The University of Texas at Austin
Department of Electrical and Computer Engineering

ECE 381V: Large-Scale Optimization II — Spring 2022

LECTURE 15

Caramanis & Mokhtari — Wednesday, March 9, 2022

Goal: In this lecture, we discuss mirror descent and go beyond the Euclidean norm for iterative methods.

1 Bregman Divergence/Distance

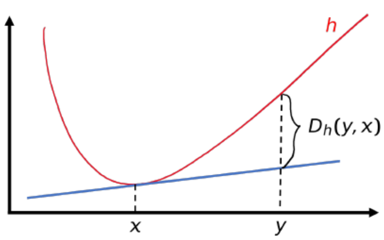
To enforce the proximity condition required in the updates of iterative methods, instead of using the Euclidean distance, one can use different distances. The non-Euclidean distances that we use are Bregman distances.

Definition 1. *Let $\omega : \mathcal{D} \to \mathbb{R}$ be a proper closed convex function that is differentiable over its domain. The Bregman distance associated with ω is the function B_ω that is formally defied as*

$$B_\omega(\mathbf{x}, \mathbf{y}) = \omega(\mathbf{x}) - \omega(\mathbf{y}) - \nabla\omega(\mathbf{y})^\top(\mathbf{x} - \mathbf{y})$$

We further assume that ω satisfies the followings assumptions:

1. ω is a proper closed convex function
2. ω is differentiable
3. ω is strongly convex with constant μ



Remark 1. *A Bregman distance is not a norm! In fact it is not (even) symmetric and it does not satisfy the triangle inequality.*

Remark 2. *A Bregman distance is not a norm, but it has the following important properties that are similar to a norm*

- $B_\omega(\mathbf{x}, \mathbf{y}) \geq 0$ *for any* $\mathbf{x}, \mathbf{y}$
- $B_\omega(\mathbf{x}, \mathbf{y}) = 0$ *if and only if* $\mathbf{x} = \mathbf{y}$

1.1 Examples

The most common choice for the function ω is $\omega(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$ which leads to the following Bregman distance

$$B_\omega(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$$

The other common choice for ω is the negative entropy function defined as $\omega : \mathbb{R}^n_{++} \to \mathbb{R}$ defined as $\omega(\mathbf{x}) = \sum_{i=1}^n x_i \log(x_i)$ which leads to the Bregman distance

$$B_\omega(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^n x_i \log \frac{x_i}{y_i} - \sum_{i=1}^n x_i + \sum_{i=1}^n y_i$$

which is known as the generalized Kullback–Leibler divergence.

1.2 Important Properties

The three point lemma is one of the most important properties of Bregman divergence.

Lemma 1 (Three Points Lemma)**.** *Suppose ω is proper convex function that is differentiable. Then, for any $\mathbf{a}, \mathbf{b}, \mathbf{c}$ in the domain of ω we have*

$$(\nabla\omega(\mathbf{b}) - \nabla\omega(\mathbf{a}))^\top(\mathbf{c} - \mathbf{a}) = B_\omega(\mathbf{c}, \mathbf{a}) + B_\omega(\mathbf{a}, \mathbf{b}) - B_\omega(\mathbf{c}, \mathbf{b})$$

Proof.

$$\begin{aligned} B_\omega(\mathbf{c}, \mathbf{a}) &= \omega(\mathbf{c}) - \omega(\mathbf{a}) - \nabla\omega(\mathbf{a})^\top(\mathbf{c} - \mathbf{a}) \\ B_\omega(\mathbf{a}, \mathbf{b}) &= \omega(\mathbf{a}) - \omega(\mathbf{b}) - \nabla\omega(\mathbf{b})^\top(\mathbf{a} - \mathbf{b}) \\ B_\omega(\mathbf{c}, \mathbf{b}) &= \omega(\mathbf{c}) - \omega(\mathbf{b}) - \nabla\omega(\mathbf{b})^\top(\mathbf{c} - \mathbf{b}) \end{aligned}$$

By computing $B_\omega(\mathbf{c}, \mathbf{a}) + B_\omega(\mathbf{a}, \mathbf{b}) - B_\omega(\mathbf{c}, \mathbf{b})$ the claim follows. □

