Beyond Blackbox Models: Structured Nonsmooth Problems

Yudong Chen

This semester we have mostly considered the "blackbox model" for optimization: we access the objective function through its function value and derivatives at each query point, but we do not make use of other internal structure of the function.

Many interesting functions have rich additional structures. For example, they are built from simple functions using simple operations. Better performance can sometimes be achieved by taking advantage of such structures. In this lecture we look at one such setting, where we minimize a nonsmooth function that is given by the maximum of smooth functions.

Readings:

- Sections 4.5, 4.6 and 5.2 of Bubeck's monograph
- Juditsky and Nemiroski, First-Order Methods for Nonsmooth Convex Large-Scale Optimization, Part I and Part II

1 Smooth saddle-point representation of non-smooth functions

Consider minimizing a nonsmooth convex function $f : \mathbb{R}^d \to \mathbb{R}$ over some convex set $\mathcal{X} \subseteq \mathbb{R}^d$. We assume that f is given by the maximum of n smooth convex functions, leading to the problem

$$\min_{x\in\mathcal{X}} \underbrace{\max_{1\le i\le n} f_i(x)}_{f(x)}. \tag{1}$$

The problem (1) can be rewritten as a saddle point problem. We concatenate the f_i's into a vector-valued function $\vec{f} : \mathbb{R}^d \to \mathbb{R}^n$ with $\vec{f}(x) = (f_1(x), \ldots, f_n(x))^\top \in \mathbb{R}^n$. Let $\Delta_n := \left\{ y \in \mathbb{R}^n : \sum_{j=1}^n y_j = 1; y_i \ge 0, \forall i \right\}$ denote the probability simplex in n dimensions. Finally, define the bivariate function $H : \mathbb{R}^d \times \mathbb{R}^n \to \mathbb{R}$ by $H(x,y) = \left\langle \vec{f}(x), y \right\rangle = \sum_{i=1}^n f_i(x) y_i$. With these notations, we can rewrite f as

$$f(x) = \max_{y\in\Delta_n} H(x,y).$$

Therefore, the problem (1) is equivalent to the min-max problem

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

where $\mathcal{Y} = \Delta_n$.

For each fixed y, H is convex and smooth in x (since the f_i's are convex and smooth). For each fixed x, H is linear (hence concave and smooth) in y. By Sion's minimax theorem, there exists a pair $(x^*, y^*) \in \mathcal{X} \times \Delta_n$ that satisfies

$$H(x^*, y^*) = \max_{y\in\mathcal{Y}} H(x^*, y) = \min_{x\in\mathcal{X}} H(x, y^*). \tag{2}$$

The pair (x^*, y^*) is called a *saddle point* of H. Note that $x^* \in \operatorname{argmin}_{x\in\mathcal{X}} f(x)$. Moreover, the order of the min and max does not matter:

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} H(x,y) = \max_{y\in\mathcal{Y}} \min_{x\in\mathcal{X}} H(x,y).$$

We have transformed a nonsmooth optimization problem (1) to the saddle point problem (2), which involves a smooth function H.

1.1 Optimality condition

Below we consider the saddle point problem (2) in its general form, where $\mathcal{X}, \mathcal{Y}$ are any convex compact sets and H is a general smooth convex-concave function. The goal is to find a saddle point (x^*, y^*), which by satisfies

$$\max_{y\in\mathcal{Y}} H(x^*, y) - \min_{x\in\mathcal{Y}} H(x, y^*) = 0.$$

This suggests that one can measure the quality of a candidate solution $(\tilde{x}, \tilde{y}) \in \mathcal{X} \times \mathcal{Y}$ by the *duality gap*

$$\max_{y\in\mathcal{Y}} H(\tilde{x}, y) - \min_{x\in\mathcal{Y}} H(x, \tilde{y}) = \left(\max_{y\in\mathcal{Y}} H(\tilde{x}, y) - H(\tilde{x}, \tilde{y})\right) + \left(H(\tilde{x}, \tilde{y}) - \min_{x\in\mathcal{Y}} H(x, \tilde{y})\right).$$

A pair $(\tilde{x}, \tilde{y})$ with a small duality gap can be viewed as an approximate saddle point.

