Intuitions for Convex Optimization

pluskid

December 4, 2015

Contents

Preface

Originally I wanted to write down some notes of the intuitions I developed when trying to understand various optimization algorithms. Gradually I realized it becomes very difficult to organize the contents in a coherent way. Since Bubeck (2015) is an excellent book (and the main reference here) that already includes most of the topics that I wanted to cover in this notes. I decided to follow the organization in that book and fill in my thoughts and notes accordingly. The original book is available at Sébastien Bubeck's homepage.

1 Introduction

2 Convex optimization in finite dimension

3 Dimension-free convex optimization

In this chapter, we study the basic gradient descent algorithm and its variants. This algorithm only requires a first-order oracle that evaluates the gradient (or subgradient) at a given point. The convergence rate can be characterized in the number of oracle calls, which is remarkably independent of the input dimension[1]. In many cases, evaluating the first order oracle involves only the same computational complexity as evaluating the function itself (e.g. the loss function for learning a structured SVM).

Remark (Lemma 3.1)**.** *This lemma will be useful when we consider constrained optimization where a projection is usually used in each iteration to make sure the parameters remain feasible. In particular, this lemma shows that projection onto a convex set $\mathcal{X}$ is a* contraction*:*

$$\|\Pi_{\mathcal{X}}(x) - \Pi_{\mathcal{X}}(x')\| \le \|x - x'\| \tag{1}$$

The inequality can be easily shown by applying Lemma 3.1:

$$\begin{aligned}
\|x - x'\|^2 &= \|x - \Pi_{\mathcal{X}}(x) + \Pi_{\mathcal{X}}(x) - \Pi_{\mathcal{X}}(x') + \Pi_{\mathcal{X}}(x') - x'\|^2 \\
&= \|\Pi_{\mathcal{X}}(x) - \Pi_{\mathcal{X}}(x')\|^2 + \|x - \Pi_{\mathcal{X}}(x)\|^2 + \|x' - \Pi_{\mathcal{X}}(x')\|^2 \\
&\quad + 2\langle x - \Pi_{\mathcal{X}}(x), \Pi_{\mathcal{X}}(x) - \Pi_{\mathcal{X}}(x')\rangle + 2\langle \Pi_{\mathcal{X}}(x) - \Pi_{\mathcal{X}}(x'), \Pi_{\mathcal{X}}(x') - x'\rangle \\
&\quad + 2\langle x - \Pi_{\mathcal{X}}(x), \Pi_{\mathcal{X}}(x') - x'\rangle \\
&\ge \|\Pi_{\mathcal{X}}(x) - \Pi_{\mathcal{X}}(x')\|^2 + \|x - \Pi_{\mathcal{X}}(x)\|^2 + \|x' - \Pi_{\mathcal{X}}(x')\|^2 \\
&\quad + 2\langle x - \Pi_{\mathcal{X}}(x), \Pi_{\mathcal{X}}(x') - x'\rangle \\
&= \|\Pi_{\mathcal{X}}(x) - \Pi_{\mathcal{X}}(x')\|^2 + \|(x - \Pi_{\mathcal{X}}(x)) + (\Pi_{\mathcal{X}}(x') - x')\|^2 \\
&\ge \|\Pi_{\mathcal{X}}(x) - \Pi_{\mathcal{X}}(x')\|^2
\end{aligned}$$

3.1 Projected subgradient descent for Lipschitz functions

Remark (Theorem 3.2)**.** *The intuition behind the proof is that by the property of subgradient,*

$$0 \le f(x_s) - f(x^*) \le g_s^\top (x_s - x^*) = -g_s^\top (x^* - x_s) \tag{2}$$

[1] Although as also noted in the book, the computational complexity remains at least linear in the dimension since we need to manipulate the gradient vector. Moreover, the constants (Lipschitz constant or radius of the feasible parameters) usually depend on the dimension when we consider specific problems in, machine learning, for example.

