Lecture 16

16.1 Application Example: Min-Max Problems

This lecture, we study general techniques, that are built on top of the basic subgradient method that we have discussed few previous lectures. These techniques, called *mirror descent* or *dual averaging* are quite general and useful. However, to see their full power, we illustrate our developments with the problems of the following structure.

16.1.1 General Min-Max Problem

A general form of a convex optimization problem is as follows:

$$\min_{x \in Q} f(x), \tag{16.1}$$

where $Q \subseteq \mathbb{R}^n$ is a convex set and f is a convex function. We can apply first-order methods in a black-box manner directly to (16.1). However, in practice, we always know something more about the problem and the actual structure of the objective. Examples include:

- *Fully-composite problems* (see Lecture 9), where we can identify their *smooth components* and then apply efficient methods from smooth optimization, such as the fast gradient method. The main assumption is that non-smooth parts of the problem as *simple* (e.g. a non-smooth regularizer or simple constraints).
- *Finite-sum minimization.* When the function f in (16.1) is represented as a finite sum of smooth objectives, we can apply efficient stochastic methods with variance reduction, such as SVRG, that preserve fast rates of the deterministic methods.

Now, we consider another specific structure of the problem, called *min-max*, that is very popular in practice:

$$f(x) \;:=\; \max_{u \in \Omega} F(x, u), \tag{16.2}$$

where $\Omega \subset \mathbb{R}^m$ is a bounded convex set, and $F(\cdot, u)$ is convex for any u and $F(x, \cdot)$ is concave for any x.

Hence, our original problem is the min-max or *saddle point* problem:

$$\begin{aligned} \min_{x \in Q} f(x) \;&=\; \min_{x \in Q} \max_{u \in \Omega} F(x, u) \\ &\geq\; \max_{u \in \Omega} \min_{x \in Q} F(x, u) \;=:\; \max_{u \in \Omega} \varphi(u), \end{aligned}$$

where $\varphi(u) := \min_{x \in Q} F(x, u)$. We call the latter problem:

$$\max_{u \in \Omega} \varphi(u), \tag{16.3}$$

as the *adjoint* or *dual* problem to (16.1). Note that the same primal problem (16.1) can have different min-max representations of the objective, and therefore the corresponding adjoint problems (16.3) depends on this representation and is not uniquely defined.

In most cases, we have so called *strong duality* (e.g., when both sets are compact and F is convex-concave and continuous) so these two problems are mathematically equivalent and symmetric:

$$\min_{x \in Q} f(x) \quad = \quad \max_{u \in \Omega} \varphi(u).$$

However, in practice, one of these problems can be much easier to solve than the other, due to different dimensionality, $x \in Q \subseteq \mathbb{R}^n$ and $u \in \Omega \subseteq \mathbb{R}^m$ and specific structure of F.

Since our initial interest was the primal problem (16.1), we assume that we can efficiently compute the *first-order oracle* for f along with the following information, for any given point $x \in Q$:

- Function value: $f(x) := \max\limits_{u \in \Omega} F(x, u)$;
- A minimizer: $u(x) := \arg\max\limits_{u \in \Omega} F(x, u)$;
- A subgradient: $f'(x) = F'_x(x, u(x)) \in \partial_x F(x, u(x))$;

where F'_x is a subgradient of F with respect to the first variable x, and $\partial_x F$ is the partial subdifferential (the set of all such subgradients). In practice, access to $u(x)$ is almost always available, once we assume an efficient way of computing the function value $f(x)$ itself.

It appears that having a method that solves the primal problem (16.1), we can also automatically generate solutions for the adjoint problem (16.3), that we will show in the next lecture.

16.1.2 Matrix Games

The simplest example of the previous setting is the following objective,

$$F(x, u) \quad = \quad \langle Ax, u \rangle + \langle b, u \rangle + \langle c, x \rangle, \tag{16.4}$$

where $A \in \mathbb{R}^{m \times n}$ is a given matrix, $b \in \mathbb{R}^m$ and $c \in \mathbb{R}^n$ are some vectors.

We assume that we have two players x and u. The *strategy* of every player belongs to the corresponding simplex:

$$x \quad \in \quad Q \quad = \quad \Delta_n \quad = \quad \{x \in \mathbb{R}^n_{\geq 0} \quad : \quad \sum_{i=1}^{n} x^{(i)} = 1\}$$

and

$$y \quad \in \quad \Omega \quad = \quad \Delta_m \quad = \quad \{y \in \mathbb{R}^m_{\geq 0} \quad : \quad \sum_{i=1}^{m} y^{(i)} = 1\}.$$

Therefore, our problem is the following one:

$$\min_{x \in \Delta_n} \max_{u \in \Delta_m} \Big[\langle Ax, u \rangle + \langle b, u \rangle + \langle c, x \rangle \Big]$$

It corresponds to choosing the best strategy of playing for the first player, which minimizes their loss under *any strategy* of the second player. In other words, we want to win as much as possible in the worst-case scenario, when our opponent plays the best (maximizing our loss).

