Lecture 11: Mirror Descent and Variable Metric Methods

May 8 - 13, 2020
1. Introduction  

In the prior lecture, we studied projected subgradient methods for solving the problem (2.1.1) by iteratively updating \( x_{k+1} = \pi_C(x_k - \alpha_k g_k) \), where \( \pi_C \) denotes Euclidean projection. The convergence of these methods, as exemplified by Corollaries 3.2.8 and 3.4.11, scales as

\[
(4.1.1) \quad f(\bar{x}_K) - f(x^*) \leq \frac{MR}{\sqrt{K}} = O(1) \frac{\text{diam}(C)\text{Lip}(f)}{\sqrt{K}},
\]

where \( R = \sup_{x \in C} \|x - x^*\|_2 \) and \( M \) is the Lipschitz constant of \( f \) over the set \( C \) with respect to the \( \ell_2 \)-norm,

\[
M = \sup_{x \in C} \sup_{g \in \partial f(x)} \left\{ \|g\|_2 = \left( \sum_{j=1}^{n} g_j^2 \right)^{\frac{1}{2}} \right\}.
\]

The convergence guarantee (4.1.1) reposes on Euclidean measures of scale—the diameter of \( C \) and norm of the subgradients \( g \) are both measured in \( \ell_2 \)-norm. It is thus natural to ask if we can develop methods whose convergence rates depend on other measures of scale of \( f \) and \( C \), obtaining better problem-dependent behavior and geometry. With that in mind, in this lecture we derive a number of methods that use either non-Euclidean or adaptive updates to better reflect the geometry of the underlying optimization problem.
**Definition:** Let $h$ be a differentiable convex function, differentiable on $C$. The *Bregman divergence* associated with $h$ is defined as

$$D_h(x, y) = h(x) - h(y) - \langle \nabla h(y), x - y \rangle.$$  

- $D_h(x, y)$ is always non-negative and convex in its first argument $x$.  

---

Mirror Descent and Variable Metrics  
DAMC Lecture 11  
May 8 - 13, 2020  
3 / 26
• $h$ is $\lambda$-strongly convex over $C$ with respect to the norm $\| \cdot \|$ if

$$h(x) \geq h(y) + \langle \nabla h(y), x - y \rangle + \frac{\lambda}{2} \| x - y \|^2 \quad \forall x, y \in C.$$  

Note that strong convexity of $h$ is equivalent to

$$D_h(x, y) \geq \frac{\lambda}{2} \| x - y \|^2 \quad \forall x, y \in C.$$  

2. Mirror Descent method

• Compute subgradient $g^k \in \partial f(x^k)$

• Update

$$x^{k+1} = \arg\min_{x \in C} \left\{ f(x^k) + \langle g^k, x - x^k \rangle + \frac{1}{\alpha_k} D_h(x, x^k) \right\}$$  

$$= \arg\min_{x \in C} \left\{ \langle g^k, x \rangle + \frac{1}{\alpha_k} D_h(x, x^k) \right\}.$$
We often attempt to choose \( h \) to better match the geometry of the underlying constraint set \( \mathcal{C} \). The goal is, roughly, to choose a strongly convex function \( h \) so that the diameter of \( \mathcal{C} \) is small in the norm \( \| \cdot \| \) (need not be \( \ell_2 \) or Euclidean norm) with respect to which \( h \) is strongly convex.

**Example:** Gradient descent is mirror descent. Consider

\[
h(x) = \frac{1}{2}\|x\|^2, \quad D_h(x, y) = \frac{1}{2}\|x - y\|^2.
\]

The update of mirror descent is

\[
x^{k+1} = \arg\min_{x \in \mathcal{C}} \left\{ f(x^k) + \langle g^k, x - x^k \rangle + \frac{1}{\alpha_k}D_h(x, x^k) \right\}
\]

\[
= \arg\min_{x \in \mathcal{C}} \left\{ \langle g^k, x \rangle + \frac{1}{\alpha_k}D_h(x, x^k) \right\}
\]

\[
= \arg\min_{x \in \mathcal{C}} \left\{ \langle g^k, x \rangle + \frac{1}{2\alpha_k}\|x - x^k\|^2 \right\}.
\]
**Example:** Solving problems on the probability simplex with exponentiated gradient methods.

