IE 598: Big Data Optimization **Fall 2016**

Lecture 18: Mirror-Prox Algorithm – October 25

Lecturer: Niao He *Scriber: Meghana Bande*

Overview: In this lecture, the mirror-prox algorithm is introduced to solve non-smooth convex functions. This involves conversion of the minimization problem to a convex concave saddle point problem. The convergence of the algorithm is also discussed.

18.1 Introduction

Nonsmooth convex optimization. Previously, we have looked at optimization problems of the form $\min_{x\in X} f(x)$, where X is a convex compact set and the function f is convex but non-smooth. We have discussed two approaches to solve this problem.

1. Subgradient descent or the more general version, Mirror descent can be used to solve this problem. These approaches give a rate of convergence of $O(\frac{1}{\sqrt{t}})$, which is indeed optimal among all subgradient-based algorithms.
2. Smoothing approach (e.g., Nesterov's smoothing technique) is to exploit the structure of the function f to find a smooth approximation f_μ and use accelerated gradient descent to solve the optimization problem for this smooth function. These approaches give a convergence rate of $O(\frac{1}{t})$.

Drawbacks of smoothing techniques. The drawbacks with the smoothing techniques discussed previously are the following:

1. The performance of the algorithm is very sensitive to the smoothness parameter μ. The optimal choice of $\mu \sim O(\frac{\epsilon}{D_Y^2})$ cannot be calculated since D_Y and ϵ may not be known. Using a large μ may lead to a bad approximation of the original function f while using a smaller μ may result in slower convergence of the algorithm used.
2. The approach uses the gradient of $\nabla f_\mu(x)$ or a prox operator which involves solving the optimization problem $\min_{x\in X}\{f(x) + \frac{1}{2\mu}||x-y||^2\}$. This can be expensive to calculate in many scenarios.

In this lecture, we will discuss the **mirror-prox method** which does not use any smoothing parameter μ.

If f can be represented as $\max_{y\in Y} \{\langle Ax+b, y\rangle - \phi(y)\}$, instead of solving a smooth approximate function, we can directly solve the minimax function i.e.,

$$\min_{x\in X} f(x) \iff \min_{x\in X} \max_{y\in Y} \{\langle Ax+b, y\rangle - \phi(y)\}.$$

Recall that we encountered minimax problems previously when we used Lagrangian dual to solve constrained optimization problems.

$$\min_{x\in X, g(x)\le 0} f(x) \iff \min_{x\in X} \max_{\lambda\ge 0} \{f(x) + \lambda^T g(x)\}.$$

In this Lagrangian setting, we discussed that i) saddle point exists if the slater condition is satisfied, and ii) if saddle point exists, then the corresponding x solves the primal problem.

18.2 Smooth Convex-Concave Saddle Point Problems

Consider the saddle point problem

$$\min_{x\in X}\max_{y\in Y}\phi(x,y), \tag{18.1}$$

under the following assumptions,

1. For each $y \in Y$, the function $\phi(x,y)$ is convex in the variable x and for each $x \in X$, the function $\phi(x,y)$ is concave in the variable y.
2. The sets X, Y are closed, convex sets.
3. $\phi(x,y)$ is a smooth function, i.e., $\nabla\phi(x,y) = [\nabla_x\phi(x,y), \nabla_y\phi(x,y)]$ is Lipschitz continuous on the domain of $X \times Y$.

A feasible point (x_*, y_*) for (18.1) is a saddle point if

$$\phi(x_*,y) \le \phi(x_*,y_*) \le \phi(x,y_*) \quad \forall x \in X, y \in Y.$$

Lemma 18.1 (Sion's minimax theorem, existence of saddle point) *If one of sets X, Y is bounded then the saddle point to* (18.1) *always exists.*

We now consider the primal and dual optimization problems induced by the convex concave saddle point problem (18.1).

$$\mathrm{Opt(P)} = \min_{x\in X}\bar{\phi}(x), \quad \bar{\phi}(x) = \max_{y\in Y}\phi(x,y) \quad (P)$$

$$\mathrm{Opt(D)} = \max_{y\in Y}\underline{\phi}(y), \quad \underline{\phi}(y) = \min_{x\in X}\phi(x,y) \quad (D)$$

