Lecture 17

17.1 Summary: Bregman Divergence

We consider the following optimization problem,

$$\min_{x \in Q} f(x), \tag{17.1}$$

where $Q \subset \mathbb{R}^n$ is a convex set and $f : Q \to \mathbb{R}$ is a convex function.

We denote by $\|\cdot\|$ an arbitrary general norm for the primal space $\mathbb{R}^n$, and the corresponding dual norm $\|\cdot\|_*$ for measuring the subgradients.

Our main complexity parameter is constant $M > 0$:

$$\|f'(x)\|_* \;\leq\; M, \qquad \forall x \in \operatorname{int} Q, \;\; \forall f'(x) \in \partial f(x).$$

We believe that using a non-Euclidean norm can be better to capture geometry of the problem.

To develop a subgradient method for an arbitrary norm, we introduce a *distance function*

$$d : \operatorname{int} Q \to \mathbb{R}$$

that measures distances between points in Q. We will use this function as a regularizer in our method. The main assumptions for this function is that it is

- *simple* — a regularizer should *help* us solving problem (17.1) rather than complicate the problem;
- at least *strictly convex*; ideally *strongly convex* with respect to our primal norm $\|\cdot\|$.

We denote the minimum of the distance function by

$$x_0 \;=\; \arg\min_{y \in Q} d(y), \tag{17.2}$$

the point from which we will start our method, a certain *center of set* Q. Thus, by default, $d(y)$ measures "how far a point $y \in \operatorname{int} Q$ from the center x_0". To measure distances between any two points $x, y \in \operatorname{int} Q$, we introduce the *Bregman divergence*:

$$\beta_d(x; y) \;:=\; d(y) - d(x) - \langle \nabla d(x), y - x \rangle.$$

Note that in general $\beta_d(x; y)$ is not symmetric: $\beta_d(x; y) \neq \beta_d(y; x)$. The property that d strictly convex implies that

$$\beta_d(x; y) \;>\; 0, \qquad x \neq y.$$

Finally, in the previous lecture, we have justified the following useful lemma.

Lemma 17.1.1. *Let $g(y) := \psi(y) + d(y)$, where both ψ and d are convex and differentiable, defined on an open convex set Q. Consider $x^+ := \arg\min_{y \in Q} g(y)$, assuming that it exists. Then, we have*

$$g(y) \quad \geq \quad g(x^+) + \beta_d(x^+; y), \qquad y \in Q. \tag{17.3}$$

This inequality looks similar to strong convexity, but it is not required. At the same time, in our examples and analysis we will use that d is *strongly convex* :

$$\beta_d(x; y) \quad \geq \quad \tfrac{1}{2}\|x - y\|^2, \qquad \forall x, y \in \operatorname{int} Q. \tag{17.4}$$

17.2 Mirror Descent

The idea is to replace the squared Euclidean norm in the subgradient method by an arbitrary Bregman divergence. We obtain the following algorithm, called *mirror descent*:

$$\boxed{x_{k+1} \quad = \quad \arg\min_{y \in Q} \Big[\eta \langle f'(x_k), y - x_k \rangle + \beta_d(x_k; y) \Big]} \tag{17.5}$$

where $\eta > 0$ is a constant step-size, starting from x_0 defined by (17.2).

Assume for a moment that $Q \equiv \mathbb{R}^n$ (unconstrained minimization). Then, the stationary condition of one mirror descent step gives

$$\eta f'(x_k) + \nabla d(x_{k+1}) - \nabla d(x_k) \quad = \quad 0,$$

or, rearranging the terms,

$$\nabla d(x_{k+1}) \quad = \quad \nabla d(x_k) - \eta g_k, \quad \text{where} \quad g_k \quad = \quad f'(x_k) \quad \in \quad \partial f(x_k).$$

This formula explains the name of the method.

