18.657: Mathematics of Machine Learning

Lecturer: Philippe Rigollet — Lecture 12
Scribe: Michael Traub — Oct. 19, 2015

2.3 Projected Gradient Descent

In the original gradient descent formulation, we hope to optimize $\min_{x\in\mathcal{C}} f(x)$ where $\mathcal{C}$ and f are convex, but we did not constrain the intermediate x_k. Projected gradient descent will incorporate this condition.

2.3.1 Projection onto Closed Convex Set

First we must establish that it is possible to always be able to keep x_k in the convex set $\mathcal{C}$. One approach is to take the closest point $\pi(x_k) \in \mathcal{C}$.

Definition: Let $\mathcal{C}$ be a closed convex subset of $\mathbb{R}^d$. Then $\forall\, x \in \mathbb{R}^d$, let $\pi(x) \in \mathcal{C}$ be the minimizer of

$$\|x - \pi(x)\| = \min_{z\in\mathcal{C}} \|x - z\|$$

where $\|\cdot\|$ denotes the Euclidean norm. Then $\pi(x)$ is unique and,

$$\langle \pi(x) - x, \pi(x) - z\rangle \le 0 \quad \forall\, z \in \mathcal{C} \tag{2.1}$$

Proof. From the definition of $\pi := \pi(x)$, we have $\|x-\pi\|^2 \le \|x - v\|^2$ for any $v \in \mathcal{C}$. Fix $w \in \mathcal{C}$ and define $v = (1-t)\pi + tw$ for $t \in (0, 1]$. Observe that since $\mathcal{C}$ is convex we have $v \in \mathcal{C}$ so that

$$\|x-\pi\|^2 \le \|x-v\|^2 = \|x - \pi - t(w-\pi)\|^2$$

Expanding the right-hand side yields

$$\|x-\pi\|^2 \le \|x-\pi\|^2 - 2t\,\langle x-\pi, w-\pi\rangle + t^2\,\|w-\pi\|^2$$

This is equivalent to

$$\langle x-\pi, w-\pi\rangle \le t\,\|w-\pi\|^2$$

Since this is valid for all $t \in (0, 1)$, letting $t \to 0$ yields (2.1).

Proof of Uniqueness. Assume π_1, $\pi_2 \in \mathcal{C}$ satisfy

$$\langle \pi_1 - x, \pi_1 - z\rangle \le 0 \quad \forall\, z \in C$$
$$\langle \pi_2 - x, \pi_2 - z\rangle \le 0 \quad \forall\, z \in C$$

Taking $z = \pi_2$ in the first inequality and $z = \pi_1$ in the second, we get

$$\langle \pi_1 - x, \pi_1 - \pi_2\rangle \le 0$$
$$\langle x - \pi_2, \pi_1 - \pi_2\rangle \le 0$$

Adding these two inequalities yields $\|\pi_1 - \pi_2\|^2 \le 0$ so that $\pi_1 = \pi_2$. ∎

2.3.2 Projected Gradient Descent

Algorithm 1 Projected Gradient Descent algorithm

Input: $x_1 \in \mathcal{C}$, positive sequence $\{\eta_s\}_{s\geq 1}$
for $s = 1$ to $k-1$ **do**
$y_{s+1} = x_s - \eta_s g_s\,, \quad g_s \in \partial f(x_s)$
$x_{s+1} = \pi(y_{s+1})$
end for
return Either $\bar{x} = \frac{1}{k}\sum_{s=1}^{k} x_s$ or $x^\circ \in \underset{x\in\{x_1,\dots,x_k\}}{\operatorname{argmin}} f(x)$

Theorem: Let $\mathcal{C}$ be a closed, nonempty convex subset of $\mathbb{R}^d$ such that $\operatorname{diam}(\mathcal{C}) \leq R$. Let f be a convex L-Lipschitz function on $\mathcal{C}$ such that $x^* \in \operatorname{argmin}_{x\in\mathcal{C}} f(x)$ exists. Then if $\eta_s \equiv \eta = \frac{R}{L\sqrt{k}}$ then

$$f(\bar{x}) - f(x^*) \leq \frac{LR}{\sqrt{k}} \quad \text{and} \quad f(\bar{x}^\circ) - f(x^*) \leq \frac{LR}{\sqrt{k}}$$

Moreover, if $\eta_s = \frac{R}{L\sqrt{s}}$, then $\exists c > 0$ such that

$$f(\bar{x}) - f(x^*) \leq c\frac{LR}{\sqrt{k}} \quad \text{and} \quad f(\bar{x}^\circ) - f(x^*) \leq c\frac{LR}{\sqrt{k}}$$

Proof. Again we will use the identity that $2a^\top b = \|a\|^2 + \|b\|^2 - \|a-b\|^2$.

