CS395T: Continuous Algorithms, Part IV
Minimax optimization

Kevin Tian

1 Minimax theorems and monotone operators

This lecture focuses on generalizing the techniques of Part III to the setting of minimax optimization, i.e., optimization problems between two competing players $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$, who respectively wish to minimize or maximize a shared objective function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. Typically, this problem is most interesting when *strong duality* holds, i.e.,

$$\inf_{\mathbf{x}\in\mathcal{X}} \sup_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y}) = \sup_{\mathbf{y}\in\mathcal{Y}} \inf_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}, \mathbf{y}). \tag{1}$$

As remarked in previous lectures, the left-hand side of (1) is always at least the right-hand side. One famous theorem (which is arguably the most important result in linear programming) states that if $f(\mathbf{x}, \mathbf{y}) = \mathbf{y}^\top \mathbf{A}\mathbf{x}$ is bilinear and $\mathcal{X} = \mathbb{R}^n_{\geq 0}$, $\mathcal{Y} = \mathbb{R}^m_{\geq 0}$ are nonnegative orthants, (1) holds. More generally, *minimax theorems* are set of tools which can be used to establish strong duality (1). Before we state a few examples which are useful in practical applications, we begin by defining the function class which we focus on in this lecture, the subject of most minimax theorems.

Definition 1 (Convex-concave)**.** *We say a function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is* convex-concave *if $\mathcal{X}$, $\mathcal{Y}$ are convex, $f(\cdot, \bar{\mathbf{y}})$ is convex in $\mathbf{x} \in \mathcal{X}$ for any $\bar{\mathbf{y}} \in \mathcal{Y}$, and $f(\bar{\mathbf{x}}, \cdot)$ is concave*[1] *in $\mathbf{y} \in \mathcal{Y}$ for any $\bar{\mathbf{x}} \in \mathcal{X}$.*

When f is convex-concave, there exist simple-to-verify conditions which guarantee strong duality. In bounded settings, these results sometimes follow from fixed-point iteration arguments, and more generally the convex analysis tools developed in prior lectures are quite useful in these endeavors. For example, extensions of Brouwer's fixed-point theorem, which states that continuous functions f from convex, compact sets $\mathcal{X}$ to themselves admit fixed points (i.e., $\mathbf{x} \in \mathcal{X}$ with $f(\mathbf{x}) = \mathbf{x}$), give simple proofs of von Neumann's minimax theorem [vN28], by considering the best response function. For convenience, we now list two of the most commonly applicable minimax theorems.[2]

1. Sion's minimax theorem [Sio58], which extends von Neumann's minimax theorem [vN28], states (1) holds if either $\mathcal{X}$ or $\mathcal{Y}$ is compact and f is convex-concave and continuous in both variables. This continues to hold under weaker assumptions, e.g., quasi-convexity-concavity.[3]
2. A result shown in e.g., [ET99] is that if f is convex-concave and continuous in both variables, but neither $\mathcal{X}$ or $\mathcal{Y}$ is compact, as long as $f(\mathbf{x}, \bar{\mathbf{y}}) \to \infty$ for any fixed $\bar{\mathbf{y}} \in \mathcal{Y}$ as $\mathbf{x} \in \mathcal{X}$ diverges, and $f(\bar{\mathbf{x}}, \mathbf{y}) \to -\infty$ for any fixed $\bar{\mathbf{x}} \in \mathcal{X}$ as $\mathbf{y} \in \mathcal{Y}$ diverges, (1) holds.

When it exists, we refer to a point $(\mathbf{x}^\star, \mathbf{y}^\star) \in \mathcal{X} \times \mathcal{Y}$ such that $f(\mathbf{x}^\star, \mathbf{y}^\star)$ realizes the value in (1), and $\mathbf{x}^\star = \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}, \mathbf{y}^\star)$, $\mathbf{y}^\star = \operatorname{argmax}_{\mathbf{y}\in\mathcal{X}} f(\mathbf{x}^\star, \mathbf{y})$, (i.e., these points are each others' best responses) as a *saddle point* of f. More generally we define the *duality gap* of $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$ by

$$\operatorname{Gap}(\mathbf{x}, \mathbf{y}) := \sup_{\mathbf{y}'\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y}') - \inf_{\mathbf{x}'\in\mathcal{X}} f(\mathbf{x}', \mathbf{y}).$$

Intuitively, $\operatorname{Gap}(\mathbf{x}, \mathbf{y})$ measures how different objectives in a minimax game can be if a player in $\mathcal{Y}$ is allowed to respond to $\mathbf{x}$, compared to if a player in $\mathcal{X}$ can respond to $\mathbf{y}$. Clearly, any saddle point $(\mathbf{x}^\star, \mathbf{y}^\star) \in \mathcal{X} \times \mathcal{Y}$ has $\operatorname{Gap}(\mathbf{x}^\star, \mathbf{y}^\star) = 0$, since $\mathbf{y}^\star = \operatorname{argmax}_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}^\star, \mathbf{y})$ and $\mathbf{x}^\star = \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}, \mathbf{y}^\star)$. The main observation underlying this lecture is that the mirror descent framework from Part III

[1] We say a function f is concave if its negation $-f$ is convex.

[2] To gain some intuition for the conditions listed in the minimax theorems below, a useful counterexample to keep in mind is the simple function $f(x, y) = x + y$ for $x, y \in \mathbb{R}$, for which (1) clearly does not hold.

[3] Quasiconvexity states sublevel sets of a function are convex, and quasiconcavity is quasiconvexity of the negation.

generalizes readily to a broad family of linear operators. In particular, the following generalizes Theorem 2 and Corollary 4 of Part III, in light of Remark 4 of Part III.

Proposition 1. *Let $\mathcal{Z} \subseteq \mathbb{R}^d$ be convex,*[4] *let $\varphi : \mathcal{Z} \to \mathbb{R}$ be 1-strongly convex in $\|\cdot\|$ and of Legendre type, and let $\mathbf{g} : \mathcal{Z} \to \mathbb{R}^d$ satisfy $\|\mathbf{g}(\mathbf{z})\|_* \le L$ for all $\mathbf{z} \in \mathcal{Z}$. Consider iterating the update*

$$\mathbf{z}_{t+1} \leftarrow \operatorname{argmin}_{\mathbf{z}\in\mathcal{Z}} \left\{ \langle \eta \mathbf{g}(\mathbf{z}_t), \mathbf{z} \rangle + D_\varphi(\mathbf{z} \| \mathbf{z}_t) \right\}, \text{ for } 0 \le t < T, \tag{2}$$

from $\mathbf{z}_0 \in \mathcal{Z}$ with $\eta > 0$. Then for any $\mathbf{z}^\star \in \mathcal{Z}$ with $D_\varphi(\mathbf{z}^\star \| \mathbf{z}_0) \le \Theta$,

$$\frac{1}{T} \sum_{0\le t<T} \langle \mathbf{g}(\mathbf{z}_t), \mathbf{z}_t - \mathbf{z}^\star \rangle \le \frac{D_\varphi(\mathbf{z}^\star \| \mathbf{z}_0)}{\eta T} + \frac{\eta L^2}{2} \le \frac{\sqrt{2\Theta} L}{\sqrt{T}} \text{ for } \eta \leftarrow \frac{\sqrt{2\Theta}}{L\sqrt{T}}. \tag{3}$$

Moreover, if $\tilde{\mathbf{g}} : \mathcal{Z} \to \mathbb{R}$ satisfies $\mathbb{E}\tilde{\mathbf{g}}(\mathbf{z}) = \mathbf{g}(\mathbf{z})$ and $\mathbb{E}\|\tilde{\mathbf{g}}(\mathbf{z})\|_*^2 \le L^2$ for all $\mathbf{z} \in \mathcal{Z}$, iterating the update (2) with $\tilde{\mathbf{g}}$ in place of $\mathbf{g}$ yields, for any $\mathbf{z}^\star \in \mathcal{Z}$ with $D_\varphi(\mathbf{z}^\star \| \mathbf{z}_0) \le \Theta$,

$$\mathbb{E}\left[\frac{1}{T} \sum_{0\le t<T} \langle \mathbf{g}(\mathbf{z}_t), \mathbf{z}_t - \mathbf{z}^\star \rangle \right] \le \frac{D_\varphi(\mathbf{z}^\star \| \mathbf{z}_0)}{\eta T} + \frac{\eta L^2}{2} \le \frac{\sqrt{2\Theta} L}{\sqrt{T}} \text{ for } \eta \leftarrow \frac{\sqrt{2\Theta}}{L\sqrt{T}}. \tag{4}$$

Proposition 1 becomes a useful tool when studying minimax optimization when we make the realization that there is a natural operator $\mathbf{g}$ in this setting, for which we can relate the left-hand side of (3) to the duality gap of a pair of points on a product space.

