18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet
Scribe: Mina Karzand

Lecture 13
Oct. 21, 2015

Previously, we analyzed the convergence of the projected gradient descent algorithm. We proved that optimizing the convex L-Lipschitz function f on a closed, convex set $\mathcal{C}$ with $\text{diam}(\mathcal{C}) \leq R$ with step sizes $\eta_s = \frac{R}{L\sqrt{k}}$ would give us accuracy of $f(\overline{x}) \leq f(x^*) + \frac{LR}{\sqrt{k}}$ after k iterations.

Although it might seem that projected gradient descent algorithm provides dimension-free convergence rate, it is not always true. Reviewing the proof of convergence rate, we realize that dimension-free convergence is possible when the objective function f and the constraint set $\mathcal{C}$ are well-behaved in Euclidean norm (i.e., for all $x \in \mathcal{C}$ and $g \in \partial f(x)$, we have that $|x|_2$ and $|g|_2$ are independent of the ambient dimension). We provide an examples of the cases that these assumptions are not satisfied.

- Consider the differentiable, convex function f on the Euclidean ball $B_{2,n}$ such that $\|\nabla f(x)\|_\infty \leq 1$, $\forall x \in B_{2,n}$. This implies that $|\nabla f(x)|_2 \leq \sqrt{n}$ and the projected gradient descent converges to the minimum of f in $B_{2,n}$ at rate $\sqrt{\frac{n}{k}}$. Using the method of mirror descent we can get convergence rate of $\sqrt{\frac{\log(n)}{k}}$

To get better rates of convergence in the optimization problem, we can use the Mirror Descent algorithm. The idea is to change the Euclidean geometry to a more pertinent geometry to a problem at hand. We will define a new geometry by using a function which is sometimes called potential function $\Phi(x)$. We will use Bregman projection based on Bregman divergence to define this geometry.

The geometric intuition behind the mirror Descent algorithm is the following: The projected gradient described in previous lecture works in any arbitrary Hilbert space $\mathcal{H}$ so that the norm of vectors is associated with an inner product. Now, suppose we are interested in optimization in a Banach space $\mathcal{D}$. In other words, the norm (or the measure of distance) that we use does not derive from an inner product. In this case, the gradient descent does not even make sense since the gradient $\nabla f(x)$ are elements of dual space. Thus, the term $x - \eta \nabla f(x)$ cannot be performed. (Note that in Hilbert space used in projected gradient descent, the dual space of $\mathcal{H}$ is isometric to $\mathcal{H}$. Thus, we didn't have any such problems.)

The geometric insight of the Mirror Descent algorithm is that to perform the optimization in the primal space $\mathcal{D}$, one can first map the point $x \in \mathcal{D}$ in primal space to the dual space $\mathcal{D}^*$, then perform the gradient update in the dual space and finally map the optimal point back to the primal space. Note that at each update step, the new point in the primal space $\mathcal{D}$ might be outside of the constraint set $\mathcal{C} \subset \mathcal{D}$, in which case it should be projected into the constraint set $\mathcal{C}$. The projection associate with the Mirror Descent algorithm is Bergman Projection defined based on the notion of Bergman divergence.

Definition (Bregman Divergence): For given differentiable, α-strongly convex function $\Phi(x) : \mathcal{D} \to \mathbb{R}$, we define the Bregman divergence associated with Φ to be:

$$D_\Phi(y, x) = \Phi(y) - \Phi(x) - \nabla\Phi(x)^T(y - x)$$

We will use the convex open set $\mathcal{D} \subset \mathbb{R}^n$ whose closure contains the constraint set $\mathcal{C} \subset \overline{\mathcal{D}}$. Bregman divergence is the error term of the first order Taylor expansion of the function Φ in $\mathcal{D}$.