If (x_*, y_*) is the saddle point of (18.1), then x_* is the optimal solution to (P) and y_* is the optimal solution to (D), i.e., we have,

$$\bar{\phi}(x_*) = \mathrm{Opt(P)} = \phi(x_*,y_*) = \mathrm{Opt(D)} = \underline{\phi}(y_*).$$

Given a candidate solution $z = (x,y)$, we quantify the inaccuracy or error by $\epsilon_{\mathrm{sad}}(z)$ defined as

$$\epsilon_{\mathrm{sad}}(z) = \bar{\phi}(x) - \underline{\phi}(y).$$

We note that for all $z \in X \times Y$, $\epsilon_{\mathrm{sad}}(z) \ge 0$, and $\epsilon_{\mathrm{sad}}(z) = 0$ iff z is the saddle point.

Since $\mathrm{Opt(P)} = \mathrm{Opt(D)}$, $\epsilon_{\mathrm{sad}}(z)$ can be written as

$$\epsilon_{\mathrm{sad}}(z) = \bar{\phi}(x) - \mathrm{Opt(P)} + \mathrm{Opt(D)} - \underline{\phi}(y),$$

and hence we have,

$$\bar{\phi}(x) - \mathrm{Opt(P)} \le \epsilon_{\mathrm{sad}}(z),$$

$$\mathrm{Opt(D)} - \underline{\phi}(y) \le \epsilon_{\mathrm{sad}}(z).$$

18.3 Examples

We present a few examples to illustrate the conversion of a non-smooth minimization to a smooth convex concave saddle point problem. In each of these examples, we assume that the set X is a closed convex set.

1. $f(x) = \max\limits_{1\leq i\leq m} f_i(x)$ where each $f_i(x)$ is smooth and convex for all $1 \leq i \leq m$. Note that $f(x)$ is the maximum of convex functions and is typically non-smooth.

   This can be written as $f(x) = \max\limits_{y\in\Delta_m} \sum_{i=1}^m y_i f_i(x)$, where the simplex $\Delta_m = \{y : y \geq 0, \sum y_i = 1\}$ is a compact convex set and the function $\phi(x, y)$ is given by

   $$\phi(x, y) = \sum_{i=1}^{m} y_i f_i(x).$$

   Note that $\phi(x, y)$ is a smooth function since each f_i is smooth and it is concave (linear) in y for any $x \in X$ and convex in x for any fixed $y \in \Delta_m$.

2. $f(x) = ||Ax - b||_p$ where $||.||_p$ denotes the p-norm given by $||x||_p = (\sum_{i=1}^n x_i^p)^{\frac{1}{p}}$. The function $f(x)$ is convex but non-smooth because it is not differentiable at zero.

   This can be written as $f(x) = \max\limits_{||y||_q\leq 1} \langle Ax - b, y\rangle$. Here $Y = \{y : ||y||_q \leq 1\}$ is a unit q-norm ball, where q is such that $\frac{1}{p} + \frac{1}{q} = 1$ and is a compact convex set and the function $\phi(x, y)$ is given by

   $$\phi(x, y) = \langle Ax - b, y\rangle.$$

   Note that $\phi(x, y)$ is a smooth function that is concave (linear) in y for any $x \in X$ and convex (linear) in x for any fixed $y \in Y$.

   If $p = 1$, we have the case of robust regression and $q = \infty$ in this case. If $p = 2$, we have least squares regression and in this case $q = 2$.

3. $f(x) = \sum_{i=1}^m \max(1 - (a_i^T x)b_i), 0)$ is a convex piecewise linear function which is non-smooth. This is the hinge loss function used widely in support vector machines.

