization of relatively smooth functions [15]; this was also explored more formally by [14]. It is noted in [15], as well as suggested in others (e.g. in the development of area convexity [34]) that this discrepancy may be due to acceleration fundamentally requiring conditions on relationships between groups of three points, rather than two. Our work, which presents an alternative three-point condition yielding accelerated rates, sheds light on this phenomenon and we believe it is an interesting future direction to explore the relationships between our condition and other alternatives in the literature which are known to yield acceleration.

Area convexity for bilinear box-simplex games. In Section 5, we draw a connection between relative Lipschitzness and the notion of an "area convex" regularizer, proposed by [34]. Area convexity is a property which weakens strong convexity, but is suitable for extragradient algorithms with a linear operator. It was introduced in the context of solving a formulation of approximate undirected maximum flow via box-constrained ℓ_∞ regression, or more generally approximating bilinear games between a box variable and a simplex variable. The algorithm of [34] applied to bilinear games was a variant of standard extragradient methods and analyzed via area convexity, which was proven via solving a subharmonic partial differential equation. We show that mirror prox, as analyzed by a local variant of relative Lipschitzness, yields the same rate of convergence as implied by area convexity, for box-simplex games. Our proof of this rate is straightforward and based on a simple Cauchy-Schwarz argument after demonstrating local stability of iterates.

Randomized extragradient methods via local variance reduction. In general, the use of stochastic operator estimates in the design of extragradient algorithms for solving general VIs is not as well-understood as their use in the special case of convex function minimization. The best-known known stochastic methods for solving VIs [19] with bounded-variance stochastic estimators obtain $O(T^{-1/2})$ rates of convergence; this is by necessity, from known classical lower bounds on the rate of the special case of stochastic convex optimization [25]. Towards advancing the randomized extragradient toolkit, we ask: when can improved $O(T^{-1})$ rates of convergence be achieved by stochastic algorithms for solving specific VIs and fine-grained bounds on estimator variance (i.e. more local notions of variance)? This direction is inspired by analogous results in convex optimization, where reduced-variance and accelerated rates have been obtained, matching and improving upon their deterministic counterparts [18, 33, 3, 20, 29, 5].

For the special case of bilinear games, this question was recently addressed by the works [31, 10], using proximal reductions to attain improved rates. In this work, we give a framework for direct stochastic extragradient method design bypassing the variance bottleneck limiting prior algorithms to a $O(T^{-1/2})$ rate of convergence for problems with block-separable structure. We identify a particular criterion of randomized operators used in the context of extragradient algorithms (cf. Proposition 12) which enables $O(T^{-1})$ rates of convergence. Our approach is a form of "local variance reduction", where estimators in an iteration of the method share a random seed and we take expectations over the whole iteration in the analysis. Our improved estimator design exploits the separable structure of the problem; it would be interesting to design a more general variance reduction framework for randomized extragradient methods.

Formally, we apply our local variance reduction framework in Section 6 to show that an instance of our new randomized extragradient methods recover acceleration for coordinate-smooth functions, matching the known tight rates of [5, 29]. Along the way, we give a variation of relative Lipschitzness capturing an analagous property between a locally variance-reduced randomized gradient estimator and a regularizer, which we exploit to obtain our runtime. We note that a similar approach was taken in [35] to obtain faster approximate maximum flow algorithms in the bilinear minimax setting; here, we generalize this strategy and give conditions under which our variance reduction technique obtains improved rates for extragradient methods more broadly.

Additional contributions. A minor contribution of our framework is that we show, in the full version of the paper, that relative Lipschitzness implies new rates for minimax convex-concave optimization, taking a step towards closing the gap with lower bounds with *fine-grained* dependence on problem parameters. Under operator-norm bounds on blocks of the Hessian of a convex-concave function, as well as blockwise strong convexity assumptions, [38] showed a lower bound on the convergence rate to obtain an ϵ-approximate saddle point. When the blockwise operator norms of the Hessian are roughly equal, [21] gave an algorithm matching the lower bound up to a polylogarithmic factor, using an alternating scheme repeatedly calling an accelerated proximal point reduction. Applying our condition with a strongly monotone variant of the mirror prox algorithm of [26] yields a new fine-grained rate for minimax optimization, improving upon the runtime of [21] for a range of parameters. Our algorithm is simple and the analysis follows directly from a tighter relative Lipschitzness bound; we note the same result can also be obtained by considering an operator norm bound of the problem after a rescaling of space, but we include this computation because it is a straightforward implication of our condition.

Finally, in the full version, we also discuss the relation of relative Lipschitzness to another framework for analyzing extragradient methods: namely, we note that our proof of the sufficiency of relative Lipschitzness recovers known bounds for optimistic mirror descent [32].

2 Notation

General notation. Variables are in $\mathbb{R}^d$ unless otherwise noted. e_i is the i^{th} standard basis vector. $\|\cdot\|$ denotes an arbitrary norm; the dual norm is $\|\cdot\|_*$, defined as $\|x\|_* := \max_{\|y\| \le 1} y^\top x$. For a variable on two blocks $z \in \mathcal{X} \times \mathcal{Y}$, we refer to the blocks by z^x and z^y. We denote the domain of $f : \mathbb{R}^d \to \mathbb{R}$ by $\mathcal{X}$; when unspecified, $\mathcal{X} = \mathbb{R}^d$. When f is clear from context, x^* is any minimizing argument. We call any x with $f(x) \le f(x^*) + \epsilon$ an *ϵ-approximate minimizer.*

Bregman divergences. The Bregman divergence induced by convex r is

$$V_x^r(y) := r(y) - r(x) - \langle \nabla r(x), y - x \rangle .$$

The Bregman divergence is always nonnegative, and convex in its argument. We define the following proximal operation with respect to a divergence from point z.

$$\text{Prox}_x^r(g) := \text{argmin}_y \left\{ \langle g, y \rangle + V_x^r(y) \right\} . \tag{2}$$

Functions. We say f is L-smooth in $\|\cdot\|$ if $\|\nabla f(x) - \nabla f(y)\|_* \le L \|x - y\|$, or equivalently $f(y) \le f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2$ for $x, y \in \mathcal{X}$. If f is twice-differentiable, equivalently $y^\top \nabla^2 f(x) y \le L \|y\|^2$. We say differentiable f is μ-strongly convex if for some $\mu \ge 0$, $f(y) \ge f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2$ for $x, y \in \mathcal{X}$. We also say f is μ-strongly convex with respect to a distance-generating function r if $V_x^f(y) \ge \mu V_x^r(y)$ for all $x, y \in \mathcal{X}$. Further, we use standard results from convex analysis throughout, in particular facts about Fenchel duality, and defer these definitions and proofs to the full version of the paper.

Saddle points. We call function $h(x, y)$ of two variables *convex-concave* if its restrictions to x and y are convex and concave respectively. We call (x, y) an *ϵ-approximate saddle point* if $\max_{y'} \{h(x, y')\} - \min_{x'} \{h(x', y)\} \le \epsilon$. We equip any differentiable convex-concave function with gradient operator $g(x, y) := (\nabla_x h(x, y), -\nabla_y h(x, y))$.