Lemma 1. *Let $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be convex-concave and differentiable.*[5] *Define the operator*

$$\mathbf{g}(\mathbf{x}, \mathbf{y}) := (\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}), -\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})) \text{ for all } (\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}. \tag{5}$$

Then for any $T \in \mathbb{N}$ and $\{\mathbf{z}_t := (\mathbf{x}_t, \mathbf{y}_t)\}_{0\le t<T} \subset \mathcal{X} \times \mathcal{Y}$, letting $(\bar{\mathbf{x}}, \bar{\mathbf{y}}) := \frac{1}{T}\sum_{0\le t<T} \mathbf{z}_t$,

$$\operatorname{Gap}(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \le \sup_{\mathbf{z}^\star \in \mathcal{X}\times\mathcal{Y}} \frac{1}{T} \sum_{0\le t<T} \langle \mathbf{g}(\mathbf{z}_t), \mathbf{z}_t - \mathbf{z}^\star \rangle.$$

Proof. We begin with the observation that for all $\mathbf{z} = (\mathbf{x}, \mathbf{y}), \mathbf{z}' = (\mathbf{x}', \mathbf{y}') \in \mathcal{X} \times \mathcal{Y}$,

$$\begin{aligned} \langle \mathbf{g}(\mathbf{z}), \mathbf{z} - \mathbf{z}' \rangle &= \langle \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}), \mathbf{x} - \mathbf{x}' \rangle - \langle \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}), \mathbf{y} - \mathbf{y}' \rangle \\ &\ge (f(\mathbf{x}, \mathbf{y}) - f(\mathbf{x}', \mathbf{y})) - (f(\mathbf{x}, \mathbf{y}) - f(\mathbf{x}, \mathbf{y}')) = f(\mathbf{x}, \mathbf{y}') - f(\mathbf{x}', \mathbf{y}). \end{aligned} \tag{6}$$

The inequality used convexity of $f(\cdot, \mathbf{y})$ and concavity of $f(\mathbf{x}, \cdot)$ (Lemma 1, Part I). For $\mathbf{x}^\star := \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}, \bar{\mathbf{y}})$, $\mathbf{y}^\star := \operatorname{argmax}_{\mathbf{y}\in\mathcal{Y}} f(\bar{\mathbf{x}}, \mathbf{y})$, $\mathbf{z}^\star := (\mathbf{x}^\star, \mathbf{y}^\star)$, convexity-concavity and (6) show

$$\operatorname{Gap}(\bar{\mathbf{x}}, \bar{\mathbf{y}}) = f(\bar{\mathbf{x}}, \mathbf{y}^\star) - f(\mathbf{x}^\star, \bar{\mathbf{y}}) \le \frac{1}{T} \sum_{0\le t<T} f(\mathbf{x}_t, \mathbf{y}^\star) - f(\mathbf{x}^\star, \mathbf{y}_t) \le \frac{1}{T} \sum_{0\le t<T} \langle \mathbf{g}(\mathbf{z}_t), \mathbf{z}_t - \mathbf{z}^\star \rangle.$$

If the supremizing $(\mathbf{x}^\star, \mathbf{y}^\star)$ do not exist, taking limits with the above argument suffices. □

In other words, under suitable regularity assumptions we can directly apply Proposition 1 to algorithmically establish duality gap bounds for minimax optimization. One subtlety is that in convex optimization settings, we choose $\mathbf{z}^\star$ in Proposition 1 to be the minimizer of a function, which is independent of any algorithm. On the other hand, Lemma 1 chooses the point $\mathbf{z}^\star$ in response to the iterates of an algorithm, so typically we require bounds of the form

$$\Theta \ge \sup_{\mathbf{z}^\star \in \mathcal{Z}} D_\varphi(\mathbf{z}^\star \| \mathbf{z}_0),$$

[4] Throughout this lecture we use $\mathbf{z}$ to denote variables in a set $\mathcal{Z}$ of interest, because $\mathcal{Z}$ will often be a product space $\mathcal{X} \times \mathcal{Y}$ in minimax optimization settings, so we reserve $\mathbf{x}, \mathbf{y}$ to refer to blocks of $\mathbf{z}$.

[5] For simplicity, we assume all functions in this lecture are differentiable, though the techniques we develop extend to more general settings e.g., through the subgradient machinery of Part I. We also ignore boundary issues as in all settings we discuss, operators have finite Lipschitz constants and thus are stable to infinitesimal perturbations.

i.e., uniform upper bounds on the Bregman divergence. By choosing $\mathbf{z}_0$ as the minimizer of φ, we can apply Remark 6, Part III to obtain such bounds. This subtlety causes further issues in stochastic settings, where (4) is false once $\mathbf{z}^\star$ is no longer independent of the realization $\tilde{\mathbf{g}}$, because

$$\mathbb{E}\left\langle \tilde{\mathbf{g}}(\mathbf{z}), \mathbf{z} - \mathbf{z}^\star \right\rangle \neq \left\langle \mathbf{g}(\mathbf{z}), \mathbf{z} - \mathbf{z}^\star \right\rangle \tag{7}$$

since linearity of expectation fails. We give tools to circumvent this issue in Sections 2 and 3.

Finally, we note that much of minimax optimization theory can be further generalized to the setting of solving *variational inequalities* (VIs) in operators satisfying the following property.

Definition 2 (Monotone operator)**.** *We say operator* $\mathbf{g} : \mathcal{Z} \to \mathbb{R}^d$ *is* monotone *if*

$$\langle \mathbf{g}(\mathbf{z}) - \mathbf{g}(\mathbf{z}'), \mathbf{z} - \mathbf{z}' \rangle \geq 0 \text{ for all } \mathbf{z}, \mathbf{z}' \in \mathcal{Z}.$$

We say $\mathbf{g}$ *is* m-strongly monotone *with respect to* $h : \mathcal{Z} \to \mathbb{R}$*, or* m*-strongly monotone in* h*, if*

$$\langle \mathbf{g}(\mathbf{z}) - \mathbf{g}(\mathbf{z}'), \mathbf{z} - \mathbf{z}' \rangle \geq m \left\langle \mathbf{h}(\mathbf{z}) - \mathbf{h}(\mathbf{z}'), \mathbf{z} - \mathbf{z}' \right\rangle \text{ for all } \mathbf{z}, \mathbf{z}' \in \mathcal{Z}. \tag{8}$$

Clearly, if h is also monotone then strong monotonicity implies standard monotonicity. We say that $\mathbf{z}^\star \in \mathcal{Z}$ is a solution[6] to a variational inequality in an operator $\mathbf{g} : \mathcal{Z} \to \mathbb{R}^d$ if

$$\langle \mathbf{g}(\mathbf{z}^\star), \mathbf{z}^\star - \mathbf{z} \rangle \leq 0 \text{ for all } \mathbf{z} \in \mathcal{Z}. \tag{9}$$

We have previously shown that when $\mathbf{g}$ is the gradient of a differentiable convex function f, it is monotone (Eq. (2), Part III), and further first-order optimality (Lemma 2, Part I) shows $\mathbf{z}^\star$ solves the VI in $\mathbf{g}$ iff it minimizes f. We state a similar result in the convex-concave setting.

Lemma 2. *Let* $f : \mathcal{X} \times \mathcal{Y}$ *be convex-concave and differentiable, and let* $\mathbf{g}$ *be defined as in* (5)*. Then* $\mathbf{g}$ *is monotone, and* $\mathbf{z}^\star$ *solves the VI in* $\mathbf{g}$ *iff* $\mathrm{Gap}(\mathbf{z}^\star) = 0$.

Proof. For the first claim, we add the following inequalities, derived in the same way as (6):

$$\begin{aligned} \langle \mathbf{g}(\mathbf{z}), \mathbf{z} - \mathbf{z}' \rangle &\geq f(\mathbf{x}, \mathbf{y}') - f(\mathbf{x}', \mathbf{y}), \\ \langle \mathbf{g}(\mathbf{z}'), \mathbf{z}' - \mathbf{z} \rangle &\geq f(\mathbf{x}', \mathbf{y}) - f(\mathbf{x}, \mathbf{y}'), \text{ where } \mathbf{z} = (\mathbf{x}, \mathbf{y}),\ \mathbf{z}' = (\mathbf{x}', \mathbf{y}'). \end{aligned}$$

Finally, if $\mathbf{z}^\star$ solves the VI in $\mathbf{g}$, Lemma 1 with $T = 1$ shows $\mathrm{Gap}(\mathbf{z}^\star) \leq 0$. By definition $\mathrm{Gap}(\mathbf{z}^\star)$ is nonnegative (since we can always choose the blocks of a point as responses), so $\mathrm{Gap}(\mathbf{z}^\star) = 0$. □

We conclude by noting that strong monotonicity is related to our earlier notion of strong convexity.