The duality gap is an analogue of the optimality gap in minimization problems. For convex minimization problem, the optimality gap can be upper bounded in terms of the gradient as $F(\tilde{x}) - F(x^*) \le \langle \nabla F(\tilde{x}), \tilde{x} - x^* \rangle$. This relationship can be generalized to saddle point problems, where the duality gap can be upper bounded by the gradients of H.

Let

$$g_{\mathcal{X}}(x,y) = \frac{\partial}{\partial x} H(x,y), \tag{3a}$$

$$g_{\mathcal{Y}}(x,y) = -\frac{\partial}{\partial y} H(x,y) \tag{3b}$$

be the gradient and negative gradient of H w.r.t. x and y, respectively. Note the minus sign for $g_{\mathcal{Y}}$, as we are maximizing over y. Define the product set $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$, the joint variables $z = (x,y) \in \mathcal{Z}, \tilde{z} = (\tilde{x}, \tilde{y}) \in \mathcal{Z}$, and the joint vector field $g(\tilde{z}) = (g_{\mathcal{X}}(\tilde{x}, \tilde{y}), g_{\mathcal{Y}}(\tilde{x}, \tilde{y}))$. Note that g need not be the gradient field of any function due to the minus sign above.

By convexity of $H(\cdot, \tilde{y})$ for each fixed $\tilde{y}$, we have

$$H(\tilde{x}, \tilde{y}) - H(x, \tilde{y}) \le \langle g_{\mathcal{X}}(\tilde{x}, \tilde{y}), \tilde{x} - x \rangle, \qquad \forall x \in \mathcal{X}.$$

Similarly by concavity of $H(\tilde{x}, \cdot)$ we have

$$H(\tilde{x}, y) - H(\tilde{x}, \tilde{y}) \le \langle g_{\mathcal{Y}}(\tilde{x}, \tilde{y}), \tilde{y} - y \rangle, \qquad \forall y \in \mathcal{X}.$$

Adding up, we see that there exists some $(x,y) \in \mathcal{X} \times \mathcal{Y}$ such that the duality gap can be controlled by

$$\begin{aligned} \max_{y\in\mathcal{Y}} H(\tilde{x}, y) - \min_{x\in\mathcal{Y}} H(x, \tilde{y}) &\le \langle g_{\mathcal{X}}(\tilde{x}, \tilde{y}), \tilde{x} - x \rangle + \langle g_{\mathcal{Y}}(\tilde{x}, \tilde{y}), \tilde{y} - y \rangle \\ &= \langle g(\tilde{z}), \tilde{z} - z \rangle. \end{aligned} \tag{4}$$

Consequently, if a pair $z^* \in \mathcal{Z}$ satisfies

$$\langle g(z^*), z^* - z \rangle \leq 0, \quad \forall z \in \mathcal{Z}, \tag{5}$$

then z^* has zero duality gap and is thus a saddle point of H. Equation (5) is equivalent to $-g(z^*) \in N_{\mathcal{Z}}(z^*)$, generalizing the optimality condition for constrained optimization.

To solve for an approximate saddle point, it suffices to find a solution $\tilde{z}$ for which the RHS of (4) is small *for all* $z \in \mathcal{Z}$. Our method of choice is the *mirror-prox* algorithm, which generalizes *mirror descent* and is applicable even when g is not a gradient field.

Remark 1. Equation (5) is called a *Variational Inequality* (VI) associated with the vector field g. Many important problems are special cases of VIs. Examples include computing Nash equilibria in two-player zero-sum games, Bellman equations in reinforcement learning, KKT conditions in optimization, and nonlinear fixed point equations in computational physics.