Monotone operators. We call operator $g : \mathcal{Z} \to \mathcal{Z}^*$ monotone if $\langle g(w) - g(z), w - z \rangle \ge 0$ for all w, $z \in \mathcal{Z}$. Examples include the gradient of a convex function and gradient operator of a convex-concave function. We call g m-strongly monotone with respect to r if $\langle g(w) - g(z), w - z \rangle \ge m \left(V_w^r(z) + V_z^r(w) \right)$. We call $z^* \in \mathcal{Z}$ the solution to the variational inequality (VI) in a monotone operator g if $\langle g(z^*), z^* - z \rangle \le 0$ for all $z \in \mathcal{Z}$.[2] Examples include the minimizer of a convex function and the saddle point of a convex-concave function.

3 Extragradient convergence under relative Lipschitzness

We give a brief presentation of mirror prox [26], and a convergence analysis under relative Lipschitzness. Our results also hold for dual extrapolation [28], which can be seen as a “lazy” version of mirror prox updating a state in dual space (see [9]); we defer details to the full version of the paper.

▶ **Definition 1** (Relative Lipschitzness). *For convex $r : \mathcal{Z} \to \mathbb{R}$, we call operator $g : \mathcal{Z} \to \mathcal{Z}^*$* λ-relatively Lipschitz with respect to r *if for every three $z, w, u \in \mathcal{Z}$,*

$$\langle g(w) - g(z), w - u \rangle \le \lambda \left(V_z^r(w) + V_w^r(u) \right)$$

[2] This is also known as a “strong solution”. A “weak solution” is a z^* with $\langle g(z), z^* - z \rangle \le 0$ for all z.

Algorithm 1 Mirror-Prox(z_0, T): Mirror prox [26].

```
Input: Distance generating $r$, $\lambda$-relatively Lipschitz monotone $g : \mathcal{Z} \to \mathcal{Z}^*$, initial point $z_0 \in \mathcal{Z}$
for $0 \le t < T$ do
    $w_t \leftarrow \mathrm{Prox}^r_{z_t}(\frac{1}{\lambda} g(z_t))$
    $z_{t+1} \leftarrow \mathrm{Prox}^r_{z_t}(\frac{1}{\lambda} g(w_t))$
end for
```

Definition 1 can be thought of as an alternative to a natural nonlinear analog of the area convexity condition of [34] displayed below:

$$\langle g(w) - g(z), w - u \rangle \le \lambda \left(r(z) + r(w) - r(u) - 3r\left(\frac{z+w+u}{3}\right)\right).$$

Our proposed alternative is well-suited for the standard analyses of extragradient methods such as mirror prox and dual extrapolation. For the special case of bilinear minimax problems in a matrix A, the left hand side of Definition 1 measures the area of a triangle in a geometry induced by A.

Relative Lipschitzness encapsulates the more standard assumptions that g is Lipschitz and r is strongly convex (Lemma 2), as well as the more recent assumptions that f is convex and relatively smooth with respect to r [6, 23] (Lemma 3).

▶ **Lemma 2.** *If g is L-Lipschitz and r is μ-strongly convex in $\|\cdot\|$, g is L/μ-relatively Lipschitz with respect to r.*

Proof. By Cauchy-Schwarz, Lipschitzness of g, and strong convexity of r,

$$\begin{aligned}\langle g(w) - g(z), w - u \rangle &\le \|g(w) - g(z)\|_* \|w - u\| \le L \|w - z\| \|w - u\| \\ &\le \frac{L}{2}\left(\|w - z\|^2 + \|w - u\|^2\right) \le \frac{L}{\mu}\left(V^r_z(w) + V^r_w(u)\right).\end{aligned}$$ ◀

▶ **Lemma 3.** *If f is L-relatively smooth with respect to r, i.e. $V^f_x(y) \le L V^r_x(y)$ for all x and y, then g, defined by $g(x) := \nabla f(x)$ for all x, is L-relatively Lipschitz with respect to r.*

Proof. By assumption of relative smoothness of f and the definition of divergence,

$$\begin{aligned}L\left(V^r_z(w) + V^r_w(u)\right) &\ge V^f_z(w) + V^f_w(u) \\ &= f(w) - \left[f(z) + \nabla f(z)^\top (w - z)\right] + f(u) - \left[f(w) + \nabla f(w)^\top (u - w)\right] \\ &= V^f_z(u) - \nabla f(z)^\top (z - u) - \nabla f(z)^\top (w - z) + \nabla f(w)^\top (w - u) \\ &= V^f_z(u) + \langle g(w) - g(z), u - z \rangle .\end{aligned}$$

The result follows from the fact that $V^f_z(u) \ge 0$ by convexity of f. ◀

We now give an analysis of Algorithm 1 showing the average "regret" $\langle g(w_t), w_t - u \rangle$ of iterates decays at a $O(T^{-1})$ rate. This strengthens Lemma 3.1 of [26].

▶ **Proposition 4.** *The iterates $\{w_t\}$ of Algorithm 1 satisfy for all $u \in \mathcal{Z}$,*

$$\sum_{0 \le t < T} \langle g(w_t), w_t - u \rangle \le \lambda V^r_{z_0}(u).$$

Proof. First-order optimality conditions of w_t, z_{t+1} with respect to u imply

$$\begin{aligned}\frac{1}{\lambda}\langle g(z_t), w_t - z_{t+1}\rangle &\le V^r_{z_t}(z_{t+1}) - V^r_{w_t}(z_{t+1}) - V^r_{z_t}(w_t),\\ \frac{1}{\lambda}\langle g(w_t), z_{t+1} - u\rangle &\le V^r_{z_t}(u) - V^r_{z_{t+1}}(u) - V^r_{z_t}(z_{t+1}).\end{aligned} \tag{3}$$

Adding and manipulating gives, via relative Lipschitzness (Definition 1),

$$\begin{aligned}\frac{1}{\lambda}\langle g(w_t), w_t - u\rangle &\le V^r_{z_t}(u) - V^r_{z_{t+1}}(u) + \frac{1}{\lambda}\langle g(w_t) - g(z_t), w_t - z_{t+1}\rangle - V^r_{w_t}(z_{t+1}) - V^r_{z_t}(w_t)\\ &\le V^r_{z_t}(u) - V^r_{z_{t+1}}(u).\end{aligned} \tag{4}$$

Finally, summing and telescoping (4) yields the desired conclusion. ◀

We briefly comment on how to use Proposition 4 to approximately solve convex-concave games in a function $f(x, y)$. By applying convexity and concavity appropriately to the regret guarantee (and dividing by T, the iteration count), one can replace the left hand side of the guarantee with the duality gap of an average point $\bar{w}$ against a point u, namely $f(w^x, u^y) - f(u^x, w^y)$. By maximizing the right hand side over u, this can be converted into an overall duality gap guarantee. For some of our applications in following sections, u will be some fixed point (rather than a best response) and the regret statement will be used in a more direct manner to prove guarantees.

4 Acceleration via relative Lipschitzness

We show that directly applying Algorithm 1 to the optimization problem (1) recovers an accelerated rate for first-order convex function minimization (for simplicity, we focus on the ℓ_2 norm here; our methods extend to general norms, discussed in the full version of the paper). Our main technical result, Lemma 5, shows the gradient operator of (1) is relatively Lipschitz in the natural regularizer induced by f, which combined with Proposition 4 gives our main result, Theorem 7. Crucially, our method regularizes the dual variable with f^*, the Fenchel dual of f, which we show admits efficient implementation, allowing us to obtain our improved bound on the relative Lipschitzness parameter.

