Intuitions for Optimization

pluskid

October 12, 2015

1 Projected Subgradient Descent for Lipschitz Functions

1.1 Problem Setup

Problem 1. *Given $f : \mathbb{R}^n \to \mathbb{R}$, and $\mathcal{K} \subset \mathbb{R}^n$. Assume $\mathcal{K}$ is compact and convex, included in a Euclidean ball of radius R:*

$$\mathcal{K} \subset \mathcal{B}_2(0; R) \tag{1}$$

and f is convex and L-Lipschitz on $\mathcal{K}$:

$$|f(x) - f(y)| \leq L\|x - y\|_2, \quad x, y \in \mathcal{K} \tag{2}$$

Find a minimizer of f on $\mathcal{K}$:

$$\operatorname*{minimize}_{x \in \mathcal{K}} f(x)$$

1.2 Algorithm and its Bounds

Algorithm 1 Projected Subgradient Descent

randomly initialize $x^0 \in \mathcal{K}$
for $t \leftarrow 0, \ldots, T-1$ **do**
    $y^{t+1} \leftarrow x^t - \eta_t g_t$, where $g_t \in \partial f(x^t)$
    $x^{t+1} \leftarrow \Pi_{\mathcal{K}}(y^{t+1})$
end for

Theorem 1. *Running Algorithm 1 on Problem 1 for T iterations gives*

$$\min_{0 \leq \tau \leq T} f(x^\tau) - f(x^*) \leq \frac{R^2 + G^2 \sum_{t=0}^{T-1} \eta_t^2}{\sum_{t=0}^{T-1} \eta_t} \tag{3}$$

In general, the optimal bound is around $O(GR/\sqrt{T})$ with stepsizes around $\eta_t \approx R/(G\sqrt{t})$. That means in order to get an approximate error of ε, we will need to run the algorithm for $O(1/\varepsilon^2)$ iterations.

1.3 Intuitions and Analysis

1.3.1 Problem Assumptions

We consider the unconstrained case first, i.e. $\mathcal{K} = \mathbb{R}^n$. Since f is convex know that x^* is a minimizer of f if and only if $0 \in \partial f(x^*)$. However, without making any extra assumptions, we have no idea of the behavior of f even in a small neighborhood of x^*. Consider for example $f(x) = C|x|$, with $C > 0$ a large constant. Assume we are currently very close to the optimal $x^* = 0$, say $x^t = \varepsilon$, $\varepsilon > 0$. f is differentiable at ε, so the only subgradient is $g_t = C$. Therefore,

$$x^{t+1} = x^t - \eta_t g_t = \varepsilon - \eta_t C < -\varepsilon, \quad \forall \eta_t > \frac{2\varepsilon}{C}$$

As we can see, unless the stepsize η_t is very tiny, we will overshoot, $f(x^{t+1}) > f(x^t)$. Moreover, if we use a constant stepsize $\eta_t = \eta$, then if $\eta > \varepsilon/C$, we will be jumping back and forth at ε and $\varepsilon - \eta C$ indefinitely.

In order to fix this, we need to make additional assumptions. We will see later in the case of smooth functions, the gradient changes continuously. So we know that at a local neighborhood of the optimal (gradient is 0), the gradient is also small. But here, we are working with non-differentiable functions, we will just assume f is L-Lipschitz.

Note f being L-Lipschitz in $\mathcal{K}$ implies that $\|g\|_2 \leq L$, $\forall g \in \partial f(x), \forall x \in \mathcal{K}$. In order to avoid overshooting, we will have to move with tiny stepsizes.

1.3.2 Convergence Analysis

By the property of subgradient, we have

$$0 \leq f(x^t) - f(x^*) \leq g_t^\top (x^t - x^*) = -g_t^\top (x^* - x^t) \tag{4}$$

where the left hand side is due to the optimality of x^*. This inequality indicates that the vector $x^* - x^t$ is non-negatively correlated with $-g_t$, the direction of a subgradient descent.