2 Solving convex constrained problems using Mirror Descent Method

Recall the following general convex (nonsmooth) constrained problem:

$$\begin{aligned} &\min f(\mathbf{x}) \\ &\text{s.t } \mathbf{x} \in \mathcal{Q}, \end{aligned} \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex on $\mathbb{R}^n$ and the set $\mathcal{Q}$ is convex. We showed that for the projected subgradient method

$$\mathbf{x}_{k+1} = \pi_{\mathcal{Q}}(\mathbf{x}_k - \eta_k \mathbf{g}_k) = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{Q}} \left\{ \mathbf{g}_k^\top \mathbf{x} + \frac{1}{2\eta_k}\|\mathbf{x} - \mathbf{x}_k\|_2^2 \right\} = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{Q}} \left\{ \eta_k(\mathbf{g}_k - \mathbf{x}_k)^\top \mathbf{x} + \frac{1}{2}\|\mathbf{x}\|_2^2 \right\}$$

we have the following result:

$$\min_{i=0,\ldots,k} f(\mathbf{x}_i) - f(\mathbf{x}^*) \le \frac{R^2 + \sum_{i=0}^k \eta_i^2 \|\mathbf{g}_i\|_2^2}{2\sum_{i=0}^k \eta_i}$$

Next, we show that how this bound can be improved (in some settings) by moving to non-Euclidean norms.

2.1 Mirror Descent

Consider the following update which we call the mirror descent method:

$$\mathbf{x}_{k+1} = \underset{\mathbf{x}\in\mathcal{Q}}{\operatorname{argmin}} \left\{ \mathbf{g}_k^\top \mathbf{x} + \frac{1}{\eta_k} B_\omega(\mathbf{x}, \mathbf{x}_k) \right\}$$

As we observe, in this case, we use Bregman divergence instead of ℓ_2 distance for our proximity condition. Note that this update can be further simplified as

$$\mathbf{x}_{k+1} = \underset{\mathbf{x}\in\mathcal{Q}}{\operatorname{argmin}} \left\{ (\eta_k \mathbf{g}_k - \nabla\omega(\mathbf{x}_k))^\top \mathbf{x} + \omega(\mathbf{x}) \right\}$$

2.2 Interesting Interpretation

Let us define projection using the Bregman divergence onto the set $\mathcal{Q}$ as the following

$$\pi_{\mathcal{Q}}^{\omega}(\mathbf{y}) = \underset{\mathbf{x}\in\mathcal{Q}\cap\mathcal{D}}{\operatorname{argmin}} \ B_\omega(\mathbf{x}, \mathbf{y})$$

which is equivalent to

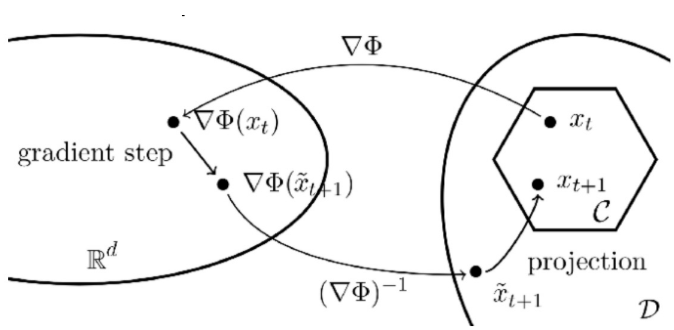
$$\pi_{\mathcal{Q}}^{\omega}(\mathbf{y}) = \underset{\mathbf{x}\in\mathcal{Q}\cap\mathcal{D}}{\operatorname{argmin}} \ \{\omega(\mathbf{x}) - \nabla\omega(\mathbf{y})^\top \mathbf{x}\}$$

Using the above definition we can write the update of mirror descent as

$$\nabla\omega(\mathbf{y}_{k+1}) = \nabla\omega(\mathbf{x}_k) - \eta_k \mathbf{g}_k$$

and

$$\mathbf{x}_{k+1} = \pi_{\mathcal{Q}}^{\omega}(\mathbf{y}_{k+1})$$



If we assume that the problem is differentiable and unconstrained then we have

$$\nabla\omega(\mathbf{x}_{k+1}) = \nabla\omega(\mathbf{x}_k) - \eta_k \nabla f(\mathbf{x}_k)$$

which implies that

$$\mathbf{x}_{k+1} = (\nabla\omega)^{-1}(\nabla\omega(\mathbf{x}_k) - \eta_k \nabla f(\mathbf{x}_k))$$

Note that $(\nabla\omega)^{-1} = \nabla(\omega^*)$, where ω^* is the conjugate of ω

2.3 Convergence Analysis

We know that

$$f(\mathbf{x}_k) - f(\mathbf{u}) \leq \mathbf{g}_k^\top(\mathbf{x}_k - \mathbf{u}) = \frac{1}{\eta_k}(\nabla\omega(\mathbf{x}_k) - \nabla\omega(\mathbf{y}_{k+1}))^\top(\mathbf{x}_k - \mathbf{u})$$