▶ **Lemma 5** (Relative Lipschitzness for the Fenchel game). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and μ-strongly convex in the Euclidean norm $\|\cdot\|_2$. Let $g(x, y) = (y, \nabla f^*(y) - x)$ be the gradient operator of the convex-concave problem (1), and define the distance-generating function $r(x, y) := \frac{\mu}{2}\|x\|_2^2 + f^*(y)$. Then, g is $1 + \sqrt{\frac{L}{\mu}}$-relatively Lipschitz with respect to r.*

Proof. Consider three points $z = (z^x, z^y)$, $w = (w^x, w^y)$, $u = (u^x, u^y)$. By direct calculation,

$$\langle g(w) - g(z), w - u\rangle = \langle w^y - z^y, w^x - u^x\rangle + \langle -w^x + z^x + \nabla f^*(w^y) - \nabla f^*(z^y), w^y - u^y\rangle. \tag{5}$$

By Cauchy-Schwarz and L^{-1}-strong convexity of f^* respectively, we have

$$\begin{aligned}\langle w^y - z^y, w^x - u^x\rangle + \langle z^x - w^x, w^y - u^y\rangle &\le \|w^y - z^y\|_2\|w^x - u^x\|_2 + \|z^x - w^x\|_2\|w^y - u^y\|_2\\ &\le \sqrt{\frac{L}{\mu}}\left(\frac{\mu}{2}\|w^x - z^x\|_2^2 + \frac{\mu}{2}\|w^x - u^x\|_2^2 + \frac{1}{2L}\|w^y - z^y\|_2^2 + \frac{1}{2L}\|w^y - u^y\|_2^2\right)\\ &\le \sqrt{\frac{L}{\mu}}\left(V^r_z(w) + V^r_w(u)\right).\end{aligned} \tag{6}$$

The second line used Young's inequality twice. Furthermore, by convexity of f^* from z^y to u^y,

$$\begin{aligned} &\langle \nabla f^*(w^y) - \nabla f^*(z^y), w^y - u^y \rangle \\ &= \langle \nabla f^*(z^y), u^y - z^y \rangle - \langle \nabla f^*(w^y), u^y - w^y \rangle - \langle \nabla f^*(z^y), w^y - z^y \rangle \\ &\le f^*(u^y) - f^*(z^y) - \langle \nabla f^*(w^y), u^y - w^y \rangle - \langle \nabla f^*(z^y), w^y - z^y \rangle \\ &= V^{f^*}_{z^y}(w^y) + V^{f^*}_{w^y}(u^y) \le V^r_z(w) + V^r_w(u). \end{aligned} \tag{7}$$

The last inequality used separability of r and nonnegativity of divergences. Summing the bounds (6) and (7) and recalling (5) yields the conclusion, where we use Definition 1. ◀

We also state a convenient fact about the form our iterates take.

▶ **Lemma 6.** *In the setting of Lemma 5, let $z_t = (x_t, y_t)$, $w_t = (x_{t+\frac{1}{2}}, y_{t+\frac{1}{2}})$ be iterates produced by running Algorithm 1 on the pair g, r. Suppose $y_0 = \nabla f(v_0)$ for some v_0. Then, $y_{t+\frac{1}{2}}$ and y_{t+1} can be recursively expressed as $y_{t+\frac{1}{2}} = \nabla f(v_{t+\frac{1}{2}})$, $y_{t+1} = \nabla f(v_{t+1})$, for*

$$v_{t+\frac{1}{2}} \leftarrow v_t + \frac{1}{\lambda}(x_t - v_t),\ v_{t+1} \leftarrow v_t + \frac{1}{\lambda}\left(x_{t+\frac{1}{2}} - v_{t+\frac{1}{2}}\right).$$

Proof. We prove this inductively; consider some iteration t. Assuming $y_t = \nabla f(v_t)$, by definition

$$\begin{aligned} y_{t+\frac{1}{2}} &= \text{argmin}_y \left\{ \left\langle \frac{1}{\lambda}\left(\nabla f^*(y_t) - x_t\right), y \right\rangle + V^{f^*}_{y_t}(y) \right\} \\ &= \text{argmax}_y \left\{ \left\langle \frac{1}{\lambda}(x_t - v_t) + v_t, y \right\rangle - f^*(y) \right\} = \nabla f\left(v_t + \frac{1}{\lambda}(x_t - v_t) \right). \end{aligned}$$

Here, we used standard facts about convex conjugates. A similar argument shows that we can compute implicitly $y_{t+1} = \nabla f(v_t + \frac{1}{\lambda}(x_{t+\frac{1}{2}} - v_{t+\frac{1}{2}}))$. ◀

We now prove Theorem 7, i.e. that we can halve function error in $O\left(\sqrt{\frac{L}{\mu}}\right)$ iterations of Algorithm 1. Simply iterating Theorem 7 yields a linear rate of convergence for smooth, strongly convex functions, yielding an ϵ-approximate minimizer in $O\left(\sqrt{\frac{L}{\mu}} \log \frac{f(x_0) - f(x^*)}{\epsilon}\right)$ iterations.

▶ **Theorem 7.** *In the setting of Lemma 5, run $T \ge 4\lambda$ iterations of Algorithm 1 initialized at $z_0 = (x_0, \nabla f(x_0))$ on the pair g, r with $\lambda = 1 + \sqrt{\frac{L}{\mu}}$, and define*

$$\bar{v} = \frac{1}{T} \sum_{0 \le t < T} v_{t+\frac{1}{2}} \text{ where } w_t = \left(x_{t+\frac{1}{2}}, \nabla f(v_{t+\frac{1}{2}}) \right).$$

Then we have $f(\bar{v}) - f(x^*) \le \frac{1}{2}(f(x_0) - f(x^*))$, where x^* minimizes f.

Proof. First, we remark that this form of w_t follows from Lemma 6, and correctness of λ follows from Lemma 5. By an application of Proposition 4, letting $u = (x^*, \nabla f(x^*))$,

$$\begin{aligned} \frac{1}{T} \sum_{0 \le t < T} \langle g(w_t), w_t - u \rangle &\le \frac{\lambda}{T} \cdot V^r_{z_0}(u) \le \frac{1}{4}\left(\frac{\mu}{2} \|x_0 - x^*\|_2^2 + V^{f^*}_{\nabla f(x_0)}(\nabla f(x^*)) \right) \\ &= \frac{1}{4}\left(\frac{\mu}{2} \|x_0 - x^*\|_2^2 + f(x_0) - f(x^*) \right) \le \frac{1}{2}\left(f(x_0) - f(x^*) \right). \end{aligned}$$

The second line used the definition of divergence in f^* and strong convexity of f, which implies $f(x_0) \geq f(x^*) + \frac{\mu}{2}\left\|x_0 - x^*\right\|_2^2$. Moreover, by the definition of g and $\nabla f(x^*) = 0$,

$$\frac{1}{T}\sum_{0\leq t<T}\langle g(w_t), w_t - u\rangle = \frac{1}{T}\sum_{0\leq t<T}\left\langle \nabla f(v_{t+\frac{1}{2}}), x_{t+\frac{1}{2}} - x^*\right\rangle + \left\langle v_{t+\frac{1}{2}} - x_{t+\frac{1}{2}}, \nabla f(v_{t+\frac{1}{2}})\right\rangle$$
$$\geq \frac{1}{T}\sum_{0\leq t<T} f(v_{t+\frac{1}{2}}) - f(x^*) \geq f(\bar{v}) - f(x^*).$$

The last line used convexity twice. Combining these two derivations yields the conclusion. ◀

For convenience, we state the full algorithm of Theorem 7 as Algorithm 2.