By convexity, we have

$$\begin{aligned} f(x_s) - f(x^*) &\leq g_s^\top (x_s - x^*) \\ &= \frac{1}{\eta}(x_s - y_{s+1})^\top (x_s - x^*) \\ &= \frac{1}{2\eta}\left[\|x_s - y_{s+1}\|^2 + \|x_s - x^*\|^2 - \|y_{s+1} - x^*\|^2\right] \end{aligned}$$

Next,

$$\begin{aligned} \|y_{s+1} - x^*\|^2 &= \|y_{s+1} - x_{s+1}\|^2 + \|x_{s+1} - x^*\|^2 + 2\left\langle y_{s+1} - x_{s+1}, x_{s+1} - x^*\right\rangle \\ &= \|y_{s+1} - x_{s+1}\|^2 + \|x_{s+1} - x^*\|^2 + 2\left\langle y_{s+1} - \pi(y_{s+1}), \pi(y_{s+1}) - x^*\right\rangle \\ &\geq \|x_{s+1} - x^*\|^2 \end{aligned}$$

where we used that $\langle x - \pi(x), \pi(x) - z\rangle \geq 0 \ \forall\, z \in \mathcal{C}$, and $x^* \in \mathcal{C}$. Also notice that $\|x_s - y_{s+1}\|^2 = \eta^2\|g_s\|^2 \leq \eta^2 L^2$ since f is L-Lipschitz with respect to $\|\cdot\|$. Using this we find

$$\begin{aligned} \frac{1}{k}\sum_{s=1}^{k} f(x_s) - f(x^*) &\leq \frac{1}{k}\sum_{s=1}^{k}\frac{1}{2\eta}\left[\eta^2 L^2 + \|x_s - x^*\|^2 - \|x_{s+1} - x^*\|^2\right] \\ &\leq \frac{\eta L^2}{2} + \frac{1}{2\eta k}\|x_1 - x^*\|^2 \leq \frac{\eta L^2}{2} + \frac{R^2}{2\eta k} \end{aligned}$$

Minimizing over η we get $\frac{L^2}{2} = \frac{R^2}{2\eta^2 k} \implies \eta = \frac{R}{L\sqrt{k}}$, completing the proof

$$f(\bar{x}) - f(x^*) \le \frac{RL}{\sqrt{k}}$$

Moreover, the proof of the bound for $f(\sum_{s=\frac{k}{2}}^{k} x_s) - f(x^*)$ is identical because $\left\|x_{\frac{k}{2}} - x^*\right\|^2 \le R^2$ as well. $\square$

2.3.3 Examples

Support Vector Machines

The SVM minimization as we have shown before is

$$\min_{\substack{\alpha \in \mathbb{R}^n \\ \alpha^\top \mathbb{K}\alpha \le C^2}} \frac{1}{n} \sum_{i=1}^{n} \max\left(0, 1 - Y_i f_\alpha(X_i)\right)$$

where $f_\alpha(X_i) = \alpha^\top \mathbb{K} e_i = \sum_{j=1}^n \alpha_j K(X_j, X_i)$. For convenience, call $g_i(\alpha) = \max\left(0, 1 - Y_i f_\alpha(X_i)\right)$. In this case executing the projection onto the ellipsoid $\{\alpha : \alpha^\top \mathbb{K}\alpha \le C^2\}$ is not too hard, but we do not know about C, R, or L. We must determine these we can know that our bound is not exponential with respect to n. First we find L and start with the gradient of $g_i(\alpha)$:

$$\nabla g_i(\alpha) = \mathbb{1}(1 - Y_i f_\alpha(X_i) \ge 0) Y_i \mathbb{K} e_i$$

With this we bound the gradient of the φ-risk $\hat{R}_{n,\varphi}(f_\alpha) = \frac{1}{n}\sum_{i=1}^n g_i(\alpha)$.

$$\left\| \frac{\partial}{\partial \alpha} \hat{R}_{n,\varphi}(f_\alpha) \right\| = \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla g_i(\alpha) \right\| \le \frac{1}{n} \sum_{i=1}^{n} \|\mathbb{K} e_i\|_2$$

by the triangle inequality and the fact that that $\mathbb{1}(1 - Y_i f_\alpha(X_i) \ge 0) Y_i \le 1$. We can now use the properties of our kernel K. Notice that $\|\mathbb{K} e_i\|$ is the ℓ_2 norm of the i^{th} column so $\|\mathbb{K} e_i\|_2 = \left(\sum_{j=1}^n K(X_j, X_i)^2\right)^{\frac{1}{2}}$. We also know that

$$K(X_j, X_i)^2 = \langle K(X_j, \cdot), K(X_i, \cdot)\rangle \le \|K(X_j, \cdot)\|_H \|K(X_i, \cdot)\|_H \le k_{\max}^2$$

