2.3 Projected Gradient Descent

In the original gradient descent formulation, we hope to optimize \( \min_{x \in C} f(x) \) where \( C \) 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 \( C \).

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

\[
\|x - \pi(x)\| = \min_{z \in C} \|x - z\|
\]

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

\[
\langle \pi(x) - x, \pi(x) - z \rangle \leq 0 \quad \forall \ z \in C \tag{2.1}
\]

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

\[
\|x - \pi\|^2 \leq \|x - v\|^2 = \|x - \pi - t(w - \pi)\|^2
\]

Expanding the right-hand side yields

\[
\|x - \pi\|^2 \leq \|x - \pi\|^2 - 2t \langle x - \pi, w - \pi \rangle + t^2 \|w - \pi\|^2
\]

This is equivalent to

\[
\langle x - \pi, w - \pi \rangle \leq t \|w - \pi\|^2
\]

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

**Proof of Uniqueness.** Assume \( \pi_1, \pi_2 \in C \) satisfy

\[
\langle \pi_1 - x, \pi_1 - z \rangle \leq 0 \quad \forall \ z \in C \\
\langle \pi_2 - x, \pi_2 - z \rangle \leq 0 \quad \forall \ z \in C
\]

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

\[
\langle \pi_1 - x, \pi_1 - \pi_2 \rangle \leq 0 \\
\langle x - \pi_2, \pi_1 - \pi_2 \rangle \leq 0
\]

Adding these two inequalities yields \( \|\pi_1 - \pi_2\|^2 \leq 0 \) so that \( \pi_1 = \pi_2 \). \( \square \)
2.3.2 Projected Gradient Descent

Algorithm 1 Projected Gradient Descent algorithm

Input: $x_1 \in C$, positive sequence $\{\eta_s\}_{s \geq 1}$
for $s = 1$ to $k - 1$ do

$y_{s+1} = x_s - \eta_s g_s$, $g_s \in \partial f(x_s)$
$x_{s+1} = \pi(y_{s+1})$

end for

return Either $\bar{x} = \frac{1}{k} \sum_{s=1}^{k} x_s$ or $x^* \in \arg\min_{x \in \{x_1, \ldots, x_k\}} f(x)$

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

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

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

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

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

By convexity, we have

$$f(x_s) - f(x^*) \leq g_s^\top (x_s - x^*)$$

$$= \frac{1}{\eta} (x_s - y_{s+1})^\top (x_s - x^*)$$

$$= \frac{1}{2\eta} \left[ \|x_s - y_{s+1}\|^2 + \|x_s - x^*\|^2 - \|y_{s+1} - x^*\|^2 \right]$$

Next,

$$\|y_{s+1} - x^*\|^2 = \|y_{s+1} - x_{s+1}\|^2 + \|x_{s+1} - x^*\|^2 + 2 \langle y_{s+1} - x_{s+1}, x_{s+1} - x^* \rangle$$

$$= \|y_{s+1} - x_{s+1}\|^2 + \|x_{s+1} - x^*\|^2 + 2 \langle y_{s+1} - \pi(y_{s+1}), \pi(y_{s+1}) - x^* \rangle$$

$$\geq \|x_{s+1} - x^*\|^2$$

where we used that $\langle x - \pi(x), \pi(x) - z \rangle \geq 0 \ \forall z \in C$, and $x^* \in C$. 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

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

$$\leq \frac{\eta L^2}{2} + \frac{1}{2\eta k} \|x_1 - x^*\|^2 \leq \frac{\eta L^2}{2} + \frac{R^2}{2\eta k}$$
Minimizing over $\eta$ we get $\frac{L^2}{2} = \frac{R^2}{2\eta_k} \implies \eta = \frac{R}{L\sqrt{k}}$, completing the proof

\[ f(\hat{x}) - f(x^*) \leq \frac{RL}{\sqrt{k}} \]

Moreover, the proof of the bound for $f(\sum_{s=\frac{1}{2}}^{k} x_s) - f(x^*)$ is identical because $\|x_{\frac{1}{2}} - x^*\|^2 \leq R^2$ as well. \hfill \square

2.3.3 Examples

**Support Vector Machines**

The SVM minimization as we have shown before is

\[
\min_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} \max(0, 1 - Y_i f_{\alpha}(X_i))
\]

where $f_{\alpha}(X_i) = \alpha^T K e_i = \sum_{j=1}^{n} \alpha_j K(X_j, X_i)$. For convenience, call $g_i(\alpha) = \max(0, 1 - Y_i f_{\alpha}(X_i))$.