Then, in the terminology of primal problem (16.1), our objective is

$$\begin{aligned} f(x) &= \max_{u \in \Delta_m} \Big[\langle Ax, u \rangle + \langle b, u \rangle \Big] + \langle c, x \rangle \\ &= \max_{u \in \Delta_m} \sum_{i=1}^{m} u^{(i)} \big[\langle a_i, x \rangle + b_i \big] + \langle c, x \rangle \\ &= \max_{1 \le i \le m} \big[\langle a_i, x \rangle + b_i \big] + \langle c, x \rangle, \end{aligned} \tag{16.5}$$

where $a_1, \ldots, a_m \in \mathbb{R}^n$ are the rows of our matrix A.

Therefore, to compute the function value $f(x)$, we need to find the variable

$$u(x) \;=\; e_i \;\in\; \Delta_m$$

with $1 \le i \le m$ such that $f(x) = \langle a_i, x \rangle + b_i + \langle c, x \rangle$. The corresponding subgradient can be set as follows:

$$f'(x) \;=\; a_i.$$

16.1.3 Performance of the Subgradient Method

In previous lectures, we analyzed the following subgradient method that we can directly apply to primal problem (16.1),

$$x_{k+1} \;=\; \pi_Q(x_k - \eta_k f'(x_k)) \tag{16.6}$$

and proved the following convergence guarantee:

$$f(\bar{x}_k) - f^\star \;\le\; \tfrac{MD}{\sqrt{K}}, \tag{16.7}$$

for example, using the normalized stepsizes, $\eta_k := \frac{\gamma}{\|f'(x_k)\|}$, or the constant $\eta_k \equiv \frac{\gamma}{M}$, where $\gamma := \frac{D}{\sqrt{K}}$ (Lecture 12), or, using the *adaptive stepsizes* (Lecture 14) that do not depend on the fixed number of iterations. Here,

$$D \;\ge\; \|x_0 - x^\star\|_2 \qquad \text{and} \qquad \|f'(x)\|_2 \;\le\; M, \qquad \forall x \in Q, \quad \forall f'(x) \in \partial f(x). \tag{16.8}$$

We also proved that the rate of (16.7) is *optimal* (Lecture 13), and the function from the lower bound construction actually matches the form of our problem (16.5). Therefore, it seems like the end of story.

How can the convergence result (16.7) be further improved?

- We do not know **when to stop the method**? Even though we can use adaptive stepsizes, that do not use a fixed number of iterations K in the method, we do not have a *computable stopping condition* in the algorithm, which would ensure that

  $$f(\bar{x}_K) - f^\star \;\le\; \varepsilon$$

  for a given accuracy $\varepsilon > 0$, unless we know $f^\star$.

- A related to the previous question: can we say anything about **solving the adjoint problem** (16.3)?

- Another crucial observation is that we **fix the geometry** by choosing $\|\cdot\|_2$ norm in the method (16.6): and this is the norm in which parameters M and R are measured in (16.8).

What can be wrong with $\|\cdot\|_2$ norm?

Example 16.1.1. Consider the objective (16.5) from the previous section. Hence $Q = \Delta_n$. Then,

$$\mathrm{diam}_{\|\cdot\|_2}(\Delta_n) \;=\; \|e_1 - e_2\|_2 \;=\; \sqrt{2}.$$

Now, assume that we have $f'(x) = a_i + c = (1, \ldots, 1)^\top \in \mathbb{R}^n$. Then

$$\|f'(x)\|_2 \;=\; \sqrt{n}.$$

Hence, $M_{\|\cdot\|_2} \geq \sqrt{n}$, and the convergence rate in (16.7) *does depend on the dimension!* If we increase n, the method will need significantly more time to solve the problem.