Lemma 3. *If* $g = \nabla f$ *and* $h = \nabla \varphi$ *for differentiable, convex* $f : \mathcal{Z} \to \mathbb{R}$ *and* $\varphi : \mathcal{Z} \to \mathbb{R}$*, and* f *is* μ*-relatively strongly convex in* φ *(Definition 2, Part II), then* g *is* μ*-strongly monotone in* h*.*

Proof. Relative strong convexity implies $f - \mu\varphi$ is convex, so $\nabla f - \mu \nabla \varphi$ is a monotone operator. The conclusion follows since (8) is implied by monotonicity of $\nabla f - \mu \nabla \varphi$ upon rearranging. □

2 Matrix games

One of the most canonical examples of a structured minimax optimization problem is the setting of bilinear minimax optimization, where the function of interest is bilinear, i.e., $f(\mathbf{x}, \mathbf{y}) = \mathbf{y}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{y} + \mathbf{c}^\top \mathbf{x}$ for some $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{b} \in \mathbb{R}^m$, $\mathbf{c} \in \mathbb{R}^n$. Here, for $\mathcal{X} \subseteq \mathbb{R}^n$, $\mathcal{Y} \subseteq \mathbb{R}^m$, we wish to solve

$$\min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} \mathbf{y}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{y} + \mathbf{c}^\top \mathbf{x}. \tag{10}$$

In light of Lemma 1, the natural monotone operator associated with (10) is

$$g(\mathbf{x}, \mathbf{y}) = \left(\mathbf{A}^\top \mathbf{y} + \mathbf{c},\ -\mathbf{A}\mathbf{x} + \mathbf{b} \right). \tag{11}$$

The following property of (11) is the result of a straightforward expansion.

[6] Sometimes in the literature, this definition is called a *strong solution* to the variational inequality.

Fact 1. *Defining* $\mathbf{g}$ *as in* (11) *over* $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$, $\langle \mathbf{g}(\mathbf{z}) - \mathbf{g}(\mathbf{z}'), \mathbf{z} - \mathbf{z}' \rangle = 0$ *for any* $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}$, *and*

$$\frac{1}{T} \sum_{0 \le t < T} \mathbf{g}(\mathbf{z}_t) = \mathbf{g}(\bar{\mathbf{z}}) \text{ for } \{\mathbf{z}_t\}_{0 \le t < T} \subset \mathcal{Z}, \ \bar{\mathbf{z}} := \frac{1}{T} \sum_{0 \le t < T} \mathbf{z}_t.$$

For example, when $\mathcal{X}$ is a Euclidean ball and $\mathcal{Y}$ is an (ℓ_1-constrained) probability simplex, (10) generalizes hard-margin support vector machines (SVMs) by taking rows of $\mathbf{A}$ to be labeled examples signed by their label [CHW12]. Moreover, when $\mathcal{X}$ and $\mathcal{Y}$ are both Euclidean balls, (10) generalizes constrained linear regression by considering the dual formulation, which shows

$$\min_{\mathbf{x} \in \mathbb{B}(\mathbf{0}_n, 1)} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 = \left(\min_{\mathbf{x} \in \mathbb{B}(\mathbf{0}_n, 1)} \max_{\mathbf{y} \in \mathbb{B}(\mathbf{0}_m, 1)} \mathbf{y}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{y} \right)^2.$$

To illustrate some techniques suited for stochastic minimax optimization, we focus on the specific setting where $\mathcal{X} = \Delta^n$ and $\mathcal{Y} = \Delta^m$ are both probability simplices, and $\mathbf{b} = \mathbf{0}_m$, $\mathbf{c} = \mathbf{0}_n$. The corresponding minimax optimization problem is called a *matrix game*, and has the form

$$\min_{\mathbf{x} \in \Delta^n} \max_{\mathbf{y} \in \Delta^m} \mathbf{y}^\top \mathbf{A}\mathbf{x}. \tag{12}$$

The problem (12) is well-studied in the game theory literature, because it has a natural interpretation as a two-player zero-sum game. Specifically, we can let $\mathbf{A}_{ij}$ encode a score for the game if a minimizing player selects action $j \in [n]$ and a maximizing player selects action $i \in [m]$. The game is called zero-sum because negating $\mathbf{A}$ reverses the roles of the players, so the minimizing player can alternatively maximize their score with respect to a payoff matrix $-\mathbf{A}$, so the sums of scores in this equivalent "max-max" game is always zero. More generally, we view Δ^n as specifying a probability distribution over an action space for the minimizing player identified with $[n]$, and similarly Δ^m is a distribution over an action space $[m]$. A saddle point for (12) is then called a *mixed Nash equilibrium* for the zero-sum game. A powerful consequence of the von Neumann minimax theorem [vN28] is that mixed Nash equilbria always exist in zero-sum games.[7] We will see an algorithmic proof of this fact, by leveraging Proposition 1 with the following claims.

Lemma 4. *Let* $\mathcal{X} \subseteq \mathbb{R}^n$ *and* $\mathcal{Y} \subseteq \mathbb{R}^m$, *and let* $\|\cdot\|_{\mathcal{X}} : \mathcal{X} \to \mathbb{R}$ *and* $\|\cdot\|_{\mathcal{Y}} : \mathcal{Y} \to \mathbb{R}$ *be norms. Then* $\|\cdot\| : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ *is a norm, where*

$$\|(\mathbf{x}, \mathbf{y})\| := \sqrt{\|\mathbf{x}\|_{\mathcal{X}}^2 + \|\mathbf{y}\|_{\mathcal{Y}}^2}. \tag{13}$$

Moreover, the dual norm to the norm defined in (13) *is, for* $\mathbf{g} \in \mathbb{R}^n$ *and* $\mathbf{h} \in \mathbb{R}^m$,

$$\|(\mathbf{g}, \mathbf{h})\|_* = \sqrt{\|\mathbf{g}\|_{\mathcal{X},*}^2 + \|\mathbf{h}\|_{\mathcal{Y},*}^2}. \tag{14}$$

Proof. To see the first claim, of the three properties of norms (positive definiteness, absolute homogeneity, and the triangle inequality), only the triangle inequality is not immediately obvious upon using that ℓ_2 is a norm on $\mathbb{R}^2$. To see the triangle inequality, we have

$$\begin{aligned} \|(\mathbf{x} + \mathbf{x}', \mathbf{y} + \mathbf{y}')\| &= \sqrt{\|\mathbf{x} + \mathbf{x}'\|_{\mathcal{X}}^2 + \|\mathbf{y} + \mathbf{y}'\|_{\mathcal{Y}}^2} \le \sqrt{\left(\|\mathbf{x}\|_{\mathcal{X}}^2 + \|\mathbf{x}'\|_{\mathcal{X}}\right)^2 + \left(\|\mathbf{y}\|_{\mathcal{Y}}^2 + \|\mathbf{y}'\|_{\mathcal{Y}}\right)^2} \\ &\le \sqrt{\|\mathbf{x}\|_{\mathcal{X}}^2 + \|\mathbf{y}\|_{\mathcal{Y}}^2} + \sqrt{\|\mathbf{x}'\|_{\mathcal{X}}^2 + \|\mathbf{y}'\|_{\mathcal{Y}}^2} = \|(\mathbf{x}, \mathbf{y})\| + \|(\mathbf{x}', \mathbf{y}')\|. \end{aligned}$$

The first inequality followed from the triangle inequality on $\|\cdot\|_{\mathcal{X}}$, $\|\cdot\|_{\mathcal{Y}}$, and the second followed from the triangle inequality on ℓ_2 applied to the points $(\|\mathbf{x}\|_{\mathcal{X}}, \|\mathbf{y}\|_{\mathcal{Y}}), (\|\mathbf{x}'\|_{\mathcal{X}}, \|\mathbf{y}'\|_{\mathcal{Y}}) \in \mathbb{R}^2$. To see the second claim, let $\|(\mathbf{x}, \mathbf{y})\| \le 1$ so that $\|\mathbf{x}\|_{\mathcal{X}} = \alpha$ and $\|\mathbf{y}\|_{\mathcal{Y}} = \beta$ for $\alpha^2 + \beta^2 \le 1$. Then,

$$\langle \mathbf{g}, \mathbf{x} \rangle + \langle \mathbf{h}, \mathbf{y} \rangle \le \alpha \|\mathbf{g}\|_{\mathcal{X},*} + \beta \|\mathbf{h}\|_{\mathcal{Y},*} \le \alpha \|\mathbf{g}\|_{\mathcal{X},*} + \sqrt{1 - \alpha^2} \|\mathbf{h}\|_{\mathcal{Y},*} \le \sqrt{\|\mathbf{g}\|_{\mathcal{X},*}^2 + \|\mathbf{h}\|_{\mathcal{Y},*}^2},$$

where the last inequality follows by expanding and completing the square. Equality is achieved by choosing $\mathbf{x}$ and $\mathbf{y}$ in directions realizing the definitions of the dual norms $\|\mathbf{g}\|_{\mathcal{X},*}$, $\|\mathbf{h}\|_{\mathcal{Y},*}$, and with lengths induced by $\alpha \in (0, 1)$ satisfying $\sqrt{1 - \alpha^2} \|\mathbf{g}\|_{\mathcal{X},*} = \alpha \|\mathbf{h}\|_{\mathcal{Y},*}$ in the above display. □

[7] In fact, Nash famously established that mixed Nash equilbria always exist in far more general settings, with an arbitrary finite number of players with finite (non-identical) action spaces [Nas51].