Unlike in the case of differentiable functions, in which we can guarantee that the function values decreases when moving along the direction of negative gradient with a small enough step size; here we do not know much about the function values. But as we can see from Figure 1, when the angle between $-g_t$ and $x^* - x^t$ is greater than or equal to $\pi/2$, we will move away from x^* with any positive step size. However, if $f(x^t) - f(x^*) > 0$, i.e. we are not already at the optimal, the angle is strictly less than $\pi/2$, so if we move with a small enough step size, we will get closer to x^*. Algebraically,

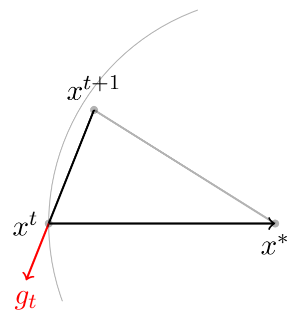

Figure 1: Demonstration of subgradient descent.

$$\begin{aligned}\|x^{t+1} - x^*\|^2 = \|x^{t+1} - x^t + x^t - x^*\|^2 &= \eta_t^2 \|g_t\|^2 + \|x^t - x^*\|^2 - 2\eta_t g_t^\top (x^t - x^*) \\ &\leq \eta_t^2 \|g_t\|^2 + \|x^t - x^*\|^2 - 2\eta_t \left(f(x^t) - f(x^*)\right)\end{aligned} \tag{5}$$

Since $\|g_t\| \leq L$ by our assumption, the red term decays quadratically, while the blue term only decays linearly. So when η_t is small enough, we will have $\|x^{t+1} - x^*\|^2 \leq \|x^t - x^*\|^2$. Furthermore, the progress we make by moving towards x^* is characterized by $f(x^t) - f(x^*)$. So if we are still far away from the optimal, we will be making quite a lot progress in each step. On the other hand, when $f(x^t) - f(x^*)$ is small, our progress might be small, but at that point we are already close to the optimal function value $f(x^*)$.

Actually, to get a bound on the algorithm, we can just sum up the previous inequality for all $t = 0, \ldots, T-1$,

$$\begin{aligned} 0 \leq \|x^T - x^*\|^2 &\leq \|x^0 - x^*\|^2 + \sum_{t=0}^{T-1} \eta_t^2 \|g_t\|^2 - 2 \sum_{t=0}^{T-1} \eta_t \left(f(x^t) - f(x^*) \right) \\ &\leq R^2 + G^2 \sum_{t=0}^{T-1} \eta_t^2 - 2 \left(\min_{0 \leq \tau \leq T} f(x^\tau) - f(x^*) \right) \sum_{t=0}^{T-1} \eta_t \end{aligned}$$

It then implies

$$\min_{0 \leq \tau \leq T} f(x^\tau) - f(x^*) \leq \frac{R^2 + G^2 \sum_{t=0}^{T-1} \eta_t^2}{\sum_{t=0}^{T-1} \eta_t} \tag{6}$$

which proved Theorem 1 for the case of unconstrained optimization ($\mathcal{K} = \mathbb{R}^n$).

1.4 Choosing Stepsizes

Note (3) holds for any choices of stepsizes η_t (though some of them will give completely trivial bounds). So we could actually optimize the right hand side to get an "optimal" bound. To make the problem easier, we choose a fixed stepsize $\eta_t = \eta$ for $t = 0, \ldots, T-1$. So the right hand side becomes

$$\frac{R^2 + G^2 T \eta^2}{T\eta} = \frac{R^2}{T\eta} + G^2 \eta \geq \frac{2GR}{\sqrt{T}} \tag{7}$$

where the inequality holds with equality when

$$\eta = \frac{R}{G\sqrt{T}} \tag{8}$$

Note the choice of step size depends on several factors:

G: As we described in Section 1.3.1, large G will force us to be careful and move with small stepsizes. Our intuition is consistent here.

R: In our analysis, we only use R to bound $\|x^0 - x^*\|$. When R is large, we want to use large step size, otherwise we might never reach the optimal in the given time budget T. Generally when x^* is unknown, R can be bounded by the size of $\mathcal{K}$ for the case of constrained optimization.