Consider the negative entropy ($h$ is convex)

$$h(x) = \sum_{j=1}^{n} x_j \log x_j$$

restricting $x$ to the probability simplex $x \geq 0$, $\sum_{j=1}^{n} x_j = 1$. Then

$$D_h(x, y) = \sum_{j=1}^{n} [x_j (\log x_j - \log y_j) + (y_j - x_j)] = \sum_{j=1}^{n} x_j \log \frac{x_j}{y_j},$$

the entropic or Kullback-Leibler divergence.
The subproblem in mirror descent is

$$\min_{\mathbf{x}} \langle \mathbf{g}, \mathbf{x} \rangle + \sum_{j=1}^{n} x_j \log \frac{x_j}{y_j} \quad \text{s.t.} \quad \mathbf{x} \geq 0, \quad \sum_{j=1}^{n} x_j = 1$$

By introducing Lagrange multipliers $\tau \in \mathbb{R}$ for the equality constraint and $\lambda \in \mathbb{R}_+^n$ for the inequality constraint. Then we obtain Lagrangian

$$\mathcal{L}(\mathbf{x}; \tau, \lambda) = \langle \mathbf{g}, \mathbf{x} \rangle + \sum_{j=1}^{n} \left[ x_j \log \frac{x_j}{y_j} + \tau x_j - \lambda_j x_j \right] - \tau$$

Taking derivatives with respect to $\mathbf{x}$ and setting them to zero yield

$$x_j = y_j \exp(-g_j - 1 - \tau + \lambda_j).$$
Taking $\lambda_j = 0$, by $\sum_{j=1}^{n} x_j = 1$, we have (a special solution)

\[
x_i = \frac{y_i \exp(-g_i)}{\sum_{j=1}^{n} y_j \exp(-g_j)}.
\]

The update in mirror descent is

\[
x_i^{k+1} = \frac{x_i^k \exp(-\alpha_k g_i^k)}{\sum_{j=1}^{n} x_j^k \exp(-\alpha_k g_j^k)}.
\]

This is the so-called exponentiated gradient update, also known as entropic mirror descent.
Example: using $\ell_p$ norms ($p \in (1, 2]$). Consider

$$h(x) = \frac{1}{2(p-1)} \|x\|_p^2$$

We have

$$\nabla h(x) = \frac{1}{p-1} \|x\|_p^{2-p} \left[ \text{sign}(x_1)|x_1|^{p-1} \cdots \text{sign}(x_n)|x_n|^{p-1} \right]^T.$$ 

Define

$$\phi(x) = (p-1)\nabla h(x)$$

and

$$\psi(y) = \|y\|_q^{2-q} \left[ \text{sign}(y_1)|y_1|^{q-1} \cdots \text{sign}(y_n)|y_n|^{q-1} \right]^T,$$

where $q = \frac{p}{p-1}$. We have $\psi(\phi(x)) = x$, and hence, $\psi = \phi^{-1}$ and $\phi = \psi^{-1}$. 
The mirror descent for $\mathcal{C} = \mathbb{R}^n$ is

$$x^{k+1} = \arg\min_{x \in \mathbb{R}^n} \left\{ f(x^k) + \langle g^k, x - x^k \rangle + \frac{1}{\alpha_k} D_h(x, x^k) \right\}$$

$$= \arg\min_{x \in \mathbb{R}^n} \left\{ \langle g^k, x \rangle + \frac{1}{\alpha_k} D_h(x, x^k) \right\}$$

$$= (\nabla h)^{-1}(\nabla h(x^k) - \alpha_k g^k)$$

$$= \psi(\phi(x^k) - \alpha_k (p - 1) g^k).$$

The update involving the inverse of the gradient mapping $(\nabla h)^{-1}$, holds more generally, that is, for any differentiable and strictly convex $h$.