Lemma 5. *Let* $\mathbf{A} \in \mathbb{R}^{m \times n}$*, and let* $\|\cdot\|_{\mathcal{X}} = \|\cdot\|_p$ *and* $\|\cdot\|_{\mathcal{Y}} = \|\cdot\|_q$ *be norms over* $\mathcal{X} \subseteq \mathbb{B}_{\|\cdot\|_{\mathcal{X}}}(1)$*,* $\mathcal{Y} \subseteq \mathbb{B}_{\|\cdot\|_{\mathcal{Y}}}(1)$ *respectively, for* $p, q \geq 1$*. Let* $p^*, q^* \geq 1$ *satisfy* $\frac{1}{p} + \frac{1}{p^*} = 1$*,* $\frac{1}{q} + \frac{1}{q^*} = 1$*. Then defining* $\|\cdot\|$ *and* $\|\cdot\|_*$ *as in* (14)*, and letting* $\mathbf{g}(\mathbf{x}, \mathbf{y}) := (\mathbf{A}^\top \mathbf{y}, -\mathbf{A}\mathbf{x})$*, we have*

$$\sup_{\mathbf{z} \in \mathcal{X} \times \mathcal{Y}} \|\mathbf{g}(\mathbf{z})\|_* \leq \sqrt{2} \|\mathbf{A}\|_{p \to q^*}.$$

Proof. Fix some $\mathbf{z} = (\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$, and recall the definition $\|\mathbf{A}\|_{p \to q} := \max_{\|\mathbf{x}\|_p \leq 1} \|\mathbf{A}\mathbf{x}\|_q$. The conclusion follows because the ℓ_2 norm in $\mathbb{R}^2$ is a $\sqrt{2}$-approximation to the ℓ_∞ norm, and

$$\left\|\mathbf{A}^\top \mathbf{y}\right\|_{\mathcal{X},*} \leq \left\|\mathbf{A}^\top\right\|_{q \to p^*} = \|\mathbf{A}\|_{p \to q^*}, \ \|-\mathbf{A}\mathbf{x}\|_{\mathcal{Y},*} \leq \|\mathbf{A}\|_{p \to q^*}.$$

□

Lemmas 4 and 5 induce the following application of Proposition 1 for solving (12). Let

$$\mathcal{Z} := \underbrace{\Delta^n}_{:=\mathcal{X}} \times \underbrace{\Delta^m}_{:=\mathcal{Y}}, \ \|\cdot\|_{\mathcal{X}} := \|\cdot\|_1, \ \|\cdot\|_{\mathcal{Y}} := \|\cdot\|_1, \tag{15}$$

and define the joint norms $\|\cdot\|$, $\|\cdot\|_*$ over $\mathcal{Z}$ following (13), (14). Moreover, consider the regularizer

$$\varphi(\mathbf{x}, \mathbf{y}) := \sum_{j \in [n]} \mathbf{x}_j \log \mathbf{x}_j + \sum_{i \in [m]} \mathbf{y}_i \log \mathbf{y}_i \tag{16}$$

over $\mathcal{Z}$. By applying Proposition 1, we immediately deduce the following.

Corollary 1. *Let* $\mathbf{A} \in \mathbb{R}^{m \times n}$ *have* $\|\mathbf{A}\|_{\max} := \max_{i \in [m], j \in [n]} |\mathbf{A}_{ij}| \leq L$*, and let* $\epsilon > 0$*. Following notation in* (15)*, there is an algorithm computing* $\mathbf{z} := (\mathbf{x}, \mathbf{y}) \in \mathcal{Z}$ *satisfying* $\mathrm{Gap}(z) \leq \epsilon$*, in time*

$$O\left(\mathrm{nnz}(\mathbf{A}) \cdot \frac{L^2 \log(mn)}{\epsilon^2}\right).$$

Proof. Consider running the algorithm in Proposition 1 for T iterations, starting from $\mathbf{z}_0 = (\frac{1}{n}\mathbf{1}_n, \frac{1}{m}\mathbf{1}_m)$, and with the operator $\mathbf{g}(\mathbf{x}, \mathbf{y}) := (\mathbf{A}^\top \mathbf{y}, -\mathbf{A}\mathbf{x})$. We showed (Lemmas 5 and 6, Part III) that by definition of the joint norm $\|\cdot\|$, the regularizer φ in (16) is 1-strongly convex in $\|\cdot\|$ and has additive range bounded by $\log(mn)$. Moreover, Lemma 5 shows that for all $(\mathbf{x}, \mathbf{y}) \in \mathcal{Z}$,

$$\|\mathbf{g}(\mathbf{x}, \mathbf{y})\|_* \leq \sqrt{2} \|\mathbf{A}\|_{1 \to \infty} = \sqrt{2} \|\mathbf{A}\|_{\max} \leq \sqrt{2} L.$$

Hence, Proposition 1 and Lemma 1 show it suffices to take $T = \Theta(\frac{L^2 \log(mn)}{\epsilon^2})$. It is clear from the entropic updates derived in Section 4, Part III that each iteration can be implemented in time $O(m+n)$ plus the time it takes to compute $\mathbf{g}(\mathbf{x}, \mathbf{y})$ for some iterate $(\mathbf{x}, \mathbf{y})$. The latter computation amounts to two matrix-vector multiplications performable in time $O(\mathrm{nnz}(\mathbf{A}))$. □

By taking $\epsilon \to 0$, Corollary 1 serves as an "algorithmic minimax theorem" in that it proves strong duality for (12), via an iterative method converging to a point with zero duality gap. Indeed, as illustrated by the relationship between mirror descent and proximal point methods explored in Part III (see also discussion in the following Section 3), our algorithmic proof can be viewed as an approximate fixed-point argument, which is similar to standard approaches in mathematics used to prove minimax theorems. This technique illustrates the utility of no-regret algorithms (Remark 4, Part III) in establishing existence of equilibria, which is a qualitative guarantee.

One can also ask the question of whether Corollary 1 is improvable in a quantitative sense, i.e., whether there exist faster algorithms computing $\mathbf{z} \in \mathcal{Z}$ with $\mathrm{Gap}(\mathbf{z}) \leq \epsilon$. We will see two such improved methods in this lecture: one presently and one in Section 3. As in Section 5, Part III, the first idea for improvement is that we can compute significantly cheaper unbiased estimates of the operator $\mathbf{g}(\mathbf{x}, \mathbf{y}) := (\mathbf{A}^\top \mathbf{y}, -\mathbf{A}\mathbf{x})$ enjoying the same boundedness properties. This is desirable because the runtime per iteration in Corollary 1 was bottlenecked by the $\mathrm{nnz}(\mathbf{A}) = \Omega(m+n)$ runtime cost of computing $\mathbf{g}(\mathbf{x}, \mathbf{y})$. Consider instead using the estimate

$$\tilde{\mathbf{g}}(\mathbf{x}, \mathbf{y}) := (\mathbf{A}_{i:}, -\mathbf{A}_{:j}) \ \text{ for } j \sim \mathbf{x}, \ i \sim \mathbf{y} \text{ independently}, \tag{17}$$

where $\mathbf{x} \in \Delta^n$ and $\mathbf{y} \in \Delta^m$ are viewed as distributions over $[n]$, $[m]$ respectively. It is simple to check that $\mathbb{E}\tilde{\mathbf{g}} = \mathbf{g}$ everywhere in $\mathcal{Z}$, and given $(\mathbf{x}, \mathbf{y}) \in \mathcal{Z}$, we can compute a sample $\tilde{\mathbf{g}}$ in time $O(m+n)$ using standard techniques.[8] Moreover, applying Lemma 5 once again shows that

$$\|\tilde{\mathbf{g}}(\mathbf{x}, \mathbf{y})\|_* \leq \sqrt{2} \|\mathbf{A}\|_{\max} \text{ for all } (\mathbf{x}, \mathbf{y}) \in \mathcal{Z}, \text{ with probability 1.}$$