Also, note that the function $\Phi(x)$ is said to be α-strongly convex w.r.t. a norm $\|.\|$ if

$$\Phi(y) - \Phi(x) - \nabla\Phi(x)^T(y-x) \geq \frac{\alpha}{2}\|y-x\|^2 .$$

We used the following property of the Euclidean norm:

$$2a^\top b = \|a\|^2 + \|b\|^2 - \|a-b\|^2$$

in the proof of convergence of projected gradient descent, where we chose $a = x_s - y_{s+1}$ and $b = x_s - x^*$.

To prove the convergence of the Mirror descent algorithm, we use the following property of the Bregman divergence in a similar fashion. This proposition shows that the Bregman divergence essentially behaves as the Euclidean norm squared in terms of projections:

Proposition: Given α-strongly differentiable convex function $\Phi : \mathcal{D} \to \mathbb{R}$, for all $x, y, z \in \mathcal{D}$,

$$\left[\nabla\Phi(x) - \nabla\Phi(y)\right]^\top (x-z) = D_\Phi(x,y) + D_\Phi(z,x) - D_\Phi(z,y) .$$

As described previously, the Bregman divergence is used in each step of the Mirror descent algorithm to project the updated value into the constraint set.

Definition (Bregman Projection): Given α-strongly differentiable convex function $\Phi : \mathcal{D} \to \mathbb{R}$ and for all $x \in \mathcal{D}$ and closed convex set $\mathcal{C} \subset \overline{\mathcal{D}}$

$$\Pi_{\mathcal{C}}^{\Phi}(x) = \operatorname*{argmin}_{z \in \mathcal{C} \cap \mathcal{D}} D_\Phi(z,x)$$

2.4.2 Mirror Descent Algorithm

Algorithm 1 Mirror Descent algorithm

Input: $x_1 \in \operatorname{argmin}_{\mathcal{C}\cap\mathcal{D}} \Phi(x)$, $\zeta : \mathbb{R}^d \to \mathbb{R}^d$ such that $\zeta(x) = \nabla\Phi(x)$
for $s = 1, \cdots, k$ **do**
  $\zeta(y_{s+1}) = \zeta(x_s) - \eta g_s$ for $g_s \in \partial f(x_s)$
  $x_{s+1} = \Pi_{\mathcal{C}}^{\Phi}(y_{s+1})$
end for
return Either $\overline{x} = \frac{1}{k}\sum_{s=1}^{k} x_s$ or $x^\circ \in \operatorname{argmin}_{x \in \{x_1, \cdots, x_k\}} f(x)$

Proposition: Let $z \in \mathcal{C} \cap \mathcal{D}$, then $\forall y \in \mathcal{D}$,

$$\left(\nabla\Phi(\pi(y)) - \nabla\Phi(y)\right)^\top (\pi(y) - z) \leq 0$$

Moreover, $D_\Phi(z, \pi(y)) \leq D_\Phi(z, y)$.

Proof. Define $\pi = \Pi_{\mathcal{C}}^{\Phi}(y)$ and $h(t) = D_\Phi(\pi + t(z - \pi), y)$. Since $h(t)$ is minimized at $t = 0$ (due to the definition of projection), we have

$$h'(0) = \nabla_x D_\Phi(x, y)|_{x=\pi}(z - \pi) \geq 0$$

where suing the definition of Bregman divergence,

$$\nabla_x D_\Phi(x, y) = \nabla\Phi(x) - \nabla\Phi(y)$$

Thus,

$$(\nabla\Phi(\pi) - \nabla\Phi(y))^\top (\pi - z) \leq 0\,.$$

Using proposition 1, we know that

$$(\nabla\Phi(\pi) - \nabla\Phi(y))^\top (\pi - z) = D_\Phi(\pi, y) + D_\Phi(z, \pi) - D_\Phi(z, y) \leq 0\,,$$

and since $D_\Phi(\pi, y) \geq 0$, we would have $D_\Phi(z, \pi) \leq D_\Phi(z, y)$. □