Algorithm 2 EG-Accel(x_0, ϵ): Extragradient accelerated smooth minimization.

```
Input: $x_0 \in \mathbb{R}^d$, $f$ $L$-smooth and $\mu$-strongly convex in $\|\cdot\|_2$, and $\epsilon_0 \geq f(x_0) - f(x^*)$
Output: $\epsilon$-approximate minimizer of $f$
$\lambda \leftarrow 1 + \sqrt{L/\mu}$, $x^{(0)} \leftarrow x_0$, $T \leftarrow 4\lceil\lambda\rceil$, $K \leftarrow \lceil\log_2 \frac{\epsilon_0}{\epsilon}\rceil$
for $0 \leq k < K$ do
  $x_0 \leftarrow x^{(k)}$, $v_0 \leftarrow x_0$
  for $0 \leq t < T$ do
    $x_{t+\frac{1}{2}} \leftarrow x_t - \frac{1}{\mu\lambda}\nabla f(v_t)$ and $v_{t+\frac{1}{2}} \leftarrow v_t + \frac{1}{\lambda}(x_t - v_t)$
    $x_{t+1} \leftarrow x_t - \frac{1}{\mu\lambda}\nabla f(v_{t+\frac{1}{2}})$ and $v_{t+1} \leftarrow v_t + \frac{1}{\lambda}(x_{t+\frac{1}{2}} - v_{t+\frac{1}{2}})$
  end for
  $x^{(k+1)} \leftarrow \frac{1}{T}\sum_{0\leq t<T} v_{t+\frac{1}{2}}$
end for
return $x^{(K)}$
```

In the full version of the paper, we give an alternative proof of acceleration leveraging relative Lipschitzness, as well as a variant of extragradient methods suited for strongly monotone operators, by applying these tools to the saddle point problem (to be contrasted with (1))

$$\min_{x\in\mathbb{R}^d} f(x) = \min_{x\in\mathbb{R}^d}\max_{y\in\mathbb{R}^d} \frac{\mu}{2}\left\|x\right\|_2^2 + \langle y, x\rangle - h^*(y), \text{ where } h(x) := f(x) - \frac{\mu}{2}\left\|x\right\|_2^2.$$

This alternative proof strategy readily generalizes the accelerated rate of Theorem 7 to general norms. While the rates attained by this alternative method are slightly less sharp (losing a $\frac{L}{\mu}$ factor in the logarithm) when compared to Theorem 7, the analysis is arguably simpler. This is in the sense that our alternative strategy shows a potential function decreases at a linear rate in every iteration, rather than requiring $O(\sqrt{L/\mu})$ iterations to halve it.

5 Area convexity rates for box-simplex games via relative Lipschitzness

In this section, we show that a local variant of Definition 1 recovers the improved convergence rate achieved by [34] for box-constrained ℓ_∞-regression, and more generally box-simplex bilinear games. Specifically, we will use the following result, a simple extension to Proposition 4 which states that relative Lipschitzness only must hold with respect to triples of algorithm iterates.

▶ **Corollary 8.** *Suppose Algorithm 1 is run with a monotone operator g and a distance generating r satisfying, for all iterations t,*

$$\langle g(w_t) - g(z_t), w_t - z_{t+1}\rangle \leq \lambda\left(V^r_{z_t}(w_t) + V_{w_t}(z_{t+1})\right). \quad (8)$$

Then, the conclusion of Proposition 4 holds.

Proof. Observe that the only applications of relative Lipschitzness in the proof of Proposition 4 are of the form (8) (namely, in (4)). Thus, the same conclusion still holds. ◀

We use Corollary 8 to give an alternative algorithm and analysis recovering the rates implied by the use of area convexity in [34], for box-simplex games, which we now define.

▶ **Definition 9** (Box-simplex game). *Let $A \in \mathbb{R}^{m\times n}$ be a matrix and let $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$ be vectors. The associated box-simplex game, and its induced monotone operator g, are*

$$\min_{x\in[-1,1]^n} \max_{y\in\Delta^m} f(x,y) := y^\top Ax - \langle b, y\rangle + \langle c, x\rangle\,,\ g(x,y) := \left(A^\top y + c, b - Ax\right). \tag{9}$$

Here, $\Delta^m := \{y \in \mathbb{R}^m_{\geq 0} : \sum_{i\in[m]} y_i = 1\}$ is the nonnegative probability simplex in m dimensions.

By a simple reduction that at most doubles the size of the input (stacking A, b with negated copies, cf. Section 3.1 of [35]), Definition 9 is a generalization of the box-constrained ℓ_∞-regression problem

$$\min_{x\in[-1,1]^m} \left\|Ax - b\right\|_\infty .$$

The work of [34] proposed a variant of extragradient algorithms, based on taking primal-dual proximal steps in the following regularizer:[3]

$$r(x,y) := y^\top |A|(x^2) + 10 \left\|A\right\|_{\infty\to\infty} \sum_{i\in[m]} y_i \log y_i. \tag{10}$$

Here, $|A|$ is the entrywise absolute value of A. The convergence rate of this algorithm was proven in [34] via an analysis based on "area convexity" of the pair (g, r), which required a somewhat sophisticated proof based on solving a partial differential equation over a triangle. We now show that the same rate can be obtained by the extragradient algorithms of [26, 28], and analyzed via local relative Lipschitzness (8).[4] We first make the following simplication without loss of generality.

▶ **Lemma 10.** *For all $x \in [-1,1]^n$ the value of $\max_{y\in\Delta^m} f(x,y)$ in* (9) *is unchanged if we remove all coordinates of b with $b_i \geq \min_{i^*\in[m]} b_{i^*} + 2\left\|A\right\|_{\infty\to\infty}$, and the corresponding rows of A. Therefore, in designing an algorithm to solve* (9) *to additive error with linear pre-processing it suffices to assume that $b_i \in [0, 2\left\|A\right\|_{\infty\to\infty}]$ for all $i \in [m]$.*

Proof. For any $x \in [-1,1]^n$, letting $i^* \in \text{argmin}_{i\in[m]} b_i$ we have

$$\max_{y\in\Delta^m} y^\top (Ax - b) = \max_{i\in[m]} [Ax - b]_i \geq -\left\|A\right\|_{\infty\to\infty} \left\|x\right\|_\infty - \min_{i^*\in[m]} b_i \geq -\left\|A\right\|_{\infty\to\infty} - b_{i^*}\ .$$

However, $[Ax-b]_i \leq \left\|A\right\|_{\infty\to\infty} - b_i$ for all $i \in [m]$. Consequently, any coordinate $i \in [m]$ that satisfies $b_i \geq b_{i^*} + 2\left\|A\right\|_{\infty\to\infty}$ has $[Ax - b]_i \leq [Ax - b]_{i_*}$ and the value of $\max_{y\in\Delta^m} f(x,y)$ is unchanged if this entry of b_i and the corresponding row of A is removed. Further, note that $\langle y, \mathbf{1}\rangle$ is a constant for all $y \in \Delta^m$. Consequently, in linear time we can remove all the coordinates i with $b_i \geq \min_{i^*\in[m]} b_{i^*} + 2\left\|A\right\|_{\infty\to\infty}$ and shift all the coordinates by an additive constant so that the minimum coordinate of a remaining b_i is 0 without affecting additive error of any x. ◀

[3] We let $\left\|A\right\|_{\infty\to\infty} := \sup_{\left\|x\right\|_\infty=1} \left\|Ax\right\|_\infty$, i.e. the ℓ_∞ operator norm of A or max ℓ_1 norm of any row.