2 Mirror-prox

To motivate the mirror-prox algorithm, let us consider the simpler minimization problem $\min_{x \in \mathcal{X}} F$. Consider the so-called *proximal-point* method:

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ F(x) + \frac{1}{2\alpha} \|x - x_t\|_2^2 \right\}, \tag{6}$$

$$\Updownarrow \text{ (by optimality condition)}$$

$$x_{t+1} = P_{\mathcal{X}} \left\{ x_t - \alpha \nabla F(x_{t+1}) \right\}. \tag{7}$$

This method has very good convergence performance, but it is not immediately implementable. In particular, the optimization problem in (6) seems as hard as minimizing F itself, and (7) is an "implicit update" where x_{t+1} appears on both sides.

As an implementable approximation, we consider

$$y_{t+1} = P_{\mathcal{X}} \left(x_t - \alpha \nabla F(x_t) \right) = \operatorname*{argmin}_{y \in \mathcal{X}} \left\{ \langle \nabla F(x_t), y \rangle + \frac{1}{2\alpha} \|y - x_t\|_2^2 \right\}, \tag{8a}$$

$$x_{t+1} = P_{\mathcal{X}} \left(x_t - \alpha \nabla F(y_{t+1}) \right) = \operatorname*{argmin}_{y \in \mathcal{X}} \left\{ \langle \nabla F(y_{t+1}), x \rangle + \frac{1}{2\alpha} \|x - x_t\|_2^2 \right\}, \tag{8b}$$

which is known as the *extragradient* method. In this method, we first perform a standard gradient descent step (8a) on F to compute the intermediate iterate y_{t+1}. We then run an "extra" gradient descent step (8b), in which y_{t+1} is used as an approximation of the x_{t+1} on the RHS of the implicit update (7).

Mirror-prox is a generalization of extragradient. Let $g : \mathbb{R}^d \to \mathbb{R}^d$ be a vector field. Let B_ψ be the Bregman divergence associated with a differentiable convex function ψ. Mirror-prox is given by the following equations:

$$y_{t+1} = \operatorname*{argmin}_{y \in \mathcal{X}} \left\{ \langle g(x_t), y \rangle + \frac{1}{\alpha} B_\psi(y, x_t) \right\}, \tag{9a}$$

$$x_{t+1} = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \langle g(y_{t+1}), x \rangle + \frac{1}{\alpha} B_\psi(x, x_t) \right\}. \tag{9b}$$

One can verify that (9) is equivalent to

$$\nabla\psi(y'_{t+1}) = \nabla\psi(x_t) - \alpha g(x_t), \tag{10a}$$
$$y_{t+1} \in \operatorname*{argmin}_{y\in\mathcal{X}} B_\psi(y, y'_{t+1}), \tag{10b}$$
$$\nabla\psi(x'_{t+1}) = \nabla\psi(x_t) - \alpha g(y_{t+1}), \tag{10c}$$
$$x_{t+1} \in \operatorname*{argmin}_{x\in\mathcal{X}} B_\psi(x, x'_{t+1}). \tag{10d}$$

Note that extragradient is a special case with $\psi(x) = \frac{1}{2}\|x\|_2^2$ and $B_\psi(x, x') = \frac{1}{2}\|x - x'\|_2^2$.

2.1 Convergence guarantee for mirror-prox

We say that the vector field g is L-Lipschitz w.r.t. $\|\cdot\|$ if

$$\|g(x) - g(y)\|_* \le L\|x - y\|, \quad \forall x, y.$$

The following theorem establishes that mirror-prox achieves an $O(1/t)$ rate for Lipschitz g.

Theorem 1. *Suppose ψ is μ-strongly convex with respect to the norm $\|\cdot\|$ with dual norm $\|\cdot\|_*$, and g is L-Lipschitz with respect to $\|\cdot\|$. Let $x_1 = \operatorname{argmin}_{x\in\mathcal{X}} \psi(x)$ and $R^2 := \sup_{x\in\mathcal{X}} \psi(x) - \psi(x_1)$. Then mirror-prox with stepsize $\alpha = \frac{\mu}{L}$ and initialized at x_1 satisfies*

$$\frac{1}{T}\sum_{t=1}^{T} \langle g(y_{t+1}), y_{t+1} - x\rangle \le \frac{LR^2}{\mu T}, \qquad \forall T \ge 1, \forall x \in \mathcal{X}.$$

When $g = \nabla f$ is the gradient field of an (L-smooth) convex function f, the LHS above is an upper bound on the optimality gap $f\left(\frac{1}{t}\sum_{s=1}^{t} y_{s+1}\right) - f(x^*)$ (by convexity and Jensen's). Such an $O(1/t)$ bound for minimizing smooth f can also be achieved by the standard mirror descent method. Why do we need the more sophiscated mirror-prox algorithm?