   This can be written as $f(x) = \max\limits_{0\leq y_i\leq 1} \sum_{i=1}^m y_i(1 - (a_i^T x)b_i)$, where the set $Y = \{y : 0 \leq y_i \leq 1, 1 \leq i \leq m\}$ is a compact convex set and the function $\phi(x, y)$ is given by

   $$\phi(x, y) = \sum_{i=1}^{m} y_i(1 - (a_i^T x)b_i).$$

   Note that $\phi(x, y)$ is a smooth function that is concave (linear) in y for any $x \in X$ and convex (linear) in x for any fixed $y \in Y$.

18.4 Mirror-Prox Algorithm

18.4.0 High-level Idea.

If we have access to the gradient of the function $\phi(x, y)$, we can use gradient descent type algorithms to solve the saddle point problem, just like what we do for convex minimization problem.

Consider the "gradient type" vector field $F(z)$ defined for each $z = (x, y)$ as

$$F(z) = [\nabla_x \phi(x, y), -\nabla_y \phi(x, y)].$$

Note that since $\phi(x, y)$ is convex in x and concave in y, $\nabla_x \phi(x, y), -\nabla_y \phi(x, y)$ are descent directions.

It can be shown that the first order optimality condition for the saddle point problem (18.1) is given by

$$z_* \text{ is optimal} \iff \langle F(z_*), z - z_* \rangle \geq 0, \quad \forall z \in X \times Y.$$

This is similar to the optimality condition for convex minimization problem where F stands for the gradient or subgradient. Intuitively, we could apply mirror descent algorithm to solve (18.1) as if we were solving a convex minimization problem, by replacing the subgradient with the above vector field. That is, at each iteration, we run

$$z_{t+1} = \operatorname*{argmin}_{z \in X \times Y} \{V(z, z_t) + \langle \gamma_t F(z_t), z \rangle\}, \tag{18.2}$$

where $V(z, z_t)$ is some Bregman distance defined on $X \times Y$. Extending the analysis we have earlier on mirror descent, we can show that $\epsilon_{\text{sad}}(z) \leq O(\frac{1}{\sqrt{t}})$, which implies a slow $O(1/t)$ rate of convergence, similar as what we obtain when using mirror descent to solve convex minimization problems. In the following, we show that with a slight modification of Mirror Descent, we can achieve the $O(\frac{1}{t})$ convergence rate, matching the results given by Nesterov's smoothing technique.

18.4.1 Mirror Prox

Setup. Let $\omega(z) : X \times Y \to \mathbb{R}$ be a distance generating function where ω is 1-strongly convex function w.r.t some norm $||.||$ on the underlying space and is continuously differentiable. The Bregman distance induced by $\omega(\cdot)$ is given as

$$V(z, z') = \omega(z) - \omega(z') - \nabla\omega(z')^T(z - z') \geq \frac{1}{2}||z - z'||^2.$$

Recall we also have the Bregman three-point identity which states that for any $x, y, z \in \text{dom}(\omega)$, we have

$$V(x, z) = V_\omega(x, y) + V_\omega(y, z) - \langle \nabla\omega(z) - \nabla\omega(y), x - y \rangle.$$

We assume that the vector field $F(z) = [\nabla_x \phi(x, y), -\nabla_y \phi(x, y)]$ is Lipschitz continuous with respect to the norm $\| \cdot \|$, namely,

$$\|F(z) - F(z')\|_* \leq L\|z - z'\|$$

where $\| \cdot \|_*$ is the dual norm of $\| \cdot \|$.

Mirror Prox algorithm:

- Initialization:

$$z_1 = (x_1, y_1) \in X \times Y.$$

- Update at each iteration t:

$$\hat{z}_t = (\hat{x}_t, \hat{y}_t) = \operatorname*{argmin}_{z \in X \times Y} \{V(z, z_t) + \langle \gamma_t F(z_t), z \rangle\}, \tag{18.3}$$

$$z_{t+1} = (x_{t+1}, y_{t+1}) = \operatorname*{argmin}_{z \in X \times Y} \{V(z, z_t) + \langle \gamma_t F(\hat{z}_t), z \rangle\}. \tag{18.4}$$

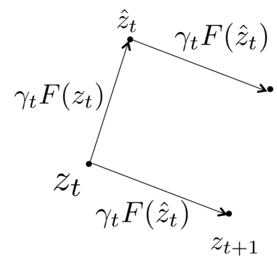

Figure 18.1:

Note that this is not the same as two consecutive steps of the mirror descent algorithm since the first term in the minimization, $V(z, z_t)$ is same in both the steps of the update. This is illustrated in Figure 18.1.