This is the original form of the mirror descent update, and it justifies the name *mirror* descent, as the gradient is “mirrored” through the distance-generating function $h$ and back again.
The case $\mathcal{C} = \{\mathbf{x} \in \mathbb{R}_+^n : \sum_{j=1}^n x_j = 1\}$

The subproblem in mirror descent is

$$\min_{\mathbf{x} \in \mathcal{C}} \langle \mathbf{v}, \mathbf{x} \rangle + \frac{1}{2} \| \mathbf{x} \|^2_p$$

where

$$\mathbf{v} = \alpha_k (p - 1) \mathbf{g}^k - \phi(x^k).$$

**Exercise:** Show, by KKT conditions, that the solution is given by

$$\mathbf{x} = \psi(\tau \mathbf{e} - \mathbf{v}^+)$$

where $\tau \in \mathbb{R}$ satisfies

$$\sum_{j=1}^n \psi_j(\tau \mathbf{e} - \mathbf{v}^+) = 1.$$

Binary search can be used to find $\tau.$
**Theorem 1**

Let $\alpha_k > 0$ be any sequence of non-increasing stepsizes. Assume that $h$ is strongly convex with respect to $\| \cdot \|$. Let $\| \cdot \|_\ast$ denote the dual norm. Let $x^k$ be generated by the mirror descent iteration. If $D_h(x, x_\ast) \leq R^2$ for all $x \in C$, then for all $K \geq 1$

$$\sum_{k=1}^{K} [f(x^k) - f(x_\ast)] \leq \frac{R^2}{\alpha K} + \sum_{k=1}^{K} \frac{\alpha_k}{2} \|g^k\|_*^2.$$ 

If $\alpha_k = \alpha$ is constant, then for all $K \geq 1$

$$\sum_{k=1}^{K} [f(x^k) - f(x_\ast)] \leq \frac{1}{\alpha} D_h(x_\ast, x^1) + \frac{\alpha}{2} \sum_{k=1}^{K} \|g^k\|_*^2.$$ 

- If $\bar{x}^K = \frac{1}{K} \sum_{k=1}^{K} x^k$ or $\bar{x}^K = \text{argmin}_{x^k} f(x^k)$, $\|g\|_* \leq M$ for all $g \in \partial f(x)$ for $x \in C$, then $f(\bar{x}^K) - f(x_\ast) \leq \frac{1}{K\alpha} D_h(x_\ast, x^1) + \frac{\alpha}{2} M^2.$
**Exercise:** Prove that the negative entropy

\[ h(x) = \sum_{j=1}^{n} x_j \log x_j, \]

is 1-strongly convex with respect to the \( \ell_1 \)-norm.

**Example:** Let \( C = \{ x \in \mathbb{R}^n_+ : \sum_{j=1}^{n} x_j = 1 \} \). Let \( x^1 = \frac{1}{n} e \). The exponentiated gradient method with fixed stepsize \( \alpha \) guarantees

\[ f(\overline{x}^K) - f(x^*) \leq \frac{\log n}{K \alpha} + \frac{\alpha}{2K} \sum_{k=1}^{K} \| g^k \|_\infty^2, \]

where \( \overline{x}^K = \frac{1}{K} \sum_{k=1}^{K} x^k. \)
Example: Let $\mathcal{C} \subset \{x \in \mathbb{R}^n : \|x\|_1 \leq R_1 \}$. Let $h(x) = \frac{1}{2(p - 1)} \|x\|^2_p$

with $p = 1 + \frac{1}{\log(2n)}$. Then

$$
\sum_{k=1}^{K} [f(x^k) - f(x_\star)] \leq \frac{2R_1^2 \log(2n)}{\alpha K} + \frac{e^2}{2} \sum_{k=1}^{K} \alpha_k \|g^k\|^2_\infty.
$$