[4] Although our analysis suffices to recover the rate of [34] for ℓ_∞ regression, the analysis of [34] is in some sense more robust (and possibly) more broadly applicable than ours, as it does not need to reason directly about how much the iterates vary in a step. Understanding or closing this gap is an interesting open problem.

We now prove our main result regarding the use of mirror prox to solve box-simplex games, using the regularizer analyzed (with a slightly different algorithm) in [34].

▶ **Theorem 11.** *Assume the preprocessing of Lemma 10 so that $b \in [0, 2\|A\|_{\infty\to\infty}]^m$. Consider running Algorithm 1 on the operator in* (9) *with $\lambda = 3$, using the regularizer in* (10)*. The resulting iterates satisfy* (8)*, and thus satisfy the conclusion of Proposition 4.*

Proof. Fix a particular iteration t. We first claim that the simplex variables w_t^y and z_{t+1}^y obey the following multiplicative stability property: entrywise,

$$w_t^y, z_{t+1}^y \in \left[\frac{1}{2} z_t^y, 2z_t^y\right]. \tag{11}$$

We will give the proof for w_t^y as the proof for z_{t+1}^y follows from the same reasoning. Recall that

$$w_t = \text{argmin}_{w\in\Delta^n\times[-1,1]^m} \left\{\left\langle\frac{1}{\lambda}g(z_t), w\right\rangle + V_{z_t}^r(w)\right\},$$

and therefore, defining $(x)^2$ and $(z_t^x)^2$ as the entrywise square of these vectors,

$$w_t^y = \text{argmin}_{y\in\Delta^m} \langle\gamma_t^y, y\rangle + 10\|A\|_{\infty\to\infty} \sum_{i\in[m]} y_i \log\frac{y_i}{[z_t^y]_i}$$
$$\text{where } \gamma_t^y := \frac{1}{\lambda}(b - Az_t^x) + |A|\left[(x)^2 - (z_t^x)^2\right].$$

Consequently, applying log and exp entrywise we have

$$w_t^y \propto \exp\left(\log z_t^y - \frac{1}{10\|A\|_{\infty\to\infty}}\gamma_t^y\right).$$

This implies the desired (11), where we use that $\|\gamma_t^y\|_\infty \le 3\|A\|_{\infty\to\infty}$, and $\exp(0.6) \le 2$. Next, we have by a straightforward calculation (Lemma 3.4, [34] or Lemma 6, [17]) that

$$\nabla^2 r(x,y) \succeq \begin{pmatrix} \textbf{diag}\left(|A_{:j}|^\top y\right) & 0 \\ 0 & \|A\|_{\infty\to\infty}\textbf{diag}\left(\frac{1}{y_i}\right)\end{pmatrix}. \tag{12}$$

By expanding the definition of Bregman divergence, we have

$$V_{z_t}^r(w_t) = \int_0^1\int_0^\alpha \|w_t - z_t\|^2_{\nabla^2 r(z_t+\beta(w_t-z_t))}\, d\beta d\alpha.$$

Fix some $\beta \in [0,1]$, and let $z_\beta := z_t + \beta(w_t - z_t)$. Since the coordinates of z_β also satisfy the stability property (11), by the lower bound of (12), we have

$$\|w_t - z_t\|^2_{\nabla^2 r(z_\beta)} \ge \sum_{i\in[m],j\in[n]} |A_{ij}|\left([z_\beta^y]_i\,[w_t^x - z_t^x]_j^2 + \frac{1}{[z_\beta^y]_i}[w_t^y - z_t^y]_i^2\right)$$
$$\ge \frac{1}{2}\sum_{i\in[m],j\in[n]} |A_{ij}|\left([z_t^y]_i\,[w_t^x - z_t^x]_j^2 + \frac{1}{[z_t^y]_i}[w_t^y - z_t^y]_i^2\right).$$

By using a similar calculation to lower bound $V_{w_t}^r(z_{t+1})$, we have by Young's inequality the desired

$$\begin{aligned}V^r_{z_t}(w_t) + V^r_{w_t}(z_{t+1}) &\geq \frac{1}{4}\sum_{i\in[m],j\in[n]} |A_{ij}| \left([z^y_t]_i \left[w^x_t - z^x_t\right]^2_j + \frac{1}{[z^y_t]_i}\left[w^y_t - z^y_t\right]^2_i \right) \\ &+ \frac{1}{4}\sum_{i\in[m],j\in[n]} |A_{ij}| \left([z^y_t]_i \left[w^x_t - z^x_{t+1}\right]^2_j + \frac{1}{[z^y_t]_i}\left[w^y_t - z^y_{t+1}\right]^2_i \right) \\ &\geq \frac{1}{3}\sum_{i\in[m],j\in[n]} A_{ij}\left(\left[w^y_t - z^y_t\right]_i \left[w^x_t - z^x_{t+1}\right]_j - \left[w^y_t - z^y_{t+1}\right]_i \left[w^x_t - z^x_t\right]_j \right) \\ &= \frac{1}{\lambda}\left\langle g(w_t) - g(z_t), w_t - z_{t+1} \right\rangle . \qquad \blacktriangleleft\end{aligned}$$

The range of the regularizer r is bounded by $O(\|A\|_{\infty\to\infty} \log m)$, and hence the iteration complexity to find an ϵ additively-approximate solution to the box-simplex game is $O(\frac{\|A\|_{\infty\to\infty} \log m}{\epsilon})$. Finally, we comment that the iteration complexity of solving the subproblems required by extragradient methods in the regularizer r to sufficiently high accuracy is logarithmically bounded in problem parameters via a simple alternating minimization scheme proposed by [34]. Here, we note that the error guarantee e.g. Proposition 4 is robust up to constant factors to solving each subproblem to ϵ additive accuracy, and appropriately using approximate optimality conditions (for an example of this straightforward extension, see Corollary 1 of [17]).

6 Randomized coordinate acceleration via expected relative Lipschitzness

We show relative Lipschitzness can compose with randomization. Specifically, we adapt Algorithm 2 to coordinate smoothness, recovering the accelerated rate first obtained in [5, 29]. We recall f is L_i-coordinate-smooth if its coordinate restriction is smooth, i.e. $|\nabla_i f(x + ce_i) - \nabla_i f(x)| \leq L_i|c|$ $\forall x \in \mathcal{X}$, $c \in \mathbb{R}$; for twice-differentiable coordinate smooth f, $\nabla^2_{ii} f(x) \leq L_i$.