Example 17.2.1. Let $d(x) := \frac{1}{2}\|x\|_2^2$. Then, $\beta_d(x; y) = \frac{1}{2}\|y - x\|^2$ and one step of the method is the classic subgradient step with projection:

$$x_{k+1} \quad = \quad \pi_Q(x_k - \eta f'(x_k)).$$

Exercise 17.2.1. Let $Q = \Delta_n$ and $d(x) = \sum\limits_{i=1}^n x^{(i)} \ln x^{(i)}$ be the negative entropy. Then, one step of the method is as follows:

$$x_{k+1} \quad = \quad \arg\min_{y \in \Delta_n} \Big\{ \eta \langle g_k, y - x_k \rangle + \sum_{i=1}^n y^{(i)} \ln \tfrac{y^{(i)}}{x_k^{(i)}} \Big\}.$$

It can be written explicitly as the *multiplicative weight update*:

$$x_{k+1}^{(i)} \quad = \quad \frac{x_k^{(i)} \exp\big(-\eta g_k^{(i)}\big)}{\sum_{j=1}^n x_k^{(j)} \exp\big(-\eta g_k^{(j)}\big)}$$

Note that we do not need to perform projection to the simplex as point x_{k+1} already belongs to it, due to normalization.

17.2.1 Analysis

By our main Lemma, we have, for any $y \in Q$:

$$\begin{aligned}
\beta_d(x_k; y) + \eta\langle g_k, y - x_k\rangle &\geq \beta_d(x_k; x_{k+1}) + \eta\langle g_k, x_{k+1} - x_k\rangle + \beta_d(x_{k+1}; y) \\
&\geq \tfrac{1}{2}\|x_k - x_{k+1}\|^2 - \eta\|g_k\|_*\|x_{k+1} - x_k\| + \beta_d(x_{k+1}; y) \\
&\geq \min_{t>0}\Big\{\tfrac{t^2}{2} - \eta\|g_k\|_* t\Big\} + \beta_d(x_{k+1}; y) \\
&= -\tfrac{\eta^2\|g_k\|_*^2}{2} + \beta_d(x_{k+1}; y) \\
&\geq -\tfrac{\eta^2 M^2}{2} + \beta_d(x_{k+1}; y).
\end{aligned}$$

Therefore, for one step of the method, we have the following inequality:

$$\tfrac{\eta^2 M^2}{2} + \beta_d(x_k; y) - \beta_d(x_{k+1}; y) \quad \geq \quad \eta\langle g_k, x_k - y\rangle.$$

Telescoping this inequality for the first $k \geq 1$ iterations, we obtain

$$\begin{aligned}
\eta k \cdot \tfrac{1}{k}\sum_{i=0}^{k-1}\langle g_i, x_i - y\rangle &\leq k\tfrac{\eta^2 M^2}{2} + \beta_d(x_0; y) - \beta_d(x_{k+1}; y) \\
&\leq k \cdot \tfrac{\eta^2 M^2}{2} + \beta_d(x_0; y).
\end{aligned}$$

Now, let us define

$$\boxed{\mathrm{Gap}_K \quad := \quad \max_{y\in Q} \tfrac{1}{K}\sum_{i=0}^{K-1}\langle g_i, x_i - y\rangle} \tag{17.6}$$

and

$$\boxed{D^2 := 2 \cdot \max_{y\in Q} \beta_d(x_0; y).}$$

This is a "diameter" of the set Q measure by our distance function.

We have proved the following bound for the new accuracy measure.

Theorem 17.2.2. *For any $\eta > 0$:*

$$\mathrm{Gap}_K \quad \leq \quad \tfrac{D^2}{2\eta K} + \tfrac{\eta M^2}{2}.$$

By choosing $\boxed{\eta := \tfrac{D}{M\sqrt{K}}}$ *we obtain*

$$\mathrm{Gap}_K \quad \leq \quad \tfrac{MD}{\sqrt{K}}. \tag{17.7}$$

We can compute the new accuracy measure (17.6) within iterations of our method, as it requires the minimization of a linear function over the set Q. This subproblem is easier that computing one step of the method (17.5). Thus, (17.6) is the accuracy certificate that we can use to stop our method:

$$\mathrm{Gap}_K \quad \leq \quad \varepsilon,$$

at it serves us an upper bound for the functional residual.

17.3 Accuracy Certificates

How to relate Gap_K to the functional residual?