In this case executing the projection onto the ellipsoid $\{ \alpha : \alpha^T \mathbb{K} \alpha \leq C^2 \}$ 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 $g_i(\alpha)$:

\[ \nabla g_i(\alpha) = \mathbb{I}(1 - Y_i f_{\alpha}(X_i) \geq 0)Y_i \mathbb{K} e_i \]

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

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

by the triangle inequality and the fact that that $\mathbb{I}(1 - Y_i f_{\alpha}(X_i) \geq 0)Y_i \leq 1$. We can now use the properties of our kernel $K$. Notice that $\| \mathbb{K} e_i \|$ is the $\ell_2$ 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

\[ K(X_j, X_i)^2 = \langle K(X_j, \cdot), K(X_i, \cdot) \rangle \leq \|K(X_j, \cdot)\|_H \|K(X_i, \cdot)\|_H \leq k_{\text{max}}^2 \]

Combining all of these we get

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

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

\[ C^2 \geq \alpha^T \mathbb{K} \alpha \geq \lambda_{\min}(\mathbb{K}) \alpha^T \alpha \implies \text{diam}\{ \alpha^T \mathbb{K} \alpha \leq C^2 \} \leq \frac{2C}{\sqrt{\lambda_{\min}(\mathbb{K})}} \]

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

\[
\lambda_{\text{min}}(I K) \simeq d \left( 1 - 2 \sqrt{\frac{n}{d}} \right) \geq \frac{d}{2}
\]

using the fact that \( d \gg n \). This means that \( \lambda_{\text{min}} \) 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

\[
\hat{R}_{n, \varphi}(f_{\alpha^*}) \leq \min_{\alpha : \|\alpha\| \leq C^2} \hat{R}_{n, \varphi}(f_{\alpha}) + C \sqrt{\frac{n}{k \lambda_{\text{min}}(I K)}} k_{\text{max}}
\]

we notice that 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 \( n \) to have precision up to \( \frac{1}{n} \) to allow for fast rates in special cases. This gives us

\[
k \geq \frac{n^3 k_{\text{max}} C^2}{\lambda_{\text{min}}(I K)}
\]

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 \( \frac{1}{k} \) 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 \( \varphi \) is \( L \)-Lipschitz for boosting because we required it before. Remember that our optimization problem is

\[
\min_{\alpha \in \mathbb{R}^N} \frac{1}{n} \sum_{i=1}^{n} \varphi(-Y_i f_{\alpha}(X_i))
\]

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 \( \varphi \)-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 \( \ell_2 \) norm of the gradient as

\[
\| \nabla \hat{R}_{n, \varphi}(f_{\alpha}) \|_2 \leq \frac{L}{n} \left\| \sum_{i=1}^{n} F(X_i) \right\| \leq \frac{L}{n} \sum_{i=1}^{n} \| F(X_i) \| \leq L \sqrt{N}
\]
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 $\ell_1$ ball is 2, $R = 2$ and the Lipschitz associated with our $\varphi$-risk is $L\sqrt{N}$ where $L$ is the Lipschitz constant for $\varphi$. Our stochastic term $\frac{KL}{\sqrt{N}}$ 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 $\ell_1$ space. While the dual of the $\ell_2$-norm is itself, the dual of the $\ell_1$ norm is the $\ell_\infty$ or sup norm. We want this appear if we have an $\ell_1$ 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 $\ell_p$ norm $\|\cdot\|_p$ is the $\ell_q$ 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 $x^\top y \leq \|x\| \|y\|_*$ 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 $g_s^\top (x - x^\ast)$, $x - x^\ast$ 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 $\Phi$ on a convex set $D$ is said to be
(i) L-Lipschitz with respect to $\|\cdot\|_*$ if $\|g\|_* \leq L$ $\forall g \in \partial \Phi(x)$ $\forall x \in D$
(ii) $\alpha$-strongly convex with respect to $\|\cdot\|$ if

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

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

**Example:** If $\Phi$ is twice differentiable with Hessian $H$ and $\|\cdot\|$ is the $\ell_2$ norm, then all $\text{eig}(H) \geq \alpha$.

**Definition (Bregman divergence):** For a given convex function $\Phi$ 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 $\Phi$ is convex then $D_\Phi(y, x) \geq 0$ because the Hessian is positive semi-definite. If $\Phi$ is $\alpha$-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 $\Phi$ 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

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

\[\square\]

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

\[\pi^\Phi(x) \in \arg\min_{z \in C} D_\Phi(x, z)\]

**References**
