Intuitions for Optimization

pluskid

October 1, 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 $\mathcal{K}$:

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

Find the 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 $t \leftarrow 0, \dots, T-1$ **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*

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

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

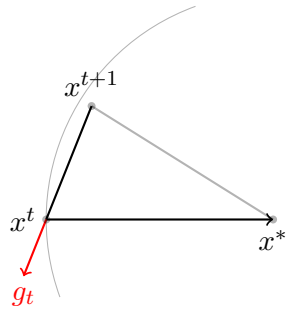

Figure 1: Demonstration of subgradient descent.

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,

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

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 $t = 0, \ldots, T-1$,

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

It then implies

$$\min_{0 \leq \tau \leq T} f(x^\tau) - f(x^*) \leq \frac{R^2 + G^2 \sum_{t=0}^{T-1} \eta_t^2}{\sum_{t=0}^{T-1} \eta_t} \tag{5}$$

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

$$\frac{R^2 + G^2 T \eta^2}{T\eta} = \frac{R^2}{T\eta} + G^2 \eta \geq \frac{2GR}{\sqrt{T}} \tag{6}$$

where the inequality holds with equality when

$$\eta = \frac{R}{G\sqrt{T}} \tag{7}$$

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

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

Plug-in to (3), we get

$$\min_{0\leq\tau\leq T} f(x^\tau) - f(x^*) \leq GR\frac{1+\log T}{2\sqrt{T+1}-2} \lesssim \frac{GR\log T}{\sqrt{T}} \tag{8}$$

Comparing with the optimal bound we get with a fixed stepsize in (6), 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