Theorem: Assume that f is convex and L-Lipschitz w.r.t. $\|.\|$. Assume that Φ is α-strongly convex on $\mathcal{C} \cap \mathcal{D}$ w.r.t. $\|.\|$ and

$$R^2 = \sup_{x \in \mathcal{C} \cap \mathcal{D}} \Phi(x) - \min_{x \in \mathcal{C} \cap \mathcal{D}} \Phi(x)$$

take $x_1 = \text{argmin}_{x \in \mathcal{C} \cap \mathcal{D}} \Phi(x)$ (assume that it exists). Then, Mirror Descent with $\eta = \frac{R}{L}\sqrt{\frac{2\alpha}{R}}$ gives,

$$f(\overline{x}) - f(x^*) \leq RL\sqrt{\frac{2}{\alpha k}} \quad \text{and} \quad f(\overline{x}^\circ) - f(x^*) \leq RL\sqrt{\frac{2}{\alpha k}},$$

Proof. Take $x^\sharp \in \mathcal{C} \cap \mathcal{D}$. Similar to the proof of the projected gradient descent, we have:

$$\begin{aligned}
f(x_s) - f(x^\sharp) &\overset{(i)}{\leq} g_s^\top (x_s - x^\sharp) \\
&\overset{(ii)}{=} \frac{1}{\eta} (\zeta(x_s) - \zeta(y_{s+1}))^\top (x_s - x^\sharp) \\
&\overset{(iii)}{=} \frac{1}{\eta} (\nabla\Phi(x_s) - \nabla\Phi(y_{s+1}))^\top (x_s - x^\sharp) \\
&\overset{(iv)}{=} \frac{1}{\eta} \Big[D_\Phi(x_s, y_{s+1}) + D_\Phi(x^\sharp, x_s) - D_\Phi(x^\sharp, y_{s+1}) \Big] \\
&\overset{(v)}{\leq} \frac{1}{\eta} \Big[D_\Phi(x_s, y_{s+1}) + D_\Phi(x^\sharp, x_s) - D_\Phi(x^\sharp, x_{s+1}) \Big] \\
&\overset{(vi)}{\leq} \frac{\eta L^2}{2\alpha^2} + \frac{1}{\eta} \Big[D_\Phi(x^\sharp, x_s) - D_\Phi(x^\sharp, x_{s+1}) \Big]
\end{aligned}$$

Where (i) is due to convexity of the function f.

Equations (ii) and (iii) are direct results of Mirror descent algorithm.
Equation (iv) is the result of applying proposition 1.
Inequality (v) is a result of the fact that $x_{s+1} = \Pi_{\mathcal{C}}^{\Phi}(y_{s+1})$, thus for $x^\sharp \in \mathcal{C} \cap \mathcal{D}$, we have $D_\Phi(x^\sharp, y_{s+1}) \geq D_\Phi(x^\sharp, x_{s+1})$.
We will justify the following derivations to prove inequality (vi):

$$\begin{aligned} D_\Phi(x_s, y_{s+1}) &\overset{(a)}{=} \Phi(x_s) - \Phi(y_{s+1}) - \nabla\Phi(y_{s+1})^\top (x_s - y_{s+1}) \\ &\overset{(b)}{\leq} \left[\nabla\Phi(x_s) - \nabla\Phi(y_{s+1})\right]^\top (x_s - y_{s+1}) - \frac{\alpha}{2}\|y_{s+1} - x_s\|^2 \\ &\overset{(c)}{\leq} \eta\|g_s\|_*\|x_s - y_{s+1}\| - \frac{\alpha}{2}\|y_{s+1} - x_s\|^2 \\ &\overset{(d)}{\leq} \frac{\eta^2 L^2}{2\alpha}\,. \end{aligned}$$

Equation (a) is the definition of Bregman divergence.
To show inequality (b), we used the fact that Φ is α-strongly convex which implies that $\Phi(y_{s+1}) - \Phi(x_s) \geq \nabla\Phi(x_s)^T(y_{s+1} - x_s)\frac{\alpha}{2}\|y_{s+1} - x_s\|^2$.
According to the Mirror descent algorithm, $\nabla\Phi(x_s) - \nabla\Phi(y_{s+1}) = \eta g_s$. We use Hölder's inequality to show that $g_s^\top(x_s - y_{s+1}) \leq \|g_s\|_*\|x_s - y_{s+1}\|$ and derive inequality (c).
Looking at the quadratic term $ax - bx^2$ for $a, b > 0$, it is not hard to show that $\max ax - bx^2 = \frac{a^2}{4b}$. We use this statement with $x = \|y_{s+1} - x_s\|$, $a = \eta\|g_s\|_* \leq L$ and $b = \frac{\alpha}{2}$ to derive inequality (d).