- **Convex Functions:**

$$\begin{aligned} \mathrm{Gap}_K &\geq \tfrac{1}{K} \sum_{i=0}^{K-1} \langle f'(x_i), x_k - x^\star \rangle \geq \tfrac{1}{K} \sum_{i=0}^{K-1} \Big[f(x_i) - f^\star \Big] \\ &\geq f(x_k) - f^\star, \end{aligned}$$

  where $x_K := \frac{1}{K} \sum_{i=0}^{K-1} x_i$.

- **Online Optimization:** in online optimization, we have a stream of functions: a function $f_k(x)$ at iteration k, and we observe the corresponding subgradient $g_k := f_k'(x_k)$. Note that our analysis of mirror descent *did not use anything* about vectors g_k, and thus the result of Theorem 17.2.2 holds in this setting as well. In online optimization, the quantity Gap_K is usually called *regret*.

- **Min-Max Structure.** As in the previous lecture, consider the following min-max problem:

$$f^\star = \min_{x \in Q} \Big[f(x) := \max_{u \in \Omega} F(x,u) \Big],$$

  where $\Omega \subset \mathbb{R}^m$ is a bounded convex set, and F is a continuous function, convex in x and concave in u.

  Denote $u(x) := \arg\max_{u \in \Omega} F(x,u)$. Then,

$$\begin{aligned} f(x) &= F(x, u(x)), \\ f'(x) &= F'_x(x, u(x)). \end{aligned}$$

  This is the primal problem. The corresponding adjoint / dual problem is

$$\varphi_\star = \max_{u \in \Omega} \Big[\varphi(u) := \min_{x \in Q} F(x,u) \Big],$$

  which can potentially be much harder (or easier) to solve than the primal one. We observe that $f^\star \geq \varphi_\star$.

  Then,

$$\begin{aligned} \mathrm{Gap}_K &:= \max_{y \in Q} \tfrac{1}{K} \sum_{i=0}^{K-1} \langle f'(x_i), x_i - y \rangle = \max_{y \in Q} \tfrac{1}{K} \sum_{i=0}^{K-1} \langle F'_x(x, u(x)), x_i - y \rangle \\ &\geq \max_{y \in Q} \tfrac{1}{K} \sum_{i=0}^{K-1} [F(x_i, u(x_i)) - F(y, u(x_i))] = \max_{y \in Q} \tfrac{1}{K} \sum_{i=0}^{K-1} [f(x_i) - F(y, u(x_i))] \\ &\geq \max_{y \in Q} \Big[f(\bar{x}_K) - F(y, \bar{u}_K) \Big] = f(\bar{x}_K) - \varphi(\bar{u}_K), \end{aligned}$$

where in the last inequality we used convexity of f and concavity of F with respect to the second argument (Jensen's inequality), denoting the average primal point:

$$\bar{x}_K \;=\; \tfrac{1}{K}\sum_{i=0}^{K-1} x_i$$

and the average dual point:

$$\bar{u}_K \;=\; \tfrac{1}{K}\sum_{i=0}^{K-1} u(x_i). \tag{17.8}$$

Hence, we can ensure convergence not only in terms of the primal residual, but in terms of the dual residual as well:

$$\begin{aligned} \tfrac{MD}{\sqrt{K}} \;\overset{(17.7)}{\geq}\; & \mathrm{Gap}_K \;\geq\; f(\bar{x}_K) - \varphi(\bar{u}_K) \\ =\; & \underbrace{f(\bar{x}_K) - f^\star}_{\geq 0} \;+\; \underbrace{\varphi_\star - \varphi(\bar{u}_K)}_{\geq 0} \;+\; \underbrace{f^\star - \varphi_\star}_{\geq 0}. \end{aligned} \tag{17.9}$$

Note that our algorithm is initially designed to a *primal problem only*, but automatically solves the dual problem as well, where the dual solution at each iteration can be computed by formula (17.8).

It is remarkable that by setting $K \to \infty$ in (17.9), we obtain an *algorithmic proof* of *strong duality* for our min-max problem:

$$\boxed{f^\star \;=\; \varphi_\star}$$

Literature

The mirror descent algorithm was developed by Nemirovski and Yudin in [NY83].

[NY83] Arkadi Nemirovski and David Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience, 1983.