Example 16.1.2. For the same problem, let us choose $\|\cdot\|_1$ norm for the primal variables $x \in \Delta_n$. Then,

$$\mathrm{diam}_{\|\cdot\|_1}(\Delta_n) \;=\; \|e_1 - e_2\|_1 \;=\; 2,$$

which is an absolute constant again. However, to measure the size of the dual objects (subgradients), we use the dual norm, which is $\|\cdot\|_\infty$. In this norm, we have

$$\|f'(x)\|_\infty \;=\; \|(1, \ldots, 1)\|_\infty \;=\; 1.$$

Therefore, $M_{\|\cdot\|_\infty} \approx 1$ and it is much better than $M_{\|\cdot\|_2}$.

Note that we always have $\|\cdot\|_\infty \leq \|\cdot\|_2 \leq \|\cdot\|_1$ and hence $M_{\|\cdot\|_\infty} \leq M_{\|\cdot\|_2}$ always, while in a particular instance, $M_{\|\cdot\|_\infty}$ can be significantly better as in our example.

16.1.4 Why Gradient Method Is Not Enough?

We want to have a modification of the subgradient method that is more suitable to the problem geometry. We assume that we have a *primal space* of variables, which we denote by $\mathbb{R}^n$:

$$x \;\in\; Q \;\subseteq\; \mathbb{R}^n.$$

We use an arbitrary norm $\|\cdot\|$ in this space (not necessary Euclidean).

Then, we treat subgradients as linear forms: $\langle f'(x), \cdot \rangle$ which are *dual objects*, and we use the corresponding dual norm for them, defined by

$$\|s\|_* \;=\; \max_{h \in \mathbb{R}^n \,:\, \|h\| \leq 1} \langle s, h \rangle.$$

Recall that we already saw the gradient method for arbitrary norms (Lecture 3). For a smooth objective (that is, the gradient of f is Lipschitz w.r.t. a fixed norm), we can perform:

$$x_{k+1} \;=\; \arg\min_{y \in Q} \Big[f(x_k) + \langle \nabla f(x_k), y - x_k \rangle + \tfrac{L}{2} \|y - x_k\|^2 \Big], \tag{16.9}$$

and we can establish global convergence rate of $O(1/k)$ in terms of the functional residual, for this process.

The first issue is the constraint set Q. When the norm is Euclidean, iteration (16.9) can be rewritten in terms of the projection operation. However, in general, for an arbitrary norm, this is a nontrivial subproblem.

The second issue is that this analysis worked only for *smooth function* (but possibly non-convex). Performing iterations of form (16.9), we can ensure a positive progress of each step:

$$f_k - f_{k+1} \;\geq\; \tfrac{1}{2LD^2} f_k^2, \tag{16.10}$$

leading to the desired rate. These iterations are based on *local relaxation*.

In non-smooth optimization ($L \to \infty$), we are not able to establish the local improvement for our objective (16.10). The methods of non-smooth convex optimization are based on the idea of building a *global model* of the objective (or building a *localizer set* that containts a solution). Our analysis was substantially based on the algebraic properties of the Euclidean norm, in particular, on

$$\textit{strong convexity of the regularizer } \tfrac{1}{2}\|x\|_2^2.$$

However, an arbitrary norm is not strongly convex in general.

Exercise 16.1.1. Consider $d(x) = \|x\|_1^2$, $x \in \mathbb{R}^n$, and show that it is not strongly convex for $n > 1$.

16.2 Arbitrary Regularizers

16.2.1 Bregman Divergence

The key idea of the *mirror descent* algorithm is to replace the square norm $\|\cdot\|^2$ by an arbitrary distance function $d : \operatorname{int} Q \to \mathbb{R}$.

The main assumption is that d is *simple* and at least *strictly convex* differentiable function. It is also convenient to define the following associated *Bregman Divergence*:

$$\beta_d(x; y) \;=\; d(y) - d(x) - \langle \nabla d(x), y - x \rangle \;>\; 0, \qquad x \neq y. \tag{16.11}$$

Strict convexity of d means that the inequality in (16.9) is strict: ">". Geometrically, $g(y) := \beta_d(x; y)$ is a "rotation" of our regularizer d such that it is minimum in x.

Example 16.2.1. Let $d(x) = \frac{1}{2}\|x\|_2^2$ (squared Euclidean norm). Then $x_0 = 0$. We have

$$\begin{aligned} \beta_d(x; y) &= \tfrac{1}{2}\|y\|_2^2 - \tfrac{1}{2}\|x\|_2^2 - \langle x, y - x \rangle \\ &= \tfrac{1}{2}\|y - x\|^2. \end{aligned}$$

This is exactly the regularized that we used in the construction of the subgradient method.

The most popular and important example is the following one.