The power of mirror-prox reveals itself when g is *not* the gradient field of any function. In this case, mirror descent may not even converge; see below for an example.

Example 1. Consider the min-max problem $\min_x \max_y \{H(x, y) = xy\}$. In this case, the vector field $g : \mathbb{R}^2 \to \mathbb{R}^2$ defined in (3) is given by $g(x, y) = \begin{pmatrix} \frac{\partial}{\partial x} H(x, y) \\ -\frac{\partial}{\partial y} H(x, y) \end{pmatrix} = \begin{pmatrix} y \\ -x \end{pmatrix}$. The mirror descent algorithm with ℓ_2 Bregman divergence is given by the update

$$\begin{pmatrix} x_{t+1} \\ y_{t+1} \end{pmatrix} = \begin{pmatrix} x_t \\ y_t \end{pmatrix} - \alpha g(x_t, y_t) = \begin{pmatrix} x_t - \alpha y_t \\ y_t + \alpha x_t \end{pmatrix},$$

which is also known as *gradient descent-ascent*. This algorithm diverges for every stepsize $\alpha > 0$, since $x_{t+1}^2 + y_{t+1}^2 = (1 + \alpha^2)(x_t^2 + y_{t+1}^2) > x_t^2 + y_{t+1}^2$.

2.2 Proof of Theorem 1

Proof. Fix an arbitrary $x \in \mathcal{X}$. We write

$$\langle g(y_{t+1}), y_{t+1} - x\rangle = \langle g(y_{t+1}), x_{t+1} - x\rangle + \langle g(x_t), y_{t+1} - x_{t+1}\rangle + \langle g(y_{t+1}) - g(x_t), y_{t+1} - x_{t+1}\rangle. \tag{11}$$

We bound these three terms separately.

For the first term, we follow the same arguments used in the analysis of mirror descent:

$$\begin{aligned}
&\langle g(y_{t+1}), x_{t+1} - x\rangle \\
\leq &\frac{1}{\alpha}\langle \nabla\psi(x_t) - \nabla\psi(x_{t+1}), x_{t+1} - x\rangle && \text{optimality condition for } x_{t+1} \text{ and (9b)} \\
= &\frac{1}{\alpha}\left(B_\psi(x, x_t) - B_\psi(x, x_{t+1}) - B_\psi(x_{t+1}, x_t)\right). && \text{3-point identity}
\end{aligned}$$

Similar, for the second term we have

$$\begin{aligned}
&\langle g(x_t), y_{t+1} - x_{t+1}\rangle \\
\leq &\frac{1}{\alpha}\langle \nabla\psi(x_t) - \nabla\psi(y_{t+1}), y_{t+1} - x_{t+1}\rangle && \text{optimality condition for } y_{t+1} \text{ and (9a)} \\
= &\frac{1}{\alpha}\left(B_\psi(x_{t+1}, x_t) - B_\psi(x_{t+1}, y_{t+1}) - B_\psi(y_{t+1}, x_t)\right) && \text{3-point identity} \\
\leq &\frac{1}{\alpha}\left(B_\psi(x_{t+1}, x_t) - \frac{\mu}{2}\|x_{t+1} - y_{t+1}\|^2 - \frac{\mu}{2}\|y_{t+1} - x_t\|^2\right). && \mu\text{-strong convexity of } \psi
\end{aligned}$$