T: The inverse dependency on T can be interpreted as: when having a large time budget, we can be a little bit more careful and move slowly.

However, in general, the fact that the stepsize depends on the total number of iterations is strange. That means if I want to compute more iterations, I will have to start over again and use a different stepsize if I want to bound the performance with formula.

In general, we will prefer to use a decaying learning rate. Specifically, as long as $\sum_t \eta_t \to \infty$ and $\sum_t \eta_t^2$ is bounded or approaches infinity at a slower rate than $\sum_t \eta_t$, (3) will give a reasonable bound. For example, take $\eta_t = R/(G\sqrt{t+1})$, since

$$\sum_{t=0}^{T-1} \frac{1}{t+1} \le 1 + \int_1^T \frac{1}{x}\,dx = 1 + \log T$$
$$\sum_{t=0}^{T-1} \frac{1}{\sqrt{t+1}} \ge \int_1^{T+1} \frac{1}{\sqrt{x}}\,dx = 2\sqrt{T+1} - 2$$

Plug-in to (3), we get

$$\min_{0\le\tau\le T} f(x^\tau) - f(x^*) \le GR\frac{1+\log T}{2\sqrt{T+1}-2} \lesssim \frac{GR\log T}{\sqrt{T}} \tag{9}$$

Comparing with the optimal bound we get with a fixed stepsize in (7), we lose a factor of $\log T$, but our stepsize does not depend on the total number of iterations any more.

1.4.1 Constrained Optimization

In the constrained case, we have an extra projection step. However, the same analysis naturally goes through because projection into convex set is a *contraction*. Specifically, we have the following lemma.

Lemma 1. *Let $\mathcal{K} \subset \mathbb{R}^n$ be a closed convex set, let $x \in \mathcal{K}$ and $y \in \mathbb{R}^n$. Then*

$$(x - \Pi_{\mathcal{K}}(y))^\top (y - \Pi_{\mathcal{K}}(y)) \le 0 \tag{10}$$

which also implies

$$\|x - \Pi_{\mathcal{K}}(y)\|^2 + \|y - \Pi_{\mathcal{K}}(y)\|^2 \le \|y - x\|^2 \tag{11}$$

This lemma could be proved with *supporting hyperplane theorem* of convex sets. (11) implies

$$\|x - \Pi_{\mathcal{K}}\|^2 \le \|y - x\|^2$$

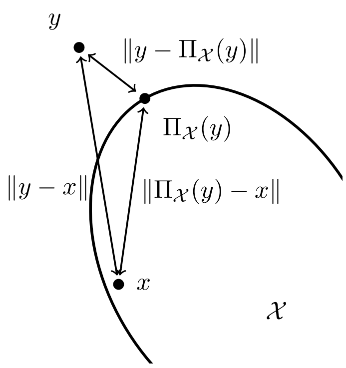

Figure 2: Illustration of convex projection. Figure source: Sébastien Bubeck, *Theory of Convex Optimization for Machine Learning*.

So if we go back to our analysis in Section 1.3.2, the only thing we need to modify is (5). Since $x^* \in \mathcal{K}$,

$$\|x^{t+1} - x^*\|^2 = \|\Pi_{\mathcal{K}}(y^{t+1}) - x^*\|^2 \le \|y^{t+1} - x^*\|^2 = \eta_t^2\|g_t\|^2 + \|x^t - x^*\|^2 - 2\eta_t g_t^\top (x^t - x^*)$$

and the rest of the analysis follows as before. So for constrained optimization, we get the same convergence rate as the unconstrained case.

2 Gradient Descent for Smooth Function

In this section, we look at smooth functions. Specifically, f is differentiable, and its gradient ∇f is Lipschitz continuous. By adding those assumptions, we know better about f than in the simple Lipschitz case. For example, we know two nearby points should have similar gradients. In particular, if x is close to the optimal x^*, since $\nabla f(x^*) = 0$, we know that $\nabla f(x)$ must be small (close to zero).