At this point, it is tempting to apply the second half of Proposition 1 to conclude that we can compute random $\mathbf{z} \in \mathcal{Z}$ with $\mathbb{E}\text{Gap}(\mathbf{z}) \leq \epsilon$, using a faster runtime of $O((m+n)\log(mn) \cdot (L\epsilon^{-1})^2)$, by speeding up iterations with the cheaper (17). This actually is true, but requires a bit more work. The second half of Proposition 1 (with Lemma 1) only shows that for any *fixed* $\mathbf{z}^\star = (\mathbf{x}^\star, \mathbf{y}^\star) \in \mathcal{Z}$,

$$\mathbb{E}\left[f(\bar{\mathbf{x}}, \mathbf{y}^\star) - f(\mathbf{x}^\star, \bar{\mathbf{y}})\right] \leq \epsilon \implies \sup_{\mathbf{z}^\star = (\mathbf{x}^\star, \mathbf{y}^\star) \in \mathcal{Z}} \mathbb{E}\left[f(\bar{\mathbf{x}}, \mathbf{y}^\star) - f(\mathbf{x}^\star, \bar{\mathbf{y}})\right] \leq \epsilon.$$

However, the definition of $\mathbb{E}$Gap requires the sup to be inside the expectation, and in general for scalar random variables, we can have $\sup \mathbb{E} < \mathbb{E} \sup$ with an arbitrarily large strict inequality.[9] The culprit here is that the best response $\mathbf{z}^\star$ in the definition of Gap is dependent on the randomness used by the algorithm, which breaks independence, as discussed in (7).

We present a technique for getting around this issue, which to our knowledge first appeared in [NJLS09]. First, we show that centered second moments are boundable via uncentered variants.

Lemma 6. *Let $\mathbf{x} \in \mathbb{R}^d$ be a random variable and let $\|\cdot\|$ be a norm on $\mathbb{R}^d$. Then*

$$\mathbb{E}\|\mathbf{x} - \mathbb{E}\mathbf{x}\|^2 \leq 4\mathbb{E}\|\mathbf{x}\|^2.$$

Proof. The composition of a convex function with a monotone function is convex, so $\|\cdot\|^2$ is convex. Thus, $\|\mathbf{a}+\mathbf{b}\|^2 \leq 2\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2$ shows the desired $\mathbb{E}\|\mathbf{x} - \mathbb{E}\mathbf{x}\|^2 \leq 2\mathbb{E}\|\mathbf{x}\|^2 + 2\|\mathbb{E}\mathbf{x}\|^2 \leq 4\mathbb{E}\|\mathbf{x}\|^2$. □

Corollary 2 (Ghost iterates)**.** *In the setting of Proposition 1, suppose $\sup_{\mathbf{z} \in \mathcal{Z}} D_\varphi(\mathbf{z}\|\mathbf{z}_0) \leq \Theta$ for $\mathbf{z}_0 \in \mathcal{Z}$. Then iterating the update (2) from $\mathbf{z}_0$ with $\tilde{\mathbf{g}}$ in place of $\mathbf{g}$ yields*

$$\mathbb{E}\left[\sup_{\mathbf{z}^\star \in \mathcal{Z}} \frac{1}{T} \sum_{0 \leq t < T} \langle \mathbf{g}(\mathbf{z}_t), \mathbf{z}_t - \mathbf{z}^\star \rangle\right] \leq \frac{2\sqrt{5\Theta}L}{\sqrt{T}} \text{ for } \eta \leftarrow \frac{2\sqrt{\Theta}}{L\sqrt{5T}}.$$

Proof. Define $\mathbf{v}_t := \mathbf{g}(\mathbf{z}_t) - \tilde{\mathbf{g}}(\mathbf{z}_t)$, for all $0 \leq t < T$ to be the difference between the true operator $\mathbf{g}$ and the estimate $\tilde{\mathbf{g}}$, conditioned on $\mathbf{z}_t$. Moreover, define a sequence $\{\mathbf{w}_t\}_{0 \leq t \leq T}$ by $\mathbf{w}_0 \leftarrow \mathbf{z}_0$, and

$$\mathbf{w}_{t+1} \leftarrow \text{argmin}_{\mathbf{w} \in \mathcal{Z}} \left\{\langle \eta \mathbf{v}_t, \mathbf{w} \rangle + D_\varphi(\mathbf{w}\|\mathbf{w}_t)\right\}, \text{ for all } 0 \leq t < T.$$

The proof of Theorem 2 or Corollary 4, Part III shows that for all $0 \leq t < T$, letting $\tilde{\mathbf{g}}_t := \tilde{\mathbf{g}}(\mathbf{z}_t)$,

$$\langle \eta \tilde{\mathbf{g}}_t, \mathbf{z}_t - \mathbf{z}^\star \rangle \leq D_\varphi(\mathbf{z}^\star\|\mathbf{z}_t) - D_\varphi(\mathbf{z}^\star\|\mathbf{z}_{t+1}) + \frac{\eta^2 \|\tilde{\mathbf{g}}_t\|_*^2}{2},$$
$$\langle \eta \mathbf{v}_t, \mathbf{w}_t - \mathbf{z}^\star \rangle \leq D_\varphi(\mathbf{z}^\star\|\mathbf{w}_t) - D_\varphi(\mathbf{z}^\star\|\mathbf{w}_{t+1}) + \frac{\eta^2 \|\mathbf{v}_t\|_*^2}{2}.$$

By adding the above bounds and rearranging, we hence have for all $\mathbf{z}^\star \in \mathcal{Z}$,

$$\begin{aligned} \langle \eta \mathbf{g}_t, \mathbf{z}_t - \mathbf{z}^\star \rangle \leq\ & D_\varphi(\mathbf{z}^\star\|\mathbf{z}_t) - D_\varphi(\mathbf{z}^\star\|\mathbf{z}_{t+1}) + \frac{\eta^2 \|\tilde{\mathbf{g}}_t\|_*^2}{2} \\ & + D_\varphi(\mathbf{z}^\star\|\mathbf{w}_t) - D_\varphi(\mathbf{z}^\star\|\mathbf{w}_{t+1}) + \frac{\eta^2 \|\mathbf{v}_t\|_*^2}{2} + \langle \eta \mathbf{v}_t, \mathbf{z}_t - \mathbf{w}_t \rangle. \end{aligned} \tag{18}$$

[8] To sample $j \in [n]$ according to $\mathbf{x}$, we can place the coordinates of $\mathbf{x}$ at the leaves of a balanced binary tree augmented with subtree sums, and descend from the root flipping appropriately-biased coins until we reach a leaf.

[9] For example, suppose one person in the world is randomly choosen to receive a $\$10^9$ prize. Then $\sup \mathbb{E}$ of the prize money received is $< \$1$, where the sup is taken over people, but $\mathbb{E} \sup$ of the prize money received is $\$10^9$.

The key observation is that the left-hand side is now deterministic conditioned on $\mathbf{z}_t$, and the right-hand side telescopes to the sum of a term which is uniformly bounded for all $\mathbf{z}^\star$, and a term which is independent of $\mathbf{z}^\star$. In particular, summing for $0 \le t < T$ and taking expectations yields

$$\mathbb{E}\left[\sup_{\mathbf{z}^\star \in \mathcal{Z}} \frac{1}{T} \sum_{0 \le t < T} \langle \mathbf{g}(\mathbf{z}_t), \mathbf{z}_t - \mathbf{z}^\star \rangle\right] \le \frac{D_\varphi(\mathbf{z}^\star \| \mathbf{z}_0) + D_\varphi(\mathbf{z}^\star \| \mathbf{w}_0)}{\eta T} + \frac{5\eta L^2}{2}$$
$$\le \frac{2\Theta}{\eta T} + \frac{5\eta L^2}{2} = \frac{2\sqrt{5\Theta} L}{\sqrt{T}},$$

where the first line took expectations over (18) using Lemma 6 and that $\mathbb{E}\mathbf{v}_t = \mathbf{0}_d$ conditioned on $\mathbf{z}_t, \mathbf{w}_t$, and the second line used $\mathbf{z}_0 = \mathbf{w}_0$ with our assumption on Θ, and our choice of η. □

The term "ghost iterates" was coined by [CJST19], to refer to the phenomenon that the sequence $\{\mathbf{w}_t\}_{0 \le t < T}$ only appears in the proof of Corollary 2, and does not affect the implementation of the algorithm. By applying Corollary 2 with the estimator in (17) in place of the deterministic operator in Corollary 1, the runtime in Corollary 1 is indeed improvable to the claimed

$$O\left((m+n) \cdot \frac{L^2 \log(mn)}{\epsilon^2}\right), \tag{19}$$

if we instead settle for a point $\mathbf{z} \in \mathcal{Z}$ with $\mathbb{E}\mathrm{Gap}(\mathbf{z}) \le \epsilon$, as first observed by [GK95].[10] This can result in significant savings when $\mathbf{A}$ is moderately dense, e.g., when $\mathrm{nnz}(\mathbf{A}) = \Omega(n^2)$ and $m = n$, using the estimator (17) yields quadratically faster iterations. We note that by using concentration tools developed in the next lecture, it is straightforward to extend this result to hold with high probability, at a mild cost in the runtime (see e.g., Proposition 1, [BGJ+23] for an example).