For the third term, we have

$$\begin{aligned}
&\langle g(y_{t+1}) - g(x_t), y_{t+1} - x_{t+1}\rangle \\
\leq &\|g(y_{t+1}) - g(x_t)\|_* \|y_{t+1} - x_{t+1}\| && \text{Holder} \\
\leq &L\|y_{t+1} - x_t\|_* \|y_{t+1} - x_{t+1}\| && L\text{-smoothness of } g \\
\leq &\frac{L}{2}\|y_{t+1} - x_t\|^2 + \frac{L}{2}\|y_{t+1} - x_{t+1}\|^2. && ab \leq \frac{1}{2}a^2 + \frac{1}{2}b^2
\end{aligned}$$

Plugging these three bounds into (11), we get

$$\begin{aligned}
\langle g(y_{t+1}), y_{t+1} - x\rangle \leq &\frac{1}{\alpha}\left(B_\psi(x, x_t) - B_\psi(x, x_{t+1}) - \frac{\mu}{2}\|x_{t+1} - y_{t+1}\|^2 - \frac{\mu}{2}\|y_{t+1} - x_t\|^2\right) \\
&+ \frac{L}{2}\|y_{t+1} - x_t\|^2 + \frac{L}{2}\|y_{t+1} - x_{t+1}\|^2.
\end{aligned}$$

We take $\alpha = \frac{\mu}{L}$ to get cancellation on the RHS, leading to

$$\langle g(y_{t+1}), y_{t+1} - x\rangle \leq \frac{1}{\alpha}\left(B_\psi(x, x_t) - B_\psi(x, x_{t+1})\right).$$

Summing over $t = 1, \ldots, T$ and dividing both sides by T gives

$$\frac{1}{T}\sum_{t=1}^{T}\langle g(y_{t+1}), y_{t+1} - x\rangle \leq \frac{B_\psi(x, x_1)}{\alpha T} = \frac{LB_\psi(x, x_1)}{\mu T}.$$

By optimality condition for $x_1 = \operatorname{argmin}_{x\in\mathcal{X}}\psi(x)$, we have $\langle -\nabla\psi(x_1), x - x_1\rangle \leq 0, \forall x \in \mathcal{X}$, hence

$$B_\psi(x, x_1) = \psi(x) - \psi(x_1) - \langle \nabla\psi(x_1), x - x_1\rangle \leq \psi(x) - \psi(x_1) \leq R^2.$$

The theorem follows. □

3 Saddle Point Mirror-Prox

We now apply mirror-prox to the saddle point problem (2), for which the vector field g is given in equation (3).

Suppose $\psi_{\mathcal{X}} : \mathcal{X} \to \mathbb{R}$ is 1-strongly convex on $\mathcal{X}$ w.r.t. some norm $\|\cdot\|_{\mathcal{X}}$. Let $R_{\mathcal{X}}^2 = \sup_{x\in\mathcal{X}} \psi_{\mathcal{X}}(x) - \min_{x\in\mathcal{X}} \psi_{\mathcal{X}}(x)$. We define $\psi_{\mathcal{Y}}, \|\cdot\|_{\mathcal{Y}}$ and $R_{\mathcal{Y}}^2$ similarly. We then define a function $\psi : \mathcal{Z} \to \mathbb{R}$ by $\psi(z) = a\psi_{\mathcal{X}}(x) + b\psi_{\mathcal{Y}}(y)$. It is immediate that ψ is 1-strongly convex on $\mathcal{Z}$ w.r.t. the norm $\|z\|_{\mathcal{Z}} = \sqrt{a\|x\|_{\mathcal{X}}^2 + b\|y\|_{\mathcal{Y}}^2}$. Below we set $a = \frac{1}{R_{\mathcal{X}}^2}, b = \frac{1}{R_{\mathcal{Y}}^2}$.