Combining all of these we get

$$\left\| \frac{\partial}{\partial \alpha} \hat{R}_{n,\varphi}(f_\alpha) \right\| \le \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{n} k_{\max}^2 \right)^{\frac{1}{2}} = k_{\max}\sqrt{n} = L$$

To find R we try to evaluate $\operatorname{diam}\{\alpha^\top \mathbb{K}\alpha \le C^2\} = 2 \max_{\alpha^\top \mathbb{K}\alpha \le C^2} \sqrt{\alpha^\top \alpha}$. We can use the condition to put bounds on the diameter

$$C^2 \ge \alpha^\top \mathbb{K}\alpha \ge \lambda_{\min}(\mathbb{K}) \alpha^\top \alpha \implies \operatorname{diam}\{\alpha^\top \mathbb{K}\alpha \le C^2\} \le \frac{2C}{\sqrt{\lambda_{\min}(\mathbb{K})}}$$

We need to understand how small $\lambda_{\min}$ can get. While it is true that these exist random samples selected by an adversary that make $\lambda_{\min} = 0$, we will consider a random sample of

$X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, I_d)$. This we can write these d-dimensional samples as a $d \times n$ matrix $\mathbb{X}$. We can rewrite the matrix $\mathbb{K}$ with entries $\mathbb{K}_{ij} = K(X_i, X_j) = \langle X_i, X_j \rangle_{\mathbb{R}^d}$ as a Wishart matrix $\mathbb{K} = \mathbb{X}^\top \mathbb{X}$ (in particular, $\frac{1}{d}\mathbb{X}^\top \mathbb{X}$ is Wishart). Using results from random matrix theory, if we take $n, d \to \infty$ but hold $\frac{n}{d}$ as a constant γ, then $\lambda_{\min}(\frac{\mathbb{K}}{d}) \to (1 - \sqrt{\gamma})^2$. Taking an approximation since we cannot take n, d to infinity, we get

$$\lambda_{\min}(\mathbb{K}) \simeq d\left(1 - 2\sqrt{\frac{n}{d}}\right) \geq \frac{d}{2}$$

using the fact that $d \gg n$. This means that $\lambda_{\min}$ becoming too small is not a problem when we model our samples as coming from multivariate Gaussians.

Now we turn our focus to the number of iterations k. Looking at our bound on the excess risk

$$\hat{R}_{n,\varphi}(f_{\alpha_R^\circ}) \leq \min_{\alpha^\top \mathbb{K}\alpha \leq C^2} \hat{R}_{n,\varphi}(f_\alpha) + C\sqrt{\frac{n}{k\lambda_{\min}(\mathbb{K})}}k_{\max}$$

we notice that our all of the constants in our stochastic term can be computed given the number of points and the kernel. Since statistical error is often $\frac{1}{\sqrt{n}}$, to be generous we want to have precision up to $\frac{1}{n}$ to allow for fast rates in special cases. This gives us

$$k \geq \frac{n^3 k_{\max}^2 C^2}{\lambda_{\min}(\mathbb{K})}$$

which is not bad since n is often not very big.

In [Bub15], the rates for many a wide rage of problems with various assumptions are available. For example, if we assume strong convexity and Lipschitz we can get an exponential rate so $k \sim \log n$. If gradient is Lipschitz, then we get get $\frac{1}{k}$ instead of $\frac{1}{\sqrt{k}}$ in the bound. However, often times we are not optimizing over functions with these nice properties.

Boosting

We already know that φ is L-Lipschitz for boosting because we required it before. Remember that our optimization problem is

$$\min_{\substack{\alpha \in \mathbb{R}^N \\ |\alpha|_1 \leq 1}} \frac{1}{n}\sum_{i=1}^n \varphi(-Y_i f_\alpha(X_i))$$

where $f_\alpha = \sum_{j=1}^N \alpha_j f_j$ and f_j is the j^{th} weak classifier. Remember before we had some rate like $c\sqrt{\frac{\log N}{n}}$ and we would hope to get some other rate that grows with $\log N$ since N can be very large. Taking the gradient of the φ-loss in this case we find

$$\nabla \hat{R}_{n,\varphi}(f_\alpha) = \frac{1}{n}\sum_{i=1}^N \varphi'(-Y_i f_\alpha(X_i))(-Y_i)F(X_i)$$

where $F(x)$ is the column vector $[f_1(x), \ldots, f_N(x)]^\top$. Since $|Y_i| \leq 1$ and $\varphi' \leq L$, we can bound the ℓ_2 norm of the gradient as

$$\begin{aligned}\left\|\nabla \hat{R}_{n,\varphi}(f_\alpha)\right\|_2 &\leq \frac{L}{n}\left\|\sum_{i=1}^n F(X_i)\right\| \\ &\leq \frac{L}{n}\sum_{i=1}^n \|F(X_i)\| \leq L\sqrt{N}\end{aligned}$$

using triangle inequality and the fact that $F(X_i)$ is a N-dimensional vector with each component bounded in absolute value by 1.