Again, we use telescopic sum to get

$$\frac{1}{k}\sum_{s=1}^{k}[f(x_s) - f(x^\sharp)] \leq \frac{\eta L^2}{2\alpha} + \frac{D_\Phi(x^\sharp, x_1)}{k\eta}\,. \tag{2.1}$$

We use the definition of Bregman divergence to get

$$\begin{aligned} D_\Phi(x^\sharp, x_1) &= \Phi(x^\sharp) - \Phi(x_1) - \nabla\Phi(x_1)(x^\sharp - x_1) \\ &\leq \Phi(x^\sharp) - \Phi(x_1) \\ &\leq \sup_{x\in\mathcal{C}\cap\mathcal{D}} \Phi(x) - \min_{x\in\mathcal{C}\cap\mathcal{D}} \Phi(x) \\ &\leq R^2\,. \end{aligned}$$

Where we used the fact $x_1 \in \operatorname{argmin}_{\mathcal{C}\cap\mathcal{D}} \Phi(x)$ in the description of the Mirror Descent algorithm to prove $\nabla\Phi(x_1)(x^\sharp - x_1) \geq 0$. We optimize the right hand side of equation (2.1) for η to get

$$\frac{1}{k}\sum_{s=1}^{k}[f(x_s) - f(x^\sharp)] \leq RL\sqrt{\frac{2}{\alpha k}}\,.$$

To conclude the proof, let $x^\sharp \to x^* \in \mathcal{C}$. □

Note that with the right geometry, we can get projected gradient descent as an instance the Mirror descent algorithm.

2.4.3 Remarks

The Mirror Descent is sometimes called Mirror Prox. We can write x_{s+1} as

$$\begin{aligned} x_{s+1} &= \operatorname*{argmin}_{x \in \mathcal{C} \cap \mathcal{D}} D_\Phi(x, y_{s+1}) \\ &= \operatorname*{argmin}_{x \in \mathcal{C} \cap \mathcal{D}} \Phi(x) - \nabla \Phi^\top(y_{s+1}) x \\ &= \operatorname*{argmin}_{x \in \mathcal{C} \cap \mathcal{D}} \Phi(x) - [\nabla \Phi(x_s) - \eta g_s]^\top x \\ &= \operatorname*{argmin}_{x \in \mathcal{C} \cap \mathcal{D}} \eta(g_s^\top x) + \Phi(x) - \nabla \Phi^\top(x_s) x \\ &= \operatorname*{argmin}_{x \in \mathcal{C} \cap \mathcal{D}} \eta(g_s^\top x) + D_\Phi(x, x_s) \end{aligned}$$

Thus, we have

$$x_{s+1} = \operatorname*{argmin}_{x \in \mathcal{C} \cap \mathcal{D}} \eta(g_s^\top x) + D_\Phi(x, x_s)\,.$$

To get x_{s+1}, in the first term on the right hand side we look at linear approximations close to x_s in the direction determined by the subgradient g_s. If the function is linear, we would just look at the linear approximation term. But if the function is not linear, the linear approximation is only valid in a small neighborhood around x_s. Thus, we penalized by adding the term $D_\Phi(x, x_s)$. We can penalized by the square norm when we choose $D_\Phi(x, x_s) = \|x - x_s\|^2$. In this case we get back the projected gradient descent algorithm as an instance of Mirror descent algorithm.

But if we choose a different divergence $D_\Phi(x, x_s)$, we are changing the geometry and we can penalize differently in different directions depending on the geometry.