Theorem 18.2 (NEM '04) *Denote the diameter of the Bregman distance* $\Omega = \max_{z \in X \times Y} V(z, z_1)$*. The Mirror Prox algorithm with step-size* $\gamma_t \leq \frac{1}{L}$ *satisfies*

$$\epsilon_{sad}(\bar{z}_T) \leq \frac{\Omega}{\sum_{t=1}^T \gamma_t}, \quad \textit{where } \bar{z}_T = \frac{\sum_{t=1}^T \gamma_t \hat{z}_t}{\sum_{t=1}^T \gamma_t}.$$

Proof: From the Bregman three-point identity and the optimality condition for $\hat{z}_t$ to be the solution of (18.3), we have

$$\langle \gamma_t F(z_t), \hat{z}_t - z \rangle \leq V(z, z_t) - V(z, \hat{z}_t) - V(\hat{z}_t, z_t), \quad \forall z \in X \times Y. \tag{18.5}$$

Similarly optimality at z_{t+1} for (18.4) gives

$$\langle \gamma_t F(\hat{z}_t), z_{t+1} - z \rangle \leq V(z, z_t) - V(z, z_{t+1}) - V(z_{t+1}, z_t), \quad \forall z \in X \times Y. \tag{18.6}$$

Set $z = z_{t+1}$ in (18.5) to obtain

$$\langle \gamma_t F(z_t), \hat{z}_t - z_{t+1} \rangle \leq V(z_{t+1}, z_t) - V(z_{t+1}, \hat{z}_t) - V(\hat{z}_t, z_t). \tag{18.7}$$

Combing (18.6) and (18.7), we have

$$\begin{aligned}
\langle \gamma_t F(\hat{z}_t), \hat{z}_t - z \rangle &= \langle \gamma_t F(\hat{z}_t), \hat{z}_t - z_{t+1} \rangle\} + \langle \gamma_t F(\hat{z}_t), z_{t+1} - z \rangle \\
&= \gamma_t \langle F(\hat{z}_t) - F(z_t), \hat{z}_t - z_{t+1} \rangle + \langle \gamma_t F(z_t), \hat{z}_t - z_{t+1} \rangle + \langle \gamma_t F(\hat{z}_t), z_{t+1} - z \rangle \\
&\leq \gamma_t \langle F(\hat{z}_t) - F(z_t), \hat{z}_t - z_{t+1} \rangle - V(z_{t+1}, \hat{z}_t) - V(\hat{z}_t, z_t) + V(z, z_t) - V(z, z_{t+1})
\end{aligned}$$

Let $\sigma_t = \gamma_t \langle F(\hat{z}_t) - F(z_t), \hat{z}_t - z_{t+1} \rangle - V(z_{t+1}, \hat{z}_t) - V(\hat{z}_t, z_t)$. By assumption of smoothness, we have $\|F(\hat{z}_t) - F(z_t)\|_* \leq L\|\hat{z}_t - z_t\|$. Invoking Cauchy-Schwatz inequality and the property of Bregman distance, $V(z, z') \geq \frac{1}{2}||z - z'||^2$, to obtain

$$\sigma_t \leq \gamma_t L||z_{t+1} - \hat{z}_t|| \cdot ||\hat{z}_t - z_t|| - \frac{1}{2}||z_{t+1} - \hat{z}_t||^2 - \frac{1}{2}||\hat{z}_t - z_t||^2.$$

Since $\gamma_t \leq 1/L$, we have $\sigma_t \leq 0$.