Using the fact that the diameter of the ℓ_1 ball is 2, $R = 2$ and the Lipschitz associated with our φ-risk is $L\sqrt{N}$ where L is the Lipschitz constant for φ. Our stochastic term $\frac{RL}{\sqrt{k}}$ becomes $2L\sqrt{\frac{N}{k}}$. Imposing the same $\frac{1}{n}$ error as before we find that $k \sim N^2 n$, which is very bad especially since we want $\log N$.

2.4 Mirror Descent

Boosting is an example of when we want to do gradient descent on a non-Euclidean space, in particular a ℓ_1 space. While the dual of the ℓ_2-norm is itself, the dual of the ℓ_1 norm is the ℓ_∞ or sup norm. We want this appear if we have an ℓ_1 constraint. The reason for this is not intuitive because we are taking about measures on the same space $\mathbb{R}^d$, but when we consider optimizations on other spaces we want a procedure that does is not indifferent to the measure we use. Mirror descent accomplishes this.

2.4.1 Bregman Projections

Definition: If $\|\cdot\|$ is some norm on $\mathbb{R}^d$, then $\|\cdot\|_*$ is its dual norm.

Example: If dual norm of the ℓ_p norm $\|\cdot\|_p$ is the ℓ_q norm $\|\cdot\|_q$, then $\frac{1}{p} + \frac{1}{q} = 1$. This is the limiting case of Hölder's inequality.

In general we can also refine our bounds on inner products in $\mathbb{R}^d$ to $x^\top y \le \|x\| \, \|y\|_*$ if we consider x to be the primal and y to be the dual. Thinking like this, gradients live in the dual space, e.g. in $g_s^\top(x - x^*)$, $x - x^*$ is in the primal space, so g_s is in the dual. The transpose of the vectors suggest that these vectors come from spaces with different measure, even though all the vectors are in $\mathbb{R}^d$.

Definition: Convex function Φ on a convex set D is said to be
(i) L-Lipschitz with respect to $\|\cdot\|$ if $\|g\|_* \le L \quad \forall g \in \partial\Phi(x) \quad \forall x \in D$
(ii) α-strongly convex with respect to $\|\cdot\|$ if

$$\Phi(y) \ge \Phi(x) + g^\top(y - x) + \frac{\alpha}{2}\|y - x\|^2$$

for all $x, y \in D$ and for $g \in \partial f(x)$

Example: If Φ is twice differentiable with Hessian H and $\|\cdot\|$ is the ℓ_2 norm, then all $\text{eig}(H) \ge \alpha$.

Definition (Bregman divergence): For a given convex function Φ on a convex set $\mathcal{D}$ with $x, y \in \mathcal{D}$, the Bregman divergence of y from x is defined as

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

This divergence is the error of the function $\Phi(y)$ from the linear approximation at x. Also note that this quantity is not symmetric with respect to x and y. If Φ is convex then $D_\Phi(y,x) \geq 0$ because the Hessian is positive semi-definite. If Φ is α-strongly convex then $D_\Phi(y,x) \geq \frac{\alpha}{2}\|y-x\|^2$ and if the quadratic approximation is good then this approximately holds in equality and this divergence behaves like Euclidean norm.

Proposition: Given convex function Φ on $\mathcal{D}$ with $x, y, z \in \mathcal{D}$

$$(\nabla\Phi(x) - \nabla\Phi(y))^\top (x-z) = D_\Phi(x,y) + D_\Phi(z,x) - D_\Phi(z,y)$$

Proof. Looking at the right hand side

$$\begin{aligned}
&= \Phi(x) - \Phi(y) - \nabla\Phi(y)^\top(x-y) + \Phi(z) - \Phi(x) - \nabla\Phi(x)^\top(z-x) \\
&\qquad - \left[\Phi(z) - \Phi(y) - \nabla\Phi(y)^\top(z-y)\right] \\
&= \nabla\Phi(y)^\top(y-x+z-y) - \nabla\Phi(x)^\top(z-x) \\
&= (\nabla\Phi(x) - \nabla\Phi(y))^\top (x-z)
\end{aligned}$$

□

Definition (Bregman projection): Given $x \in \mathbb{R}^d$, Φ a convex differentiable function on $\mathcal{D} \subset \mathbb{R}^d$ and convex $C \subset \bar{\mathcal{D}}$, the Bregman projection of x with respect to Φ is

$$\pi^\Phi(x) \in \operatorname*{argmin}_{z \in C} D_\phi(x,z)$$

References

[Bub15] Sébastien Bubeck, *Convex optimization: algorithms and complexity*, Now Publishers Inc., 2015.