Now using three point lemma we have

$$f(\mathbf{x}_k) - f(\mathbf{u}) \leq \frac{1}{\eta_k}(B_\omega(\mathbf{u}, \mathbf{x}_k) + B_\omega(\mathbf{x}_k, \mathbf{y}_{k+1}) - B_\omega(\mathbf{u}, \mathbf{y}_{k+1}))$$

Now since

$$\mathbf{x}_{k+1} = \pi_{\mathcal{Q}}^\omega(\mathbf{y}_{k+1}) = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{Q}\cap\mathcal{D}} \{\omega(\mathbf{x}) - \nabla\omega(\mathbf{y}_{k+1})^\top\mathbf{x}\}$$

we have

$$(\nabla\omega(\mathbf{x}_{k+1}) - \nabla\omega(\mathbf{y}_{k+1}))^\top(\mathbf{u} - \mathbf{x}_{k+1}) \geq 0 \quad \text{for all } \mathbf{u} \in \mathcal{Q}\cap\mathcal{D}$$

Now again using the three point lemma we have

$$B_\omega(\mathbf{u}, \mathbf{x}_{k+1}) + B_\omega(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}) - B_\omega(\mathbf{u}, \mathbf{y}_{k+1}) = (\nabla\omega(\mathbf{x}_{k+1}) - \nabla\omega(\mathbf{y}_{k+1}))^\top(\mathbf{x}_{k+1} - \mathbf{u}) \leq 0$$

Hence, we have

$$f(\mathbf{x}_k) - f(\mathbf{u}) \leq \frac{1}{\eta_k}(B_\omega(\mathbf{u}, \mathbf{x}_k) + B_\omega(\mathbf{x}_k, \mathbf{y}_{k+1}) - B_\omega(\mathbf{u}, \mathbf{x}_{k+1}) - B_\omega(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}))$$

Now using the definition of Bregman divergence we have

$$B_\omega(\mathbf{x}_k, \mathbf{y}_{k+1}) - B_\omega(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}) = \omega(\mathbf{x}_k) - \omega(\mathbf{x}_{k+1}) - \nabla\omega(\mathbf{y}_{k+1})^\top(\mathbf{x}_k - \mathbf{x}_{k+1})$$

now using the fact that ω is strongly convex with respect to norm $\|.\|$ we have

$$\begin{aligned}
B_\omega(\mathbf{x}_k, \mathbf{y}_{k+1}) - B_\omega(\mathbf{x}_{k+1}, \mathbf{y}_{k+1}) &\leq (\nabla\omega(\mathbf{x}_k) - \nabla\omega(\mathbf{y}_{k+1}))\top(\mathbf{x}_k - \mathbf{x}_{k+1}) - \frac{\sigma}{2}\|\mathbf{x}_k - \mathbf{x}_{k+1}\|^2 \\
&= \eta_k\mathbf{g}_k^\top(\mathbf{x}_k - \mathbf{x}_{k+1}) - \frac{\sigma}{2}\|\mathbf{x}_k - \mathbf{x}_{k+1}\|^2 \\
&\leq \eta_k\|\mathbf{g}_k\|_*\|\mathbf{x}_k - \mathbf{x}_{k+1}\| - \frac{\sigma}{2}\|\mathbf{x}_k - \mathbf{x}_{k+1}\|^2 \\
&\leq \frac{\eta_k^2}{2\sigma}\|\mathbf{g}_k\|_*^2 + \frac{\sigma}{2}\|\mathbf{x}_k - \mathbf{x}_{k+1}\|^2 - \frac{\sigma}{2}\|\mathbf{x}_k - \mathbf{x}_{k+1}\|^2 \\
&\leq \frac{\eta_k^2}{2\sigma}\|\mathbf{g}_k\|_*^2
\end{aligned}$$

By combining these results, setting $\mathbf{u} = \mathbf{x}^*$ and computing the sum of both sides we obtain

$$\sum_{i=0}^{k}\eta_i(f(\mathbf{x}_i) - f(\mathbf{x}^*)) \leq B_\omega(\mathbf{x}^*, \mathbf{x}_0) + \sum_{i=0}^{k}\frac{\eta_i^2}{2\sigma}\|\mathbf{g}_i\|_*^2$$

which implies that

$$\min_{i=0,\ldots,k} f(\mathbf{x}_i) - f(\mathbf{x}^*) \leq \frac{B_\omega(\mathbf{x}^*, \mathbf{x}_0) + \frac{1}{2\sigma}\sum_{i=0}^{k}\eta_i^2\|\mathbf{g}_i\|_*^2}{\sum_{i=0}^{k}\eta_i}$$