Along the way, we build a framework for randomized extragradient methods via "local variance reduction" in Proposition 12. In particular, we demonstrate how for separable domains our technique can yield $O(T^{-1})$ rates for stochastic extragradient algorithms, bypassing a variance barrier encountered by prior methods [19]. Throughout, let $f : \mathbb{R}^d \to \mathbb{R}$ be L_i-smooth in coordinate i, and μ-strongly convex in $\|\cdot\|_2$, and define the distance generating function $r(x, y) = \frac{\mu}{2}\|x\|_2^2 + f^*(y)$.

Our approach modifies that of Section 4 in the following ways. First, our iterates are defined via stochastic estimators which "share randomness" (use the same coordinate in both updates). Concretely, fix some iterate $z_t = (x_t, \nabla f(v_t))$. For a distribution $\{p_i\}_{i\in[d]}$, sample $i \sim p_i$ and let

$$\begin{aligned}&g_i(z_t) := \left(\frac{1}{p_i}\nabla_i f(v_t), v_t - x_t\right),\ w^{(i)}_t = \left(x^{(i)}_{t+\frac{1}{2}}, \nabla f(v_{t+\frac{1}{2}})\right) := \mathrm{Prox}^r_{z_t}\left(\frac{1}{\lambda} g_i(z_t)\right), \\ &g_i(w^{(i)}_t) := \left(\frac{1}{p_i}\nabla_i f(v_{t+\frac{1}{2}}), v_{t+\frac{1}{2}} - \left(x_t + \frac{1}{p_i}\Delta^{(i)}_t\right)\right) \text{ for } \Delta^{(i)}_t := x^{(i)}_{t+\frac{1}{2}} - x_t, \qquad (13) \\ &z^{(i)}_{t+1} = \left(x^{(i)}_{t+1}, \nabla f(v^{(i)}_{t+1})\right) := \mathrm{Prox}^r_{z_t}\left(\frac{1}{\lambda} g_i(w^{(i)}_t)\right).\end{aligned}$$

By observation, $g_i(z_t)$ is unbiased for $g(z_t)$; however, the same cannot be said for $g_i(w_t^{(i)})$, as the random coordinate was used in the definition of $w_t^{(i)}$. Nonetheless, examining the proof of Proposition 4, we see that the conclusion

$$\langle g(\bar{w}_t), \bar{w}_t - u\rangle \le V^r_{z_t}(u) - \mathbb{E}\left[V^r_{z^{(i)}_{t+1}}(u)\right]$$

still holds for some point $\bar{w}_t$, as long as

$$\begin{aligned} \mathbb{E}\left[\left\langle g_i(w_t^{(i)}), w_t^{(i)} - u\right\rangle\right] &= \langle g(\bar{w}_t), \bar{w}_t - u\rangle, \\ \mathbb{E}\left[\left\langle g_i(w_t^{(i)}) - g_i(z_t), w_t^{(i)} - z_{t+1}^{(i)}\right\rangle\right] &\le \lambda \mathbb{E}\left[V^r_{z_t}(w_t^{(i)}) + V^r_{w_t^{(i)}}(z^{(i)}_{t+1})\right]. \end{aligned} \tag{14}$$

We make this concrete in the following claim, a generalization of Proposition 4 which handles randomized operator estimates as well as an expected variant of relative Lipschitzness. We remark that as in Corollary 8, the second condition in (14) only requires relative Lipschitzness to hold for the iterates of the algorithm, rather than globally.

▶ **Proposition 12.** *Suppose in every iteration of Algorithm 1, steps are conducted with respect to randomized gradient operators* $\left\{g_i(z_t), g_i(w_t^{(i)})\right\}$ *satisfying* (14) *for some* $\{\bar{w}_t\}$. *Then, for all* $u \in \mathcal{Z}$,

$$\mathbb{E}\left[\sum_{0 \le t < T} \langle g(\bar{w}_t), \bar{w}_t - u\rangle\right] \le \lambda V^r_{z_0}(u).$$

Proof. The proof follows identically to that of Proposition 4, where we iterate taking expectations over (4), each time applying the two conditions in (14). ◀

For the rest of this section, we overload g_i to mean the choices used in (13). This choice is motivated via the following two properties, required by (14).

▶ **Lemma 13.** *Let* $\bar{w}_t := (x_t + \sum_{i\in[d]} \Delta_t^{(i)}, \nabla f(v_{t+\frac{1}{2}}))$. *Then* $\forall u$, *taking expectations over iteration* t,

$$\mathbb{E}\left[\left\langle g_i(w_t^{(i)}), w_t^{(i)} - u\right\rangle\right] = \langle g(\bar{w}_t), \bar{w}_t - u\rangle.$$

Proof. Note $v_{t+\frac{1}{2}}$ is deterministic regardless of the sampled $i \in [d]$. Expanding for $u = (u^x, u^y)$,

$$\begin{aligned} \mathbb{E}\left[\left\langle g_i(w_t^{(i)}), w_t^{(i)} - u\right\rangle\right] = \sum_{i\in[d]} p_i \Bigg(&\left\langle \frac{1}{p_i}\nabla_i f(v_{t+\frac{1}{2}}), x^{(i)}_{t+\frac{1}{2}} - u^x\right\rangle \\ &+ \left\langle v_{t+\frac{1}{2}} - \left(x_t + \frac{1}{p_i}\Delta_t^{(i)}\right), y_{t+\frac{1}{2}} - u^y\right\rangle\Bigg) \\ &= \langle g(\bar{w}_t), \bar{w}_t - u\rangle. \end{aligned}$$

Here, we used the fact that $\nabla_i f(v_{t+\frac{1}{2}})$ is 1-sparse. ◀

▶ **Lemma 14** (Expected relative Lipschitzness). *Let* $\lambda = 1 + S_{1/2}/\sqrt{\mu}$, *where* $S_{1/2} := \sum_{i\in[d]}\sqrt{L_i}$. *Then, for the iterates* (13) *with* $p_i = \sqrt{L_i}/S_{1/2}$, *taking expectations over iteration* t,

$$\mathbb{E}\left[\left\langle g_i(w_t^{(i)}) - g_i(z_t), w_t^{(i)} - z_{t+1}^{(i)}\right\rangle\right] \le \lambda \mathbb{E}\left[V^r_{z_t}(w_t^{(i)}) + V^r_{w_t^{(i)}}(z^{(i)}_{t+1})\right].$$

Proof. Equivalently, we wish to show that

$$\mathbb{E}\left[\left\langle g_i(w_t^{(i)}) - g_i\left(z_t\right), w_t^{(i)} - z_{t+1}^{(i)}\right\rangle\right] \le \left(1 + \frac{S_{1/2}}{\sqrt{\mu}}\right)\mathbb{E}\left[V_{z_t}^r(w_t^{(i)}) + V_{w_t^{(i)}}^r(z_{t+1}^{(i)})\right].$$

The proof is patterned from Lemma 5. By direct calculation, the left hand side is

$$\begin{aligned}
&\mathbb{E}\left[\left\langle g_i(w_t^{(i)}) - g_i\left(z_t\right), w_t^{(i)} - z_{t+1}^{(i)}\right\rangle\right]\\
&= \sum_{i\in[d]} p_i \left(\frac{1}{p_i}\left\langle \nabla_i f(v_{t+\frac{1}{2}}) - \nabla_i f\left(v_t\right), x_{t+\frac{1}{2}}^{(i)} - x_{t+1}^{(i)}\right\rangle\right.\\
&\left. + \frac{1}{p_i}\left\langle x_t - x_{t+\frac{1}{2}}^{(i)}, \nabla_i f(v_{t+\frac{1}{2}}) - \nabla_i f(v_{t+1}^{(i)})\right\rangle\right)\\
&+ \sum_{i\in[d]} p_i \left\langle v_{t+\frac{1}{2}} - v_t, \nabla f(v_{t+\frac{1}{2}}) - \nabla f(v_{t+1}^{(i)})\right\rangle.
\end{aligned} \tag{15}$$