Example 16.2.2. Let $d(x) = \sum\limits_{i=1}^{n} x^{(i)} \ln x^{(i)}$ (negative Entropy). Then,

$$x_0 \;=\; (\tfrac{1}{n}, \ldots, \tfrac{1}{n}) \;\in\; \Delta_n,$$

and $d(x_0) = -\ln n$. We have

$$[\nabla d(x)]^{(i)} \;=\; 1 + \ln x^{(i)}.$$

Hence, the Bregman divergence is equal to

$$\begin{aligned} \beta_d(x; y) &= \sum_{i=1}^{n} y^{(i)} \ln y^{(i)} - \sum_{i=1}^{n} x^{(i)} \ln x^{(i)} - \sum_{i=1}^{n} (1 + \ln x^{(i)})(y^{(i)} - x^{(i)}) \\ &= \sum_{i=1}^{n} y^{(i)} \ln y^{(i)} - \sum_{i=1}^{n} x^{(i)} \ln x^{(i)} - \sum_{i=1}^{n} \ln x^{(i)} (y^{(i)} - x^{(i)}) \\ &= \sum_{i=1}^{n} y^{(i)} \ln \tfrac{y^{(i)}}{x^{(i)}}. \end{aligned}$$

In statistics, it is called Kullback–Leibler or KL divergence between probability distributions x and y.

Note that the second derivative of d is the diagonal matrix:

$$[\nabla f(x)]^{(i,i)} \;=\; \tfrac{1}{x^{(i)}} \;>\; 0, \qquad x \;\in\; \operatorname{int}\Delta_n \;:=\; \Big\{x \in \mathbb{R}^n_{>0} \;:\; \sum_{i=1}^n x^{(i)} = 1\Big\}$$

Hence, we know that d is strictly convex and thus $\beta_d(x;y) > 0$ for $x \neq y$. It appears that we can improve this inequality. Indeed, for any $x \in \operatorname{int}\Delta_n$ and $h \in \mathbb{R}^n$:

$$\langle \nabla^2 d(x)h, h\rangle \;=\; \sum_{i=1}^n \frac{\left(h^{(i)}\right)^2}{x^{(i)}}.$$

Then, using Cauchy-Schwarz inequality, we observe that

$$\|h\|_1 \;=\; \sum_{i=1}^n |h^{(i)}| \;=\; \sum_{i=1}^n \frac{|h^{(i)}|}{\sqrt{x^{(i)}}} \cdot \sqrt{x^{(i)}} \;\leq\; \Big(\sum_{i=1}^n \frac{|h^{(i)}|^2}{x^{(i)}}\Big)^{1/2} \cdot \Big(x^{(i)}\Big)^{1/2} \;=\; \langle \nabla^2 d(x)h, h\rangle^{1/2}.$$

Hence, $d(\cdot)$ is strongly convex with respect to $\|\cdot\|_1$ norm, and we have

$$\beta_d(x;y) \;\geq\; \tfrac{1}{2}\|y-x\|_1^2.$$

16.2.2 Main Lemma

To analyze methods with arbitrary regularizers, we need the following simple lemma.

Let $\psi : Q \to \mathbb{R}$ be a convex function and $d : Q \to \mathbb{R}$ is a convex regularizer, both defined on an open convex set $Q \subset \mathbb{R}^n$. We can assume for simplicity that both ψ and d are differentiable, which will be sufficient for the analysis of the mirror descent, while this assumption can be relaxed. Consider the regularized objective and denote its minimum by

$$x^+ \;:=\; \arg\min_{y \in Q}\Big[\, g(y) \;:=\; \psi(y) + d(y)\,\Big],$$

assuming that it exists.

Lemma 16.2.3. *We have*

$$g(y) \;\geq\; g(x^+) + \beta_d(x^+;y), \qquad \forall y \in Q. \tag{16.12}$$

Proof. Note that

$$\beta_g(x;y) \;=\; \beta_\psi(x;y) + \beta_d(x;y) \;\geq\; \beta_d(x;y).$$

Rearranging the left hand side, we get

$$g(y) \;\geq\; g(x) + \langle \nabla g(x), y - x\rangle + \beta_d(x;y).$$

Substituting $x := x^+$ and using the optimality condition: $\langle \nabla g(x^+), y - x^+\rangle \geq 0$, for all $y \in Q$ (see Corollary 12.1.2 in Lecture 12), completes the proof. □

Remark 16.2.4. Note that the extra non-negative term $\beta_d(x^+;y)$ in (16.12) is an improvement of the trivial inequality: $g(y) \geq g(x^+)$. It is remarkable that we do not need any specific properties of g and d, such as strong convexity, to prove (16.12).