Now if L_* is a uniform bound on $\|\mathbf{g}_i\|_*$ and we select the stepsize as $\eta_k = \frac{\sqrt{2R\sigma}}{L_*\sqrt{k+1}}$, then we have

$$\min_{i=0,\ldots,k} f(\mathbf{x}_i) - f(\mathbf{x}^*) \leq \frac{\sqrt{2R(\mathbf{x}_0)}L_*}{\sqrt{\sigma}\sqrt{k+1}}$$

3 Convergence rate comparison

Consider the following problem

$$\min f(\mathbf{x}) \qquad \text{s.t } \mathbf{x} \in \Delta_n,$$

where Δ_n is the unit simplex.
In this case, by following the projected subgradient method which is equivalent to the case that $\omega(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$ we have $\sigma = 1$ for l_2 norm and

$$R(\mathbf{x}_0) = \max_{\mathbf{x}\in\Delta_n} \frac{1}{2}\|\mathbf{x} - \mathbf{x}_0\|_2^2 = \frac{1}{2}\left(1 - \frac{1}{n}\right), \qquad L_* = L_2$$

assuming that $\mathbf{x}_0 = (1/n, \dots, 1/n)$. Hence,

$$\min_{i=0,\dots,k} f(\mathbf{x}_i) - f(\mathbf{x}^*) \leq \frac{L_2}{\sqrt{k+1}}$$

On the other hand, if we use the negative entropy function

$$\omega(\mathbf{x}) = \sum_{i=1}^{n} x_i \log(x_i)$$

we have that the function is 1-strongly convex with respect to ℓ_1 norm. Hence, and further have

$$R(\mathbf{x}_0) = \max_{\mathbf{x}\in\Delta_n} B_\omega(\mathbf{x}^*, \mathbf{x}_0) = \max_{\mathbf{x}\in\Delta_n} \sum_{i=1}^{n} x_i \log(n x_i) = \log(n) + \sum_{i=1}^{n} x_i \log(x_i) = \log(n)$$

Hence, we have

$$\min_{i=0,\dots,k} f(\mathbf{x}_i) - f(\mathbf{x}^*) \leq \frac{\sqrt{2\log(n)}L_\infty}{\sqrt{k+1}}$$

The important observation here is that

$$\frac{1}{\sqrt{n}} \leq \frac{L_\infty}{L_2} \leq 1$$

Hence, in the best case scenario we could obtain a gain of $\mathcal{O}\left(\frac{\sqrt{n}}{\log(n)}\right)$!

4 Extension to Proximal Gradient Method for Composite Optimization

Composite Optimization. In this section, we formally study the composite optimization problem

$$\min f(\mathbf{x}) := \phi(\mathbf{x}) + h(\mathbf{x}) \tag{2}$$

- ϕ is convex and and L-smooth.
- h is convex (possibly nondifferentiable) and its prox operator is easy to compute.
- The optimal solution set is nonempty.

Perhaps the most special case of the above problem and algorithm is the ℓ_1 regularization problem:

$$\min f(\mathbf{x}) := \phi(\mathbf{x}) + \lambda \|\mathbf{x}\|_1$$

Accelerated Proximal Gradient Method (APGM). To solve the above problem, we follow the update of APGM which extends the idea of acceleration to proximal gradient methods.

Initialize: $\mathbf{x}_0$
Main Loop:

$$\mathbf{x}_{k+1} = \mathrm{prox}_{\eta_k h}(\mathbf{x}_k - \eta_k \nabla\phi(\mathbf{x}_k)) = \operatorname*{argmin}_{\mathbf{x}} \left\{ \nabla\phi(\mathbf{x}_k)^\top (\mathbf{x} - \mathbf{x}_k) + h(\mathbf{x}) + \frac{1}{2\eta_k} \|\mathbf{x} - \mathbf{x}_k\|_2^2 \right\}$$

Accelerated Proximal Gradient Method (APGM) beyond Euclidean norm.

Initialize: $\mathbf{x}_0$
Main Loop:

$$\mathbf{x}_{k+1} = \operatorname*{argmin}_{\mathbf{x}} \left\{ \nabla\phi(\mathbf{x}_k)^\top (\mathbf{x} - \mathbf{x}_k) + h(\mathbf{x}) + \frac{1}{\eta_k} B_\omega(\mathbf{x}, \mathbf{x}_k) \right\}$$