Remark 1. *There are various generalizations of the techniques in this section, beyond the setting of (12), explored in depth in e.g., [CHW12, PB16, CJST19, CJST20]. For example, stochastic mirror descent readily handles composite terms [DSST10] (analogously to Section 5.2, Part II), and this yields extensions to nonzero linear terms* $\mathbf{b}$, $\mathbf{c}$ *in (10) with worse Lipschitz constants than the bilinear portion* $\mathbf{A}$. *Moreover, in* ℓ_p-ℓ_q *geometries for* $p, q \in (1, 2)$ *the natural extension of the sampling strategy (17) does not yield uniform bounds, which requires the analysis of the bias induced by clipping the gradient operator (see e.g., Section 4.2.2, [CJST19]). Finally, faster stochastic algorithms can be achieved under further structural assumptions. For example, [PB16, CJST19] extended the variance reduction techniques of Section 6, Part III to the minimax setting, and [CJST20] proposed the use of coordinate sampling (as opposed to the row-column sampling in (17)) to achieve more fine-grained runtime guarantees depending on sparsity properties of* $\mathbf{A}$.

3 Mirror prox

We now show that the mirror descent framework of Proposition 1 can be improved when g satisfies additional stability assumptions, which are relatively mild in many cases. This improvement is analogous to that attained by smooth gradient descent over Lipschitz gradient descent (see Theorems 2 and 3, Part II). Because of the flexibility of mirror descent (i.e., its ability to handle arbitrary operators and non-Euclidean geometries), the improved mirror prox framework of this section can be used to straightforwardly design optimal algorithms in settings which are traditionally considered challenging. We give an example in the next lecture: a simple *accelerated* method for smooth convex optimization matching the lower bound of Theorem 5, Part II.[11]

Recall that Proposition 1 is parameterized by an operator $\mathbf{g}$ and a regularizer φ, with compatible regularity properties governed by a norm $\|\cdot\|$. The mirror prox algorithm of this section requires a similar compatibility assumption, but which can be stated without explicitly defining any norm.

[10] Slightly different arguments were used in [GK95], which involved verifying (using a specialized certificate) whether the duality gap is bounded to ensure termination. We choose to present the ghost iterate argument because of its generality, and to handle settings where checking the duality gap is computationally expensive.

[11] As another example, [She17] showed how to use these improved methods to design an accelerated algorithm for smooth optimization in the ℓ_∞ geometry (see Section 5, [CST21] for an alternative exposition). This is particularly surprising, because (as discussed in Section 5.1, Part II) there are known lower bounds on the additive range of strongly convex functions in the ℓ_∞ norm, which is typically a crucial ingredient for acceleration.

Definition 3 (Relative Lipschitzness)**.** *Let $\mathcal{Z} \subseteq \mathbb{R}^d$, let $\mathbf{g} : \mathcal{Z} \to \mathbb{R}^d$ be an operator, and let $\varphi : \mathcal{Z} \to \mathbb{R}$ be a convex, differentiable function. We say* $\mathbf{g}$ *is* λ-relatively Lipschitz *with respect to φ, or λ-relatively Lipschitz in φ, if for any $\mathbf{z}, \mathbf{w}, \mathbf{u} \in \mathcal{Z}$, we have*

$$\langle \mathbf{g}(\mathbf{w}) - \mathbf{g}(\mathbf{z}), \mathbf{w} - \mathbf{u} \rangle \leq \lambda \left(D_\varphi(\mathbf{w} \| \mathbf{z}) + D_\varphi(\mathbf{u} \| \mathbf{w}) \right). \tag{20}$$

The condition (20) is somewhat opaque, as it requires three points to define. To build intuition, we first show several basic settings we have already studied, where (20) is straightforward to establish.

Lemma 7. *In the setting of Definition 3, suppose* $\mathbf{g}$ *is L-Lipschitz in $\|\cdot\|$, i.e.,*[12]

$$\|\mathbf{g}(z) - \mathbf{g}(z')\|_* \leq L \|\mathbf{z} - \mathbf{z}'\| \text{ for all } \mathbf{z}, \mathbf{z}' \in \mathcal{Z},$$

and φ is μ-strongly convex in $\|\cdot\|$. Then, $\mathbf{g}$ *is $\frac{L}{\mu}$-relatively Lipschitz with respect to φ.*

Proof. We have

$$\begin{aligned} \langle \mathbf{g}(\mathbf{w}) - \mathbf{g}(\mathbf{z}), \mathbf{w} - \mathbf{u} \rangle &\leq \|\mathbf{g}(\mathbf{w}) - \mathbf{g}(\mathbf{z})\|_* \|\mathbf{w} - \mathbf{u}\| \leq L \|\mathbf{w} - \mathbf{z}\| \|\mathbf{w} - \mathbf{u}\| \\ &\leq \frac{L}{2} \left(\|\mathbf{w} - \mathbf{z}\|^2 + \|\mathbf{w} - \mathbf{u}\|^2 \right) \leq \frac{L}{\mu} \left(D_\varphi(\mathbf{w} \| \mathbf{z}) + D_\varphi(\mathbf{u} \| \mathbf{w}) \right). \end{aligned}$$

The first inequality was Cauchy-Schwarz (Lemma 12, Part II), and the second was Lipschitzness of $\mathbf{g}$. The third inequality was Young's, and the last used strong convexity of φ (Fact 5, Part III). □

To give an example of when Lemma 7 is useful, note that it generalizes the setting of smooth convex optimization (Section 5, Part II). In particular, if f is L-smooth and φ is strongly convex (both in $\|\cdot\|$), then $\mathbf{g} = \nabla f$ is relatively Lipschitz in φ by Lemma 7 and hence mirror prox applies. We remark that mirror prox is typically analyzed under the assumptions in Lemma 7 (as was done in [Nem04], where it was introduced), but (20) is a weaker condition which also suffices (see the proof of Theorem 1). This weakening becomes important in certain applications [She17, CST21].

We next give an additional important example of a relatively Lipschitz operator.

Lemma 8. *In the setting of Definition 3, suppose $f : \mathcal{Z} \to \mathbb{R}$ is L-relatively smooth in φ (i.e., $L\varphi - f$ is convex, see Definition 2, Part II). Then ∇f is L-relatively Lipschitz with respect to φ.*

Proof. The claim follows from relative smoothness and nonnegativity of the Bregman divergence:

$$\begin{aligned} L \left(D_\varphi(\mathbf{w} \| \mathbf{z}) + D_\varphi(\mathbf{u} \| \mathbf{w}) \right) &\geq D_f(\mathbf{w} \| \mathbf{z}) + D_f(\mathbf{u} \| \mathbf{w}) \\ &= f(\mathbf{w}) - f(\mathbf{z}) - \langle \nabla f(\mathbf{z}), \mathbf{w} - \mathbf{z} \rangle + f(\mathbf{u}) - f(\mathbf{w}) - \langle \nabla f(\mathbf{w}), \mathbf{u} - \mathbf{w} \rangle \\ &= D_f(\mathbf{u} \| \mathbf{z}) + \langle \nabla f(\mathbf{w}) - \nabla f(\mathbf{z}), \mathbf{w} - \mathbf{u} \rangle \geq \langle \nabla f(\mathbf{w}) - \nabla f(\mathbf{z}), \mathbf{w} - \mathbf{u} \rangle. \end{aligned}$$

□

One interesting consequence of Lemma 8 is the following basic fact.