We first bound the second and third lines of (15):

$$\begin{aligned}
&\frac{1}{p_i}\left(\left\langle \nabla_i f(v_{t+\frac{1}{2}}) - \nabla_i f\left(v_t\right), x_{t+\frac{1}{2}}^{(i)} - x_{t+1}^{(i)}\right\rangle + \left\langle x_t - x_{t+\frac{1}{2}}^{(i)}, \nabla_i f(v_{t+\frac{1}{2}}) - \nabla_i f(v_{t+1}^{(i)})\right\rangle\right)\\
&\qquad \le \frac{S_{1/2}}{\sqrt{\mu}}\left(\frac{\mu}{2}\left\|x_{t+\frac{1}{2}}^{(i)} - x_{t+1}^{(i)}\right\|_2^2 + \frac{1}{2L_i}\left\|\nabla_i f(v_{t+\frac{1}{2}}) - \nabla_i f(v_{t+1}^{(i)})\right\|_2^2\right.\\
&\qquad\qquad \left. + \frac{\mu}{2}\left\|x_t - x_{t+\frac{1}{2}}^{(i)}\right\|_2^2 + \frac{1}{2L_i}\left\|\nabla_i f(v_{t+\frac{1}{2}}) - \nabla_i f\left(v_t\right)\right\|_2^2\right)\\
&\qquad \le \frac{S_{1/2}}{\sqrt{\mu}}\left(V_{z_t}^r(w_t^{(i)}) + V_{w_t^{(i)}}^r(z_{t+1}^{(i)})\right).
\end{aligned} \tag{16}$$

The first inequality used the definition $p_i = \sqrt{L_i/S_{1/2}}$ and Cauchy-Schwarz, and the second used strong convexity and Lemma 13 of the full version of the paper. Next, we bound the fourth line of (15):

$$\begin{aligned}
&\left\langle v_{t+\frac{1}{2}} - v_t, \nabla f(v_{t+\frac{1}{2}}) - \mathbb{E}\left[\nabla f(v_{t+1}^{(i)})\right]\right\rangle\\
&\le V_{\nabla f(v_t)}^{f^*}\left(\nabla f(v_{t+\frac{1}{2}})\right) + V_{\nabla f(v_{t+\frac{1}{2}})}^{f^*}\left(\mathbb{E}\left[\nabla f(v_{t+1}^{(i)})\right]\right)\\
&\le V_{\nabla f(v_t)}^{f^*}\left(\nabla f(v_{t+\frac{1}{2}})\right) + \mathbb{E}\left[V_{\nabla f(v_{t+\frac{1}{2}})}^{f^*}\left(\nabla f(v_{t+1}^{(i)})\right)\right]\\
&\le \mathbb{E}\left[V_{z_t}^r(w_t^{(i)}) + V_{w_t^{(i)}}^r(z_{t+1}^{(i)})\right].
\end{aligned}$$

The first inequality is (7), the second is convexity of Bregman divergence, and the third used nonnegativity of $\frac{\mu}{2}\left\|\cdot\right\|_2^2$. Combining with an expectation over (16) yields the claim. ◀

Crucially, our proof of these results uses the fact that our randomized gradient estimators are 1-sparse in the x component, and the fact that we "shared randomness" in the definition of the gradient estimators. Moreover, our iterates are efficiently implementable, under the "generalized partial derivative oracle" of prior work [20, 5, 29], which computes $\nabla_i f(ax + by)$ for $x, y \in \mathbb{R}^d$ and $a, b \in \mathbb{R}$. In many settings of interest, these oracles can be implemented with a dimension-independent runtime; we defer a discussion to previous references.

▶ **Lemma 15** (Iterate maintenance). *We can implement each iteration of Algorithm 3 using two generalized partial derivative oracle queries and constant additional work.*

We defer a formal statement to the full version of the paper. Combining Lemma 13 and Lemma 14, (14) is satisfied with $\lambda = 1 + S_{1/2}/\sqrt{\mu}$. Finally, all of these pieces directly imply the following, via the proof of Theorem 7 and iterating expectations. We give our full method as Algorithm 3.

▶ **Theorem 16** (Coordinate acceleration). *Algorithm 3 produces an ϵ-approximate minimizer of f in*

$$O\left(\sum_{i\in[d]} \sqrt{\frac{L_i}{\mu}} \log\left(\frac{f(x_0) - f(x^*)}{\epsilon}\right)\right) \text{ iterations in expectation,}$$

with iteration complexity given by Lemma 15.

Proof. This follows from the proof of Theorem 7, using Proposition 12 in place of Proposition 4. ◀

Algorithm 3 EG-Coord-Accel(x_0, ϵ): Extragradient accelerated coordinate minimization.

```
Input: $x_0 \in \mathbb{R}^d$, $f$ $\{L_i\}_{i\in[d]}$-coordinate smooth and $\mu$-s.c. in $\|\cdot\|_2$, and $\epsilon_0 \geq f(x_0) - f(x^*)$
$\lambda \leftarrow 1 + \sum_{i\in[d]} \sqrt{L_i/\mu}$, $T \leftarrow 4\lceil\lambda\rceil$, $K \leftarrow \lceil \log_2 \frac{\epsilon_0}{\epsilon} \rceil$, $\mathbf{A} \leftarrow \begin{pmatrix} 1 & \frac{1}{\kappa} - \frac{1}{\kappa^2} \\ 0 & 1 - \frac{1}{\kappa} + \frac{1}{\kappa^2} \end{pmatrix}$
$p_0 \leftarrow x_0$, $q_0 \leftarrow x_0$, $\mathbf{B_0} \leftarrow \mathbf{I}_{2\times 2}$
for $0 \leq k < K$ do
  Sample $\tau$ uniformly in $[0, T-1]$
  for $0 \leq t < \tau$ do
    Sample $i \propto \sqrt{L_i}$
    Compute $\nabla_i f(v_t)$, $\nabla_i f((1-\lambda^{-1})v_t + \lambda^{-1}x_t)$ via generalized partial derivative oracle
    $s_t \leftarrow \begin{pmatrix} \frac{1}{\mu\lambda p_i}\nabla_i f((1-\lambda^{-1})v_t + \lambda^{-1}x_t) & \frac{1}{\mu\lambda^2 p_i^2}\nabla_i f(v_t) \end{pmatrix}$
    $\mathbf{B_{t+1}} \leftarrow \mathbf{B_t}\mathbf{A}$, $\begin{pmatrix} p_{t+1} & q_{t+1} \end{pmatrix} \leftarrow \begin{pmatrix} p_t & q_t \end{pmatrix} - s_t \mathbf{B_{t+1}}^{-1}$
  end for
  $\mathbf{B_0} \leftarrow \begin{pmatrix} [\mathbf{B}_\tau]_{12} & [\mathbf{B}_\tau]_{12} \\ [\mathbf{B}_\tau]_{22} & [\mathbf{B}_\tau]_{22} \end{pmatrix}$, $p_0 \leftarrow p_\tau$, $q_0 \leftarrow q_\tau$
end for
return $[\mathbf{B}_\tau]_{12} p_\tau + [\mathbf{B}_\tau]_{22} q_\tau$
```