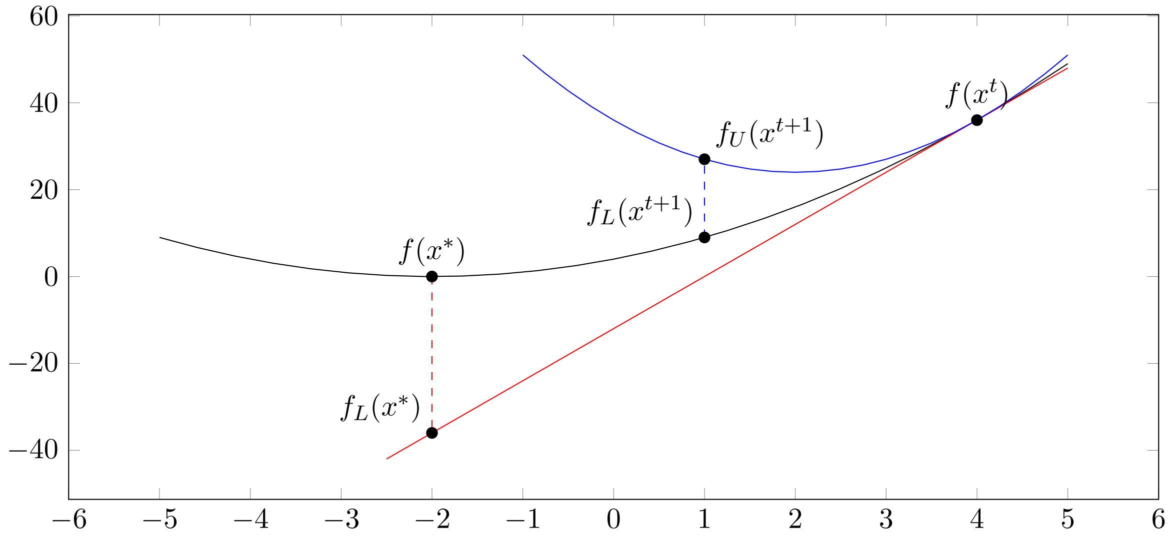

Figure 3: Illustration of a convex β-smooth function $f(x)$, with its lower bound $f^L(x) = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle$ and upper bound $f^U(x) = f(x_0) + \langle \nabla f(x_0), (x - x_0) \rangle + 0.5\beta \|x - x_0\|^2$. The lower bound makes sure $f(x^t)$ is not too far away from $f(x^*)$, while the upper bound makes sure some progress $f(x^t) - f(x^{t+1})$ are made in each iteration.

Definition 1 (β-smooth functions)**.** *A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is β-smooth if its gradient ∇f is β-Lipschitz, that is*

$$\|\nabla f(x) - \nabla f(y)\| \le \beta \|x - y\|, \quad \forall x, y \in \mathbb{R}^d \tag{12}$$

Problem 2. *Given a convex and β-smooth function $f : \mathbb{R}^n \to \mathbb{R}$. Find a minimizer of f.*

2.1 Sandwiching Smooth Convex Functions

In the case of convex Lipschitz function, we use the property of subgradient to lower bound f. $\forall x, y \in \mathbb{R}^n$ and $\forall g \in \partial f(x)$

$$f(y) \ge f(x) + g^\top (y - x)$$

this leads to (4). And in the proof, we use this to lower bound $f(x^*)$, making sure that $f(x^t) - f(x^*)$ is controlled, i.e. we are not too far away from the optimal.

In the scenario of β-smooth function, we can get the same lower bound, because $\nabla f(x)$ is always the unique subgradient at any point x. Moreover, the β-smoothness gives us an upper bound:

$$f(y) \le f(x) + \nabla f(x)^\top (y - x) + \frac{\beta}{2} \|y - x\|^2$$

and this could be used to get a lower bound on the decrement $f(x^t) - f(x^{t+1})$ at each iteration. Sef Fig. 3 for an illustration. We state the conclusions below formally.