Corollary 3. *Let $f : \mathcal{Z} \to \mathbb{R}$ be convex and differentiable. Then ∇f is 1-relatively Lipschitz in f.*

Lemmas 7 and 8 convey the idea that relative Lipschitzness of $\mathbf{g}$ is a first-order regularity condition on $\mathbf{g}$, i.e., $\mathbf{g}$ is stable in a relative sense under small perturbations. This agrees with our intuition that smoothness of a function f is a second-order condition (see e.g., Lemma 6, Part II), as this is equivalent to first-order regularity of the first derivative of f, i.e., the operator $\mathbf{g} = \nabla f$. We now give our first variant of mirror prox, leveraging the relative Lipschitzness assumption.

Theorem 1 (Mirror prox)**.** *Let $\mathcal{Z} \subseteq \mathbb{R}^d$ be convex, let $\varphi : \mathcal{Z} \to \mathbb{R}$ be convex and of Legendre type, and let $\mathbf{g} : \mathcal{Z} \to \mathbb{R}^d$ be λ-relatively Lipschitz with respect to φ. Consider iterating the update*

$$\begin{aligned} \mathbf{z}'_t &\leftarrow \operatorname{argmin}_{\mathbf{z} \in \mathcal{Z}} \left\{ \frac{1}{\lambda} \langle \mathbf{g}(\mathbf{z}_t), \mathbf{z} \rangle + D_\varphi(\mathbf{z} \| \mathbf{z}_t) \right\}, \\ \mathbf{z}_{t+1} &\leftarrow \operatorname{argmin}_{\mathbf{z} \in \mathcal{Z}} \left\{ \frac{1}{\lambda} \langle \mathbf{g}(\mathbf{z}'_t), \mathbf{z} \rangle + D_\varphi(\mathbf{z} \| \mathbf{z}_t) \right\}, \text{ for } 0 \leq t < T, \end{aligned} \tag{21}$$

[12] This definition is consistent with our definition of a Lipschitz gradient in Definition 3, Part II.

from $\mathbf{z}_0 \in \mathcal{Z}$. *Then for any* $\mathbf{z}^\star \in \mathcal{Z}$ *with* $D_\varphi(\mathbf{z}^\star \| \mathbf{z}_0) \le \Theta$,

$$\frac{1}{T} \sum_{0 \le t < T} \langle \mathbf{g}(\mathbf{z}'_t), \mathbf{z}'_t - \mathbf{z}^\star \rangle \le \frac{\lambda \Theta}{T}.$$

Proof. The proof is analogous to the mirror descent analysis in Theorem 2, Part III. By the first-order optimality conditions on the problems defining $\mathbf{z}'_t$ and $\mathbf{z}_{t+1}$, we have for any $\mathbf{u}, \mathbf{z}^\star \in \mathcal{Z}$,

$$\begin{aligned} \frac{1}{\lambda} \langle \mathbf{g}(\mathbf{z}_t), \mathbf{z}'_t - \mathbf{u} \rangle &\le D_\varphi(\mathbf{u} \| \mathbf{z}_t) - D_\varphi(\mathbf{u} \| \mathbf{z}'_t) - D_\varphi(\mathbf{z}'_t \| \mathbf{z}_t), \\ \frac{1}{\lambda} \langle \mathbf{g}(\mathbf{z}'_t), \mathbf{z}_{t+1} - \mathbf{z}^\star \rangle &\le D_\varphi(\mathbf{z}^\star \| \mathbf{z}_t) - D_\varphi(\mathbf{z}^\star \| \mathbf{z}_{t+1}) - D_\varphi(\mathbf{z}_{t+1} \| \mathbf{z}_t), \end{aligned}$$

see Eq. (10), Part III for this derivation in more detail. Combining the first equation in the above display (with $\mathbf{u} \leftarrow \mathbf{z}_{t+1}$) and the second equation, we have upon rearranging that

$$\begin{aligned} \frac{1}{\lambda} \langle \mathbf{g}(\mathbf{z}'_t), \mathbf{z}'_t - \mathbf{z}^\star \rangle &\le D_\varphi(\mathbf{z}^\star \| \mathbf{z}_t) - D_\varphi(\mathbf{z}^\star \| \mathbf{z}_{t+1}) \\ &\quad + \frac{1}{\lambda} \langle (\mathbf{g}(\mathbf{z}'_t) - \mathbf{g}(\mathbf{z}_t)), \mathbf{z}'_t - \mathbf{z}_{t+1} \rangle - D_\varphi(\mathbf{z}_{t+1} \| \mathbf{z}'_t) - D_\varphi(\mathbf{z}'_t \| \mathbf{z}_t) \\ &\le D_\varphi(\mathbf{z}^\star \| \mathbf{z}_t) - D_\varphi(\mathbf{z}^\star \| \mathbf{z}_{t+1}), \end{aligned} \tag{22}$$

where the last line used relative Lipschitzness. Summing and multiplying by $\frac{\lambda}{T}$ yields the claim. □

Remark 2. *To motivate the update* (21)*, recall from Section 2, Part III that mirror descent is a discretization of the ideal* proximal point method*, which yields a* $\frac{1}{T}$ *rate of convergence in* T *iterations (Theorem 1, Part III). The proximal point method repeatedly solves the implicit equation*

$$\mathbf{z}_{t+1} \leftarrow \{ \langle \eta g(\mathbf{z}_{t+1}), \mathbf{z} \rangle + D_\varphi(\mathbf{z} \| \mathbf{z}_t) \},$$

which is not implementable in general because $\mathbf{z}_{t+1}$ *is used in its own definition. The updates* (21) *are explained in [Nem04], where they are first introduced, as two steps of a fixed-point iteration, where the goal is to converge to a new iterate induced by its own gradient operator. Theorem 1 gives quantitative guarantees on this fixed-point iteration, showing that under relative Lipschitzness, we can recover the improved* $\frac{1}{T}$ *rate of the proximal point method (compared to the* $T^{-1/2}$ *rate in Proposition 1). Because mirror prox requires two operator computations per step, it is sometimes called an* extragradient method*. Another well-known extragradient method is the* dual extrapolation *method of [Nes07], which can be viewed as the lazy version of mirror prox (see Remark 3, Part III for more discussion on this point, and Section D.2, [CST21] for an alternate exposition).*

Theorem 1 is powerful in that it recovers many state-of-the-art results simultaneously. For example, consider the matrix game (12). It is simple to check that $g(\mathbf{x}, \mathbf{y}) = (\mathbf{A}^\top \mathbf{y}, -\mathbf{A}\mathbf{x})$ is $L := \sqrt{2} \|\mathbf{A}\|_{\max}$-Lipschitz in the norm $\|\cdot\|$ used in Corollary 1, over the set $\mathcal{Z}$ in (15). Therefore, applying Theorem 1 the same regularizer as in Corollary 1, with the relative Lipschitzness bound in Lemma 7, immediately yields an algorithm for producing $\mathbf{z} \in \mathcal{Z}$ with $\mathrm{Gap}(\mathbf{z}) \le \epsilon$ in time

$$O\left(\mathrm{nnz}(\mathbf{A}) \cdot \frac{L \log(mn)}{\epsilon} \right),$$

which remains the best deterministic runtime known in the matrix-vector multiplication model [Nem04, Nes05]. This improves nearly-quadratically over Corollary 1 in the number of iterations, and is incomparable to our stochastic matrix game solver (with runtime in (19)), which focused on decreasing the cost per iteration rather than the iteration count. Moreover, combining Lemma 8 and Theorem 1 recovers known results on relatively smooth optimization from [BBT17, LFN18].

We mention another useful property of mirror prox: an extension of it yields linear convergence rates when the relatively Lipschitz operator $\mathbf{g}$ enjoys strong monotonicity in $\nabla\varphi$ (Definition 2). This is in contrast to the setting of mirror descent (where $\mathbf{g}$ is simply bounded over a domain, rather than relatively Lipschitz), where known lower bounds hold in the strongly convex setting, as established in Remark 1, Part II. In this sense, mirror prox recovers another aspect of the proximal point method (Theorem 1, Part III) which is unattainable by mirror descent.