Thus we have

$$\langle \gamma_t F(\hat{z}_t), \hat{z}_t - z \rangle \leq V(z, z_t) - V(z, z_{t+1}).$$

Note that

$$\begin{aligned}\langle \gamma_t F(\hat{z}_t), \hat{z}_t - z\rangle &= \gamma_t[\langle \nabla_x \phi(\hat{x}_t, \hat{y}_t), \hat{x}_t - x\rangle + \langle -\nabla_y \phi(\hat{x}_t, \hat{y}_t), \hat{y}_t - y\rangle] \\ &\geq \gamma_t[\phi(\hat{x}_t, \hat{y}_t) - \phi(x, \hat{y}_t) + \phi(\hat{x}_t, y) - \phi(\hat{x}_t, \hat{y}_t)] \\ &= \gamma_t[\phi(\hat{x}_t, y) - \phi(x, \hat{y}_t)]\end{aligned}$$

where we have used that $\phi(x, \hat{y}_t)$ and $-\phi(\hat{x}_t, y)$ are convex and for any convex function f, $f(u) \geq f(v) + \langle \nabla f(v), u - v\rangle, \quad \forall u, v \in \mathrm{dom}(f)$.

Consider the sum up to T terms, and divide by $\sum_{t=1}^T \gamma_t$ to obtain

$$\frac{\sum_{t=1}^T \gamma_t[\phi(\hat{x}_t, y) - \phi(x, \hat{y}_t)]}{\sum_{t=1}^T \gamma_t} \leq \frac{V(z, z_1)}{\sum_{t=1}^T \gamma_t}.$$

Let $\bar{z}_T = (\bar{x}_T, \bar{y}_T) = \frac{\sum_{t=1}^T \gamma_t \hat{z}_t}{\sum_{t=1}^T \gamma_t}$, by convex-concavity of $\phi(x, y)$, this further implies,

$$\phi(\bar{x}_T, y) - \phi(x, \bar{y}_T) \leq \frac{V(z, z_1)}{\sum_{t=1}^T \gamma_t}, \quad \forall z = [x, y] \in X \times Y.$$

Taking the maximum over all $x \in X, y \in Y$, we obtain

$$\epsilon_{\text{sad}}(\bar{z}_T) = \bar{\phi}(\bar{x}_T) - \underline{\phi}(\bar{y}_T) \leq \max_{z \in X \times Y} \frac{V(z, z_1)}{\sum_{t=1}^T \gamma_t} = \frac{\Omega}{\sum_{t=1}^T \gamma_t}.$$

■

Remark. If the step-size is assumed to be constant, $\gamma_t = \frac{1}{L}$, then we have

$$\epsilon_{\text{sad}}(\bar{z}_T) \leq \frac{\Omega L}{T}.$$

Mirror Prox algorithm achieves a $O(1/T)$ rate of convergence using only first order information of $\phi(x, y)$.

Recall that

$$\epsilon_{\text{sad}}(z) = \bar{\phi}(x) - \mathrm{Opt(P)} + \mathrm{Opt(D)} - \underline{\phi}(y),$$

Hence, both primal and dual error is bounded by $O(1/T)$. That is, when solving a nonsmooth convex minimization problem

$$\min_{x \in X} f(x), \text{ where } f(x) = \max_{y \in Y} \phi(x, y)$$

the Mirror Prox algorithm attains the $O(1/T)$ rate, comparable to Nesterov's smoothing technique. Note that this algorithm does not require f to be simple in order to allow for easy omputation of proximal operators of f but instead only requires operator F which comes from ϕ. In this regard, it is more general than Nesterov's smoothing technique.

Beyond saddle point problems. Mirror-prox algorithm can be used to solve a large range of problems for which we have the knowledge of operator F irrespective of whether it comes from gradient of a convex minimization problem, a saddle point problem or any other optimization problems. Indeed, this algorithm has been used widely to solve convex minimization, saddle point problems, variational inequalities, and fixed point problems.

References

[NEM '04] 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.

[JN '11] Anatoli Juditsky, and Arkadi Nemirovski, First order methods for nonsmooth convex large-scale optimization, ii: utilizing problems structure, *Optimization for Machine Learning*, 149-183, 2011.