7 Discussion

We give a general condition for extragradient algorithms to converge at a $O(T^{-1})$ rate. In turn, we show that this condition (coupled with additional tools such as locality, randomization, or strong monotonicity) yields a recipe for tighter convergence guarantees in structured instances. While our condition applies generally, we find it interesting to broaden the types of instances where it obtains improved runtimes by formulating appropriate VI problems. For example, can we recover acceleration in settings such as finite-sum convex optimization (i.e. for stochastic gradient methods) [3] or composite optimization [7]? Moreover, we are interested in the interplay between (tighter analyses of) extragradient algorithms with other algorithmic frameworks. For example, is there a way to interpolate between our minimax algorithm and the momentum-based framework of [21] to obtain tight runtimes for minimax optimization? Ultimately, our hope is that our methods serve as an important stepping stone towards developing the toolkit for solving e.g. convex-concave games and variational inequalities in general.

References

1 Jacob D. Abernethy, Kevin A. Lai, Kfir Y. Levy, and Jun-Kun Wang. Faster rates for convex-concave games. In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018.*, pages 1595–1625, 2018.

2 Jacob D. Abernethy, Kevin A. Lai, and Andre Wibisono. Last-iterate convergence rates for min-max optimization. *CoRR*, abs/1906.02027, 2019. `arXiv:1906.02027`.

3 Zeyuan Allen Zhu. Katyusha: the first direct acceleration of stochastic gradient methods. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, Montreal, QC, Canada, June 19-23, 2017*, pages 1200–1205, 2017.

4 Zeyuan Allen Zhu and Elad Hazan. Optimal black-box reductions between optimization objectives. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 1606–1614, 2016.

5 Zeyuan Allen Zhu, Zheng Qu, Peter Richtárik, and Yang Yuan. Even faster accelerated coordinate descent using non-uniform sampling. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 1110–1119, 2016.

6 Heinz H. Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond lipschitz gradient continuity: First-order methods revisited and applications. *Math. Oper. Res.*, 42(2):330–348, 2017.

7 Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sciences*, 2(1):183–202, 2009.

8 Digvijay Boob, Saurabh Sawlani, and Di Wang. Faster width-dependent algorithm for mixed packing and covering lps. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 15253–15262, 2019.

9 Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

10 Yair Carmon, Yujia Jin, Aaron Sidford, and Kevin Tian. Variance reduction for matrix games. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 11377–11388, 2019.

11 Yair Carmon, Yujia Jin, Aaron Sidford, and Kevin Tian. Coordinate methods for matrix games. In *61st Annual IEEE Symposium on Foundations of Computer Science, FOCS 2020*, 2020.

12 Tatjana Chavdarova, Gauthier Gidel, François Fleuret, and Simon Lacoste-Julien. Reducing noise in GAN training with variance reduced extragradient. *CoRR*, abs/1904.08598, 2019. `arXiv:1904.08598`.

13 Jelena Diakonikolas and Lorenzo Orecchia. Accelerated extra-gradient descent: A novel accelerated first-order method. In *9th Innovations in Theoretical Computer Science Conference, ITCS 2018, January 11-14, 2018, Cambridge, MA, USA*, pages 23:1–23:19, 2018.

14 Radu-Alexandru Dragomir, Adrien Taylor, Alexandre d'Aspremont, and Jérôme Bolte. Optimal complexity and certification of bregman first-order methods. *CoRR*, abs/1911.08510, 2019. `arXiv:1911.08510`.

15 Filip Hanzely, Peter Richtarik, and Lin Xiao. Accelerated bregman proximal gradient methods for relatively smooth convex optimization. *CoRR*, abs/1808.03045, 2018. `arXiv:1808.03045`.

16 Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 6936–6946, 2019.

17 Arun Jambulapati, Aaron Sidford, and Kevin Tian. A direct $\tilde{O}(1/\epsilon)$ iteration parallel algorithm for optimal transport. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 11355–11366, 2019.

18 Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 315–323, 2013.

19 Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

20 Yin Tat Lee and Aaron Sidford. Efficient accelerated coordinate descent methods and faster algorithms for solving linear systems. In *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2013, 26-29 October, 2013, Berkeley, CA, USA*, pages 147–156, 2013.

21 Tianyi Lin, Chi Jin, and Michael I. Jordan. Near-optimal algorithms for minimax optimization. *CoRR*, abs/2002.02417, 2020. `arXiv:2002.02417`.

22 Haihao Lu. "relative-continuity" for non-lipschitz non-smooth convex optimization using stochastic (or deterministic) mirror descent. *INFORMS Journal on Optimization*, pages 288–303, 2019.

23 Haihao Lu, Robert M. Freund, and Yurii E. Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM J. Optim.*, 28(1):333–354, 2018.

24 Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019.

25 A. Nemirovski and D.$\tilde{\text{B}}$. Yudin. *Problem Complexity and Method Efficiency in Optimization.* Wiley, 1983.

26 Arkadi 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.

27 Yurii Nesterov. A method for solving a convex programming problem with convergence rate $o(1/k^2)$. *Doklady AN SSSR*, 269:543–547, 1983.

28 Yurii Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Math. Program.*, 109(2-3):319–344, 2007.

29 Yurii Nesterov and Sebastian U. Stich. Efficiency of the accelerated coordinate descent method on structured optimization problems. *SIAM Journal on Optimization*, 27(1):110–123, 2017.

30 Yuyuan Ouyang and Yangyang Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, 2019.

31 Balamurugan Palaniappan and Francis R. Bach. Stochastic variance reduction methods for saddle-point problems. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 1408–1416, 2016.

32 Alexander Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3066–3074, 2013.

33 Mark W. Schmidt, Nicolas Le Roux, and Francis R. Bach. Minimizing finite sums with the stochastic average gradient. *Math. Program.*, 162(1-2):83–112, 2017.

34 Jonah Sherman. Area-convexity, l_∞ regularization, and undirected multicommodity flow. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, Montreal, QC, Canada, June 19-23, 2017*, pages 452–460, 2017.

35 Aaron Sidford and Kevin Tian. Coordinate methods for accelerating ℓ_∞ regression and faster approximate maximum flow. In *59th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2018, 7-9 October, 2018, Paris, France*, 2018.

36 Fedor Stonyakina, Alexander Tyurin, Alexander Gasnikov, Pavel Dvurechensky, Artem Agafonov, Darina Dvinskikh, Dmitry Pasechnyuk, Sergei Artamonov, and Victorya Piskunova. Inexact relative smoothness and strong convexity for optimization and variational inequalities by inexact model. *CoRR*, abs/2001.09013, 2020. `arXiv:2001.09013`.

37 Jun-Kun Wang and Jacob D. Abernethy. Acceleration through optimistic no-regret dynamics. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 3828–3838, 2018.

38 Junyu Zhang, Minyi Hong, and Shuzhong Zhang. On lower iteration complexity bounds for the saddle point problems. *CoRR*, abs/1912.07481, 2019. `arXiv:1912.07481`.