Theorem 2 (Strongly monotone mirror prox)**.** *Let $\mathcal{Z} \subseteq \mathbb{R}^d$ be convex, let $\varphi : \mathcal{Z} \to \mathbb{R}$ be convex and of Legendre type, and let $\mathbf{g} : \mathcal{Z} \to \mathbb{R}^d$ be λ-relatively Lipschitz in φ and m-strongly monotone in $\nabla\varphi$. Consider iterating the update*

$$\begin{aligned} \mathbf{z}_t' &\leftarrow \operatorname{argmin}_{\mathbf{z}\in\mathcal{Z}} \left\{ \frac{1}{\lambda} \langle \mathbf{g}(\mathbf{z}_t), \mathbf{z} \rangle + D_\varphi(\mathbf{z} \| \mathbf{z}_t) \right\}, \\ \mathbf{z}_{t+1} &\leftarrow \operatorname{argmin}_{\mathbf{z}\in\mathcal{Z}} \left\{ \frac{1}{\lambda} \langle \mathbf{g}(\mathbf{z}_t'), \mathbf{z} \rangle + D_\varphi(\mathbf{z} \| \mathbf{z}_t) + \frac{m}{\lambda} D_\varphi(\mathbf{z} \| \mathbf{z}_t') \right\}, \text{ for } 0 \le t < T, \end{aligned} \tag{23}$$

from $\mathbf{z}_0 \in \mathcal{Z}$. Then if $\mathbf{z}^\star$ solves the variational inequality in $\mathbf{g}$,[13]

$$D_\varphi(\mathbf{z}^\star \| \mathbf{z}_T) \le \left(1 + \frac{m}{\lambda}\right)^{-T} D_\varphi(\mathbf{z}^\star \| \mathbf{z}_0).$$

Proof. As in Theorem 1, first-order optimality conditions and nonnegativity of D_φ imply

$$\begin{aligned} \frac{1}{\lambda} \langle \mathbf{g}(\mathbf{z}_t), \mathbf{z}_t' - \mathbf{u} \rangle &\le D_\varphi(\mathbf{u} \| \mathbf{z}_t) - D_\varphi(\mathbf{u} \| \mathbf{z}_t') - D_\varphi(\mathbf{z}_t' \| \mathbf{z}_t), \\ \frac{1}{\lambda} \langle \mathbf{g}(\mathbf{z}_t'), \mathbf{z}_{t+1} - \mathbf{z}^\star \rangle &\le D_\varphi(\mathbf{z}^\star \| \mathbf{z}_t) - D_\varphi(\mathbf{z}^\star \| \mathbf{z}_{t+1}) - D_\varphi(\mathbf{z}_{t+1} \| \mathbf{z}_t) \\ &\quad + \frac{m}{\lambda} \left(D_\varphi(\mathbf{z}^\star \| \mathbf{z}_t') - D_\varphi(\mathbf{z}^\star \| \mathbf{z}_{t+1}) \right). \end{aligned}$$

Rearranging and applying relative Lipschitzness as in (22), we then have

$$\begin{aligned} \frac{1}{\lambda} \langle \mathbf{g}(\mathbf{z}_t'), \mathbf{z}_t' - \mathbf{z}^\star \rangle - \frac{m}{\lambda} D_\varphi(\mathbf{z}^\star \| \mathbf{z}_t') &\le D_\varphi(\mathbf{z}^\star \| \mathbf{z}_t) - D_\varphi(\mathbf{z}^\star \| \mathbf{z}_{t+1}) - D_\varphi(\mathbf{z}_{t+1} \| \mathbf{z}_t) - \frac{m}{\lambda} D_\varphi(\mathbf{z}^\star \| \mathbf{z}_{t+1}) \\ &\le D_\varphi(\mathbf{z}^\star \| \mathbf{z}_t) - \left(1 + \frac{m}{\lambda}\right) D_\varphi(\mathbf{z}^\star \| \mathbf{z}_{t+1}). \end{aligned}$$

Because $\mathbf{z}^\star$ solves the VI in $\mathbf{g}$, and $\mathbf{g}$ is strongly monotone in $\nabla\varphi$, we further derive

$$\begin{aligned} \frac{1}{\lambda} \langle \mathbf{g}(\mathbf{z}_t'), \mathbf{z}_t' - \mathbf{z}^\star \rangle - \frac{m}{\lambda} D_\varphi(\mathbf{z}^\star \| \mathbf{z}_t') &\ge \frac{1}{\lambda} \langle \mathbf{g}(\mathbf{z}_t') - g(\mathbf{z}^\star), \mathbf{z}_t' - \mathbf{z}^\star \rangle - \frac{m}{\lambda} D_\varphi(\mathbf{z}^\star \| \mathbf{z}_t') \\ &\ge \frac{m}{\lambda} \langle \nabla\varphi(\mathbf{z}_t') - \nabla\varphi(\mathbf{z}^\star), \mathbf{z}_t' - \mathbf{z}^\star \rangle - \frac{m}{\lambda} D_\varphi(\mathbf{z}^\star \| \mathbf{z}_t') \\ &= \frac{m}{\lambda} D_\varphi(\mathbf{z}^\star \| \mathbf{z}_t') \ge 0. \end{aligned}$$

Combining the above two displays yields the desired claim upon recursion, since we have shown

$$D_\varphi(\mathbf{z}^\star \| \mathbf{z}_{t+1}) \le \left(1 + \frac{m}{\lambda}\right)^{-1} D_\varphi(\mathbf{z}^\star \| \mathbf{z}_t), \text{ for all } 0 \le t < T.$$

□

Remark 3. *One downside of mirror prox compared to mirror descent (Proposition 1) is it does not extend as straightforwardly to stochastic settings, while retaining the $\frac{1}{T}$ rate. This is because the dependencies induced by the two-stage updates (21), (23) do not play well with the analysis. In certain structured settings (see e.g., Section 6, [CST21]), however, it is possible to design "coupled" stochastic estimators in a way that directly lets us carry out the convergence analyses in Theorems 1, 2, taking into account dependences between iterates. This also gives us another strategy for handling the independence issue (7). Namely, because $\mathbf{z}^\star$ in Theorem 2 is deterministic (independent of the algorithm), and Theorem 2 yields high-accuracy convergence rates, we can typically directly argue that our iterates approximate a saddle point and bound the duality gap accordingly.*

[13] Recall from (9) that this means $\langle \mathbf{g}(\mathbf{z}^\star), \mathbf{z}^\star - \mathbf{z} \rangle \le 0$ for all $\mathbf{z} \in \mathcal{Z}$.

Source material

Portions of this lecture are based on reference material in [ET99, Bub15, Sid23], as well as the author's own experience working in the field.

References

[BBT17] 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.

[BGJ+23] Adam Bouland, Yosheb M. Getachew, Yujia Jin, Aaron Sidford, and Kevin Tian. Quantum speedups for zero-sum games via improved dynamic gibbs sampling. In *International Conference on Machine Learning, ICML 2023*, volume 202 of *Proceedings of Machine Learning Research*, pages 2932–2952. PMLR, 2023.

[Bub15] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

[CHW12] Kenneth L. Clarkson, Elad Hazan, and David P. Woodruff. Sublinear optimization for machine learning. *J. ACM*, 59(5):23:1–23:49, 2012.

[CJST19] 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*, pages 11377–11388, 2019.

[CJST20] Yair Carmon, Yujia Jin, Aaron Sidford, and Kevin Tian. Coordinate methods for matrix games. In *61st IEEE Annual Symposium on Foundations of Computer Science, FOCS 2020*, pages 283–293. IEEE, 2020.

[CST21] Michael B. Cohen, Aaron Sidford, and Kevin Tian. Relative lipschitzness in extragradient methods and a direct recipe for acceleration. In *12th Innovations in Theoretical Computer Science Conference, ITCS 2021, January 6-8, 2021, Virtual Conference*, volume 185 of *LIPIcs*, pages 62:1–62:18. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021.

[DSST10] John C. Duchi, Shai Shalev-Shwartz, Yoram Singer, and Ambuj Tewari. Composite objective mirror descent. In *COLT 2010 - The 23rd Conference on Learning Theory*, pages 14–26. Omnipress, 2010.

[ET99] Ivar Ekeland and Roger Temam. *Convex Analysis and Variational Problems*. Society for Industrial and Applied Mathematics, 1999.

[GK95] Michael D. Grigoriadis and Leonid G. Khachiyan. A sublinear-time randomized approximation algorithm for matrix games. *Operation Research Letters*, 18(2):53–58, 1995.

[LFN18] 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.

[Nas51] John Nash. Non-cooperative games. *The Annals of Mathematics*, 54(2):286–295, 1951.

[Nem04] 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.

[Nes05] Yurii Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1):127–152, 2005.

[Nes07] Yurii Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programing*, 109(2-3):319–344, 2007.

[NJLS09] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

[PB16] 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.

[She17] 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. ACM, 2017.

[Sid23] Aaron Sidford. *Optimization Algorithms*. 2023.

[Sio58] Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

[vN28] John von Neumann. Zur theorie der gesellschaftsspiele. *Math. Ann.*, 100:295–320, 1928.