In particular, taking $\alpha_k = \frac{R_1}{e} \sqrt{\frac{\log(2n)}{k}}$ and $\bar{x}^K = \frac{1}{K} \sum_{k=1}^{K} x^k$ gives

$$
f(\bar{x}^K) - f(x_\star) \leq \frac{3eR_1 \sqrt{\log(2n)}}{\sqrt{K}}.
$$
A simulated mirror-descent example: Let

\[ A = [a_1 \ldots a_m]^T \in \mathbb{R}^{m \times n} \]

have entries drawn i.i.d \( \mathcal{N}(0, 1) \). Let

\[ b_i = \frac{1}{2}(a_{i1} + a_{i2}) + \varepsilon_i, \]

where \( \varepsilon_i \) drawn i.i.d \( \mathcal{N}(0, 10^{-2}) \), and \( n = 3000, m = 20 \). Define

\[ f(x) := \|Ax - b\|_1 = \sum_{i=1}^m |\langle a_i, x \rangle - b_i|, \]

which has subgradients \( A^T \text{sign}(Ax - b) \). Consider

\[ \min_x f(x) \quad \text{s.t.} \quad x \in C = \left\{ x \in \mathbb{R}_+^n : \sum_{j=1}^n x_j = 1 \right\} \]
We compare the subgradient method to exponentiated gradient
descent for this problem, noting that the Euclidean projection of a
vector $\mathbf{v} \in \mathbb{R}^n$ to the set $\mathcal{C}$ has coordinates

$$x_j = [v_j - \tau]_+,$$

where $\tau \in \mathbb{R}$ is chosen so that

$$\sum_{j=1}^{n} x_j = \sum_{j=1}^{n} [v_j - \tau]_+ = 1.$$

We use stepsize $\alpha_k = \alpha_0 / \sqrt{k}$ where the initial stepsize $\alpha_0$ is chosen
to optimize the convergence guarantee for each of the methods.
\( f^k_{\text{best}} - f_* \) versus \( k \)
2.1 Stochastic version

**Theorem 2**

Let the conditions of Theorem 1 hold, except that instead of receiving a vector

\[ g^k \in \partial f(x^k) \]

at iteration \( k \), the vector \( g^k \) is a stochastic subgradient satisfying

\[ \mathbb{E}[g^k | x^k] \in \partial f(x^k) \).

Then for any non-increasing stepsize sequence \( \alpha_k \) (where \( \alpha_k \) may be chosen dependent on \( g^1, \cdots, g^k \)),

\[ \mathbb{E} \left[ \sum_{k=1}^{K} (f(x_k) - f(x_*)) \right] \leq \mathbb{E} \left[ \frac{R^2}{\alpha K} + \sum_{k=1}^{K} \frac{\alpha_k}{2} \|g^k\|_*^2 \right]. \]
2.2 Adaptive stepsizes

- Let $\bar{x}^K = \frac{1}{K} \sum_{k=1}^{K} x^k$. If $D_h(x, x_*) \leq R^2$ for all $x \in C$, then

$$
\mathbb{E}[f(\bar{x}^K) - f(x_*)] \leq \mathbb{E} \left[ \frac{R^2}{K\alpha K} + \frac{1}{K} \sum_{k=1}^{K} \frac{\alpha_k}{2} \|g^k\|_2^2 \right].
$$

If $\alpha_k = \alpha$ for all $k$, we see that the choice of stepsize minimizing

$$
\frac{R^2}{K\alpha} + \frac{\alpha}{2K} \sum_{k=1}^{K} \|g^k\|_2^2
$$

is $\alpha_* = \sqrt{2R} \left( \sum_{k=1}^{K} \|g^k\|_2^2 \right)^{-\frac{1}{2}}$, which yields

$$
\mathbb{E}[f(\bar{x}^K) - f(x_*)] \leq \frac{\sqrt{2R}}{K} \mathbb{E} \left[ \left( \sum_{k=1}^{K} \|g^k\|_2^2 \right)^{\frac{1}{2}} \right].
$$

- But this $\alpha_*$ is not known in advance because ...
Corollary 3

Let the conditions of Theorem 2 hold. Let

$$\alpha_k = \frac{R}{\sqrt{\sum_{i=1}^{k} \|g_i\|^2_*}},$$

which is the “up to now” optimal choice. Then

$$\mathbb{E}[f(\bar{x}^K) - f(x_*)] \leq \frac{3R}{K} \mathbb{E} \left[ \left( \sum_{k=1}^{K} \|g^k\|^2_* \right)^{\frac{1}{2}} \right],$$

where

$$\bar{x}^K = \frac{1}{K} \sum_{k=1}^{K} x^k.$$
3. Variable metric methods

- The idea is to adjust the metric with which one constructs updates to better reflect problem structure.