Lemma 2. *Assume $f : \mathbb{R}^n \to \mathbb{R}$ is β-smooth, then $\forall x, y \in \mathbb{R}^n$*

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \le \frac{\beta}{2} \|y - x\|^2$$

Proof. By the fundamental theorem for line integrals,

$$f(y) = f(x) + \int_0^1 \langle \nabla f(x + t(y-x)), y - x\rangle dt$$

Plugin $f(y) - f(x)$,

$$\begin{aligned} |f(y) - f(x) - \langle \nabla f(x), y - x\rangle| \leq \frac{\beta}{2}\|y-x\|^2 &= \left| \int_0^1 \langle \nabla f(x + t(y-x)) - \nabla f(x), y - x\rangle dt \right| \\ &\leq \|y - x\| \int_0^1 \|\nabla f(x + t(y-x)) - \nabla f(x)\| dt \\ &\leq \|y - x\| \int_0^1 \beta t \|y - x\| dt \\ &= \frac{\beta}{2}\|y - x\|^2 \end{aligned}$$

□

If we further know that f is convex, combining the lower bound from the first order condition of convexity, we get both lower bound and upper bound

$$0 \leq f(y) - f(x) - \langle \nabla f(x), y - x\rangle \leq \frac{\beta}{2}\|y-x\|^2 \tag{13}$$

See again Fig. 3 for an illustration. Actually, combining convexity and β-smoothness, the lower bound in (13) could be improved. Consider the extreme case when $f(x)$ is a linear function, then the lower bound is actually tight. In this case, we also have $\nabla f(x) = \nabla f(y)$. However, if $f(x)$ is not linear, $\nabla f(x) \neq \nabla f(y)$, we might observe a non-zero gap between $f(x)$ and its linear lower bound. It is also intuitive that the gap might be larger when the gradient $\nabla f(y)$ changed a lot from $\nabla f(x)$, so we are thinking about getting a better lower bound using the quantity $\|\nabla f(x) - \nabla f(y)\|$.

Lemma 3. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex and β-smooth function, then $\forall x, y \in \mathbb{R}^n$*

$$\frac{1}{2\beta}\|\nabla f(x) - \nabla f(y)\|^2 \leq f(y) - f(x) - \langle \nabla f(x), y - x\rangle \leq \frac{\beta}{2}\|y-x\|^2 \tag{14}$$

Proof. In order to invite both $\nabla f(x)$ and $\nabla f(y)$ into play, we consider a third point $z \in \mathbb{R}^n$, and approximate $f(z)$ from below by $\nabla f(y)$ and from above by $\nabla f(x)$, respectively. Using (13)

$$\begin{aligned} f(z) - f(x) - \langle \nabla f(x), z - x\rangle &\geq 0 \\ f(z) - f(y) - \langle \nabla f(y), z - y\rangle &\leq \frac{\beta}{2}\|z - y\|^2 \end{aligned}$$

Multiply the first inequality by -1 and sum the two inequalities, we get

$$f(x) - f(y) + \langle \nabla f(x), z - x\rangle - \langle \nabla f(y), z - y\rangle \leq \frac{\beta}{2}\|z - y\|^2$$

Re-write the inequality by moving the quantity we want to lower bound to the right,

$$\langle \nabla f(x), z - y\rangle - \langle \nabla f(y), z - y\rangle - \frac{\beta}{2}\|z - y\|^2 \leq f(y) - f(x) - \langle \nabla f(x), y - x\rangle$$

Inspecting the left hand side, if we let $z = y + \alpha(\nabla f(y) - \nabla f(x))$ for any $\alpha \in \mathbb{R}$, we get

$$\left(\alpha - \frac{\alpha^2 \beta}{2}\right) \|\nabla f(y) - \nabla f(x)\|^2 \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle$$

Since the lower bound is a quadratic function in α, we can maximize the lower bound by taking $\alpha = 1/\beta$. And the conclusion follows. $\square$