Thus, using the Mirror descent algorithm, we could replace the 2-norm in projected gradient descent algorithm by another norm, hoping to get less constraining Lipschitz constant. On the other hand, the norm is a lower bound on the strong convexity parameter. Thus, there is trade off in improvement of rate of convergence.

2.4.4 Examples

Euclidean Setup:

$\Phi(x) = \frac{1}{2}\|x\|^2, \quad \mathcal{D} = \mathbb{R}^d, \quad \nabla\Phi(x) = \zeta(x) = x$. Thus, the updates will be similar to the gradient descent.

$$\begin{aligned} D_\Phi(y, x) &= \frac{1}{2}\|y\|^2 - \frac{1}{2}\|x\|^2 - x^\top y + \|x\|^2 \\ &= \frac{1}{2}\|x - y\|^2\,. \end{aligned}$$

Thus, Bregman projection with this potential function $\Phi(x)$ is the same as the usual Euclidean projection and the Mirror descent algorithm is exactly the same as the projected descent algorithm since it has the same update and same projection operator.

Note that $\alpha = 1$ since $D_\Phi(y, x) \geq \frac{1}{2}\|x - y\|^2$.

ℓ_1 Setup:

We look at $\mathcal{D} = \mathbb{R}^d_+ \setminus \{0\}$.

Define $\Phi(x)$ to be the negative entropy so that:

$$\Phi(x) = \sum_{i=1}^{d} x_i \log(x_i), \qquad \zeta(x) = \nabla\Phi(x) = \{1 + \log(x_i)\}_{i=1}^{d}$$

Thus, looking at the update function $y^{(s+1)} = \nabla\Phi(x^{(s)}) - \eta g_s$, we get $\log(y_i^{(s+1)}) = \log(x_i^{(s)}) - \eta g_i^{(s)}$ and for all $i = 1, \cdots, d$, we have $y_i^{(s+1)} = x_i^{(s)} \exp(-\eta g_i^{(s)})$. Thus,

$$y^{(s)} = x^{(s)} \exp(-\eta g^{(s)}) \,.$$

We call this setup exponential Gradient Descent or Mirror Descent with multiplicative weights.

The Bregman divergence of this mirror map is given by

$$\begin{aligned} D_\Phi(y, x) &= \Phi(y) - \Phi(x) - \nabla\Phi^\top(x)(y - x) \\ &= \sum_{i=1}^{d} y_i \log(y_i) - \sum_{i=1}^{d} x_i \log(x_i) - \sum_{i=1}^{d} (1 + \log(x_i))(y_i - x_i) \\ &= \sum_{i=1}^{d} y_i \log(\frac{y_i}{x_i}) + \sum_{i=1}^{d} (y_i - x_i) \end{aligned}$$

Note that $\sum_{i=1}^{d} y_i \log(\frac{y_i}{x_i})$ is call the Kullback-Leibler divergence (KL-div) between y and x.

We show that the projection with respect to this Bregman divergence on the simplex $\Delta_d = \{x \in \mathbb{R}^d : \sum_{i=1}^{d} x_i = 1, x_i \geq 0\}$ amounts to a simple renormalization $y \mapsto y/|y|_1$. To prove so, we provide the Lagrangian:

$$\mathcal{L} = \sum_{i=1}^{d} y_i \log(\frac{y_i}{x_i}) + \sum_{i=1}^{d} (x_i - y_i) + \lambda(\sum_{i=1}^{d} x_i - 1) \,.$$

To find the Bregman projection, for all $i = 1, \cdots, d$ we write

$$\frac{\partial}{\partial x_i}\mathcal{L} = -\frac{y_i}{x_i} + 1 + \lambda = 0$$

Thus, for all i, we have $x_i = \gamma y_i$. We know that $\sum_{i=1}^{d} x_i = 1$. Thus, $\gamma = \frac{1}{\sum y_i}$.