1. Choose subgradient \( \mathbf{g}^k \in \partial f(\mathbf{x}^k) \) or stochastic subgradient \( \mathbf{g}^k \) satisfying

   \[
   \mathbb{E}[\mathbf{g}^k | \mathbf{x}^k] \in \partial f(\mathbf{x}^k)
   \]

2. Update positive semidefinite matrix \( \mathbf{H}_k \in \mathbb{R}^{n \times n} \)

3. Compute update

   \[
   \mathbf{x}^{k+1} = \operatorname{argmin}_{\mathbf{x} \in C} \left\{ \langle \mathbf{g}^k, \mathbf{x} \rangle + \frac{1}{2} \langle \mathbf{x} - \mathbf{x}^k, \mathbf{H}_k (\mathbf{x} - \mathbf{x}^k) \rangle \right\}.
   \]

- Special cases: Subgradient method \( \mathbf{H}_k = \frac{1}{\alpha_k} \mathbf{I} \)

  Newton method \( \mathbf{H}_k = \nabla^2 f(\mathbf{x}^k) \)
Theorem 4

Let $H_k$ be a sequence of positive definite matrices, where $H_k$ is a function of $g^1, \ldots, g^k$ (and potentially some additional randomness). Let $g^k$ be (stochastic) subgradients with

$$\mathbb{E}[g^k|x^k] \in \partial f(x^k).$$

Then

$$\mathbb{E} \left[ \sum_{k=1}^{K} (f(x^k) - f(x_*)) \right]$$

$$\leq \frac{1}{2} \mathbb{E} \left[ \sum_{k=2}^{K} \left( \|x^k - x_*\|^2_{H_k} - \|x^k - x_*\|^2_{H_{k-1}} \right) + \|x^1 - x_*\|^2_{H_1} \right]$$

$$+ \frac{1}{2} \mathbb{E} \left[ \sum_{k=1}^{K} \|g^k\|^2_{H^{-1}_k} \right].$$
3.1 Adaptive gradient variable metric method

**Theorem 5**

Let

\[ R_\infty := \sup_{x \in \mathcal{C}} \|x - x_*\|_\infty \]

be the \( \ell_\infty \) radius of the set \( \mathcal{C} \) and the conditions of Theorem 4 hold. Let

\[ H_k = \frac{1}{\alpha} \left( \text{diag} \left( \sum_{i=1}^{k} g_i g_i^T \right) \right)^{1/2} \]

where \( \alpha > 0 \) is a pre-specified constant. Then we have

\[ \mathbb{E} \left[ \sum_{k=1}^{K} (f(x^k) - f(x_*)) \right] \leq \left( \frac{R_\infty^2}{2} + \alpha^2 \right) \mathbb{E}[\text{tr}(H_K)]. \]
Example: Support vector machine classification problem

Consider

$$
\min_{x} f(x) = \frac{1}{m} \sum_{i=1}^{m} [1 - b_i \langle a_i, x \rangle]_+ \quad \text{s.t.} \quad \|x\|_{\infty} \leq 1,
$$

where the vectors $a_i \in \{-1, 0, 1\}^n$ with $m = 5000$ and $n = 1000$. For each coordinate $j \in [n]$, we set $a_{ij} \in \{\pm 1\}$ to have a random sign with probability $1/j$, and $a_{ij} = 0$ otherwise.

Let $u \in \{-1, 1\}^n$ uniformly at random. We set $b_i = \text{sign}(\langle a_i, u \rangle)$ with probability $0.95$ and $b_i = -\text{sign}(\langle a_i, u \rangle)$ otherwise.

We choose a stochastic gradient by selecting $i \in [m]$ uniformly at random, then set

$$
g^k \in \partial[1 - b_i \langle a_i, x^k \rangle]_+,
$$

and $\alpha^k = \alpha/\sqrt{k}$. 

- \( f(x^k) - f_* \) versus \( k/100 \): various initial stepsizes
\[ f(x^k) - f_* \] versus \( k/100 \): best initial stepsize