We assume that the function H is $(L_{11}, L_{12}, L_{22}, L_{21})$-smooth in the sense that

$$\begin{aligned}
\|g_{\mathcal{X}}(x,y) - g_{\mathcal{X}}(x',y)\|_{\mathcal{X}}^* &\le L_{11}\|x - x'\|_{\mathcal{X}}, \\
\|g_{\mathcal{X}}(x,y) - g_{\mathcal{X}}(x,y')\|_{\mathcal{X}}^* &\le L_{12}\|y - y'\|_{\mathcal{Y}}, \\
\|g_{\mathcal{Y}}(x,y) - g_{\mathcal{Y}}(x,y')\|_{\mathcal{Y}}^* &\le L_{22}\|y - y'\|_{\mathcal{Y}}, \\
\|g_{\mathcal{Y}}(x,y) - g_{\mathcal{Y}}(x',y)\|_{\mathcal{Y}}^* &\le L_{21}\|x - x'\|_{\mathcal{X}},
\end{aligned}$$

for all $x, x' \in \mathcal{X}$ and $y, y' \in \mathcal{Y}$. This assumption implies that the joint vector field g is $L_{\mathcal{Z}}$-Lipschitz on $\mathcal{Z}$ in the norm $\|\cdot\|_{\mathcal{Z}}$ with $L_{\mathcal{Z}} = 2\max\left\{L_{11}R_{\mathcal{X}}^2, L_{22}R_{\mathcal{Y}}^2, L_{12}R_{\mathcal{X}}R_{\mathcal{Y}}, L_{21}R_{\mathcal{X}}R_{\mathcal{Y}}\right\}$.

The mirror-prox method (9), when specialized to the above g and ψ, is called *Saddle Point Mirror-Prox* (SP-MP). It is given explicitly as follows: let $z_1 \in \operatorname{argmin}_{z\in\mathcal{Z}} \psi(z)$; compute $z_t = (x_t, y_t)$ and $w_t = (u_t, v_t)$ by

$$\begin{aligned}
w_{t+1} &= \operatorname*{argmin}_{z\in\mathcal{Z}} \left\{ \langle g(z_t), z\rangle + \frac{1}{\alpha} B_\psi(z, z_t) \right\}, \\
z_{t+1} &= \operatorname*{argmin}_{z\in\mathcal{Z}} \left\{ \langle g(w_t), z\rangle + \frac{1}{\alpha} B_\psi(z, z_t) \right\}.
\end{aligned}$$

Theorem 2. *Under the above assumptions, SP-MP with $\alpha = \frac{1}{L_{\mathcal{Z}}}$ produces a pair $\left(\frac{1}{T}\sum_{t=1}^T u_{s+1}, \frac{1}{T}\sum_{t=1}^T v_{t+1}\right)$ that satisfies the duality gap bound*

$$\max_{y\in\mathcal{Y}} H\left(\frac{1}{T}\sum_{t=1}^T u_{s+1}, y\right) - \min_{x\in\mathcal{Y}} H\left(x, \frac{1}{T}\sum_{t=1}^T v_{t+1}\right) \le \frac{2L_{\mathcal{Z}}}{T}.$$

Proof. Set $\tilde{x} = \frac{1}{T}\sum_{t=1}^T u_{s+1}$ and $\tilde{y} = \frac{1}{T}\sum_{t=1}^T v_{t+1}$. We upper bound the duality gap of $(\tilde{x}, \tilde{y})$ using (4), whose RHS can in turn be bounded using Theorem 1. □

In what follows we discuss the application of Theorem 2 to the nonsmooth optimization problem (1), as well as two other applications.

3.1 Minimizing max of smooth functions

We apply Theorem 2 to the saddle point representation (2) of minimizing the nonsmooth function $f(x) = \max_{1\le i\le n} f_i(x)$ over $\mathcal{X}$. We assume each f_i is M-Lipschitz and L-smooth w.r.t. $\|\cdot\|_{\mathcal{X}}$. Let $\mathcal{Y} = \Delta_n$ and $\|\cdot\|_{\mathcal{Y}} = \|\cdot\|_1$ and $\psi_{\mathcal{Y}} =$ negative entropy. One can verify that the above assumptions are satisfied with

$$\beta_{11} = L, \beta_{12} = M, \beta_{21} = M, \beta_{22} = 0.$$

Consequently, by Theorem 2 we obtain an

$$O\left(\frac{LR_{\mathcal{X}}^2 + MR_{\mathcal{X}}\sqrt{\log n}}{T}\right)$$

rate for minimizing f. This improves on the $1/\sqrt{T}$ rate for minimizing a general nonsmooth function.