Thus, we have $\Pi_{\Delta_d}^{\Phi}(y) = \frac{y}{|y|_1}$. The Mirror Descent algorithm with this update and projection would be:

$$\begin{aligned} y_{s+1} &= x_s \exp(-\eta g_s) \\ x_{s+1} &= \frac{y}{|y|_1} \,. \end{aligned}$$

To analyze the rate of convergence, we want to study the ℓ_1 norm on Δ_d. Thus, we have to show that for some α, Φ is α-strongly convex w.r.t $|\cdot|_1$ on Δ_d.

$$\begin{aligned} D_\Phi(y,x) &= KL(y,x) + \sum_i (x_i - y_i) \\ &= KL(y,x) \\ &\geq \frac{1}{2}|x-y|_1^2 \end{aligned}$$

Where we used the fact that $x, y \in \Delta_d$ to show $\sum_i (x_i - y_i) = 0$ and used Pinsker inequality show the result. Thus, Φ is 1-strongly convex w.r.t. $|\cdot|_1$ on Δ_d.

Remembering that $\Phi(x) = \sum_{i=1}^d x_i \log(x_i)$ was defined to be negative entropy, we know that $-\log(d) \leq \Phi(x) \leq 0$ for $x \in \Delta_d$. Thus,

$$R^2 = \max_{x \in \Delta_d} \Phi(x) - \min_{x \in \Delta_d} \Phi(x) = \log(d)\,.$$

Corollary: Let f be a convex function on Δ_d such that

$$\|g\|_\infty \leq L, \qquad \forall g \in \partial f(x), \quad \forall x \in \Delta_d\,.$$

Then, Mirror descent with $\eta = \frac{1}{L}\sqrt{\frac{2\log(d)}{k}}$ gives

$$f(\overline{x}_k) - f(x^*) \leq L\sqrt{\frac{2\log(d)}{k}}, \qquad f(x_k^\circ) - f(x^*) \leq L\sqrt{\frac{2\log(d)}{k}}$$

Boosting: For weak classifiers $f_1(x), \cdots, f_N(x)$ and $\alpha \in \Delta_n$, we define

$$f_\alpha = \sum_{j=1}^N \alpha_j f_j \qquad \text{and} \qquad F(x) = \begin{pmatrix} f_1(x) \\ \vdots \\ f_N(x) \end{pmatrix}$$

so that $f_\alpha(x)$ is the weighted majority vote classifier. Note that $|F|_\infty \leq 1$.

As shown before, in boosting, we have:

$$g = \nabla \widehat{R}_{n,\phi}(f_\alpha) = \frac{1}{n}\sum_{i=1}^n \phi'(-y_i f_\alpha(x_i))(-y_i)F(x_i)\,,$$

Since $|F|_\infty \leq 1$ and $|y|_\infty \leq 1$, then $|g|_\infty \leq L$ where L is the Lipschitz constant of ϕ (e.g., a constant like e or 2).

$$\widehat{R}_{n,\phi}(f_{\alpha_k^\circ}) - \min_{\alpha \in \Delta_n} \widehat{R}_{n,\phi}(f_\alpha) \leq L\sqrt{\frac{2\log(N)}{k}}$$

We need the number of iterations $k \approx n^2 \log(N)$.

The functions f_j's could hit all the vertices. Thus, if we want to fit them in a ball, the ball has to be radius $\sqrt{N}$. This is why the projected gradient descent would give the rate of $\sqrt{\frac{N}{k}}$. But by looking at the gradient we can determine the right geometry. In this case, the gradient is bounded by sup-norm which is usually the most constraining norm in projected

gradient descent. Thus, using Mirror descent would be most beneficial.

Other Potential Functions:

There are other potential functions which are strongly convex w.r.t ℓ_1 norm. In particular, for

$$\Phi(x) = \frac{1}{p}|x|_p^p, \qquad p = 1 + \frac{1}{\log(d)}$$

then Φ is $c\sqrt{\log(d)}$-strongly convex w.r.t ℓ_1 norm.

18.657 Mathematics of Machine Learning
Fall 2015