3.2 Two-player zero-sum games

Consider a two-player zero-sum game with the payoff matrix $A \in \mathbb{R}^{m\times n}$. Upon taking actions $i \in \{1,\dots,m\}$ and $j \in \{1,\dots,n\}$, respectively, player 1 receives a payoff A_{ij} and player 2 receives $-A_{ij}$. The goal is to compute a mixed-strategy Nash equilibrium, which is a saddle point of the bilinear min-max problem

$$\min_{x\in\Delta_m}\max_{y\in\Delta_n} \underbrace{x^\top A y}_{H(x,y)}$$

See these slides for additional background on two-player zero-sum games.

We apply SP-MP to this problem with $\psi_{\mathcal{X}}, \psi_{\mathcal{Y}}$ being the negative entropy, which is 1-strongly convex w.r.t. $\|\cdot\|_1$. Note that $g_{\mathcal{X}}(x,y) = \frac{\partial}{\partial x}H(x,y) = Ay$ and $g_{\mathcal{Y}}(x,y) = -\frac{\partial}{\partial y}H(x,y) = -A^\top x$. We immediately have $L_{11} = L_{22} = 0$. Moreover, letting A_i denotes the j-th column of A and $\|A\|_{\max} := \sum_{ij}|A_{ij}|$, we have

$$\left\|A(y-y')\right\|_\infty = \left\|\sum_{j=1}^n \left(y(i)-y'(i)\right)A_i\right\|_\infty \le \|A\|_{\max}\left\|y-y'\right\|_1, \quad \forall y, y',$$

hence $L_{12} = \|A\|_{\max}$; similarly $L_{21} = \|A\|_{\max}$. Theorem 2 implies that SP-MP finds an approximate Nash equilibrium with duality gap $\frac{\|A\|_{\max}\sqrt{\log m \log n}}{T}$ is T iterations. This improves on the $1/\sqrt{T}$ rate of the Multiplicative Weight algorithm.

3.3 Linear max-margin classification

Consider n data points of the form (A_j, ℓ_j), $i \in \{1,\dots,n\}$, where $A_j \in \mathbb{R}^m$ is the feature vector of the i-th data point and $\ell_j \in \{\pm 1\}$ is the label. Using a linear classifier given by $x \in \mathbb{B}_2^m := \{x' \in \mathbb{R}^m s : \|x'\|_2 \le 1\}$, we want to classify each data point such that $\text{sign}(x^\top A_j) = \ell_j$, or equivalently $x^\top(\ell_j A_j) > 0$. Perfect classification may not be possible. In this case we instead seek for the *maximum margin classifier* x, which is the solution to the min-max problem

$$\underbrace{\max_{x\in\mathbb{B}_2^m}\min_{1\le j\le m} \underbrace{x^\top\left(\ell_j A_j\right)}_{\text{margin for } i\text{th data point}}}_{\text{margin}} = \max_{x\in\mathbb{B}_2^m}\min_{y\in\Delta_n} \underbrace{x^\top \bar{A} y}_{H(x,y)},$$

where $\bar{A} \in \mathbb{R}^{m\times n}$ is the matrix with $\ell_j A_j$ being its j-th column. This problem is similar to two-player zero-sum games except that x lives in the unit ℓ_2 ball $\mathbb{B}_2^m$.

Suppose $\left\|A_j\right\|_2 \le B, \forall j$. It is easy to show that $\bar{A}$ is $(0, B, 0, B)$-smooth with respect to $\|\cdot\|_2$ on $\mathbb{B}_2^m$ and $\|\cdot\|_1$ on Δ_n. We apply SP-MP to this problem with $\psi_{\mathcal{X}} = \|\cdot\|_2^2$ and $\psi_{\mathcal{Y}}$ =negative entropy. Theorem 2 implies an $\frac{B\sqrt{\log m}}{T}$ rate.

The above algorithm can be adapted to compute the closely related Support Vector Machine (SVM) classifier.