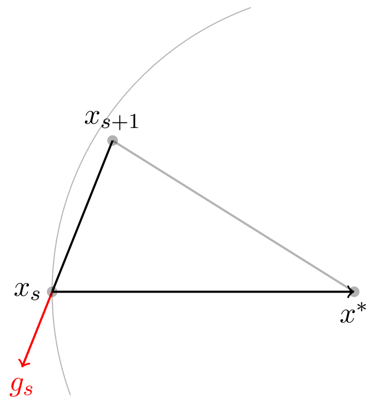

Figure 1: Demonstration of subgradient descent.

where the left hand side is due to the optimality of x^*. This inequality indicates that the vector $x^* - x_s$ is non-negatively correlated with $-g_s$, 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 $-g_s$ and $x^* - x_s$ is greater than or equal to $\pi/2$, we will move away from x^* with any positive step size. However, if $f(x_s) - f(x^*) > 0$, 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,

$$\begin{aligned} \|x_{s+1} - x^*\|^2 &\leq \|y_{s+1} - x^*\|^2 = \|y_{s+1} - x_s + x_s - x^*\|^2 \\ &= \eta_s^2\|g_s\|^2 + \|x_s - x^*\|^2 - 2\eta_s g_s^\top (x_s - x^*) \\ &\leq \eta_s^2\|g_s\|^2 + \|x_s - x^*\|^2 - 2\eta_s \left(f(x_s) - f(x^*)\right) \end{aligned} \tag{3}$$

Since $\|g_s\| \leq L$ by our assumption, the red term decays quadratically as $\eta_t \to 0$, while the blue term only decays linearly. So when η_s is small enough, we will have $\|x_{s+1} - x^*\|^2 \leq \|x_s - x^*\|^2$. Furthermore, the progress we make by moving towards x^* is characterized by $f(x_s) - f(x^*)$. 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 $f(x_s) - f(x^*)$ 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 $s = 0, \ldots, t-1$,

$$\begin{aligned} 0 \leq \|x_t - x^*\|^2 &\leq \|x_0 - x^*\|^2 + \sum_{s=0}^{t-1} \eta_s^2\|g_s\|^2 - 2\sum_{s=0}^{t-1} \eta_s \left(f(x_s) - f(x^*)\right) \\ &\leq R^2 + L^2 \sum_{s=0}^{t-1} \eta_s^2 - 2\left(\min_{0\leq s\leq t} f(x_s) - f(x^*)\right) \sum_{s=0}^{t-1} \eta_s \end{aligned}$$

It then implies

$$\min_{0\leq s\leq t} f(x_s) - f(x^*) \leq \frac{R^2 + L^2 \sum_{s=0}^{t-1} \eta_s^2}{2\sum_{s=0}^{t-1} \eta_s} \tag{4}$$

Theorem assumptions

This theorem deals with almost the most general case: the function is assumed to be convex but not necessarily differentiable. Several minimum assumptions are still needed for both the algorithm and convergence rate.

Firstly, it is assumed that $\mathcal{X} \subset \mathcal{B}(0; R)$. This quantity is mostly to characterize how far the initial guess is from the optimal solution. The step size will also depend on (proportional to) the radius R: intuitively, if R is too large, using relatively too small step sizes will take a long time to converge.

The second assumption is the L-Lipschitzness of f.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 $x_s = \varepsilon$, $\varepsilon > 0$. f is differentiable at ε, so the only subgradient is $g_s = C$. Therefore,

$$x_{s+1} = x_s - \eta_s g_s = \varepsilon - \eta_s C < -\varepsilon, \quad \forall \eta_s > \frac{2\varepsilon}{C}$$

As we can see, unless the stepsize η_s is very tiny, we will overshoot, $f(x_{s+1}) > f(x_s)$. Moreover, if we use a constant stepsize $\eta_s = \eta$. Then if $\eta > \varepsilon/C$, we will be jumping back and forth at ε and $\varepsilon - \eta C$, not able to make any further progress.

In order to fix this, we need to make additional assumptions. 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, so we will just assume f is L-Lipschitz.

Note f being L-Lipschitz in $\mathcal{X}$ implies that $\|g\| \le L$, $\forall g \in \partial f(x), \forall x \in \mathcal{X}$. Knowing an upper bound of $\|g\|$, we could then choose a small enough (according to L) step size to avoid overshooting. This choice guarantees convergence, but tiny step size is also the source of the slow convergence of subgradient descent algorithms.

Both R and L can typically be estimated for real world problems. Estimating does not need to be exact, loose upper bounds will still have the convergence guarantee, but the convergence will be slower.

Choosing the step size

The text after the theorem mentioned that the step size depending on the total number of iterations t is a bit strange — if I want to run the algorithm for more iterations, I will have to start over from scratch with a different step size. Using non-constant decaying step sizes independent of t is actually possible.

We first describe how the “magic” step size in Theorem 3.2 $\eta = R/(L\sqrt{t})$ is chosen, and then generalize it to decaying step sizes. Note (4) holds for any choices of step sizes η_s (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 $\eta_s = \eta$ for $s = 0, \dots, t-1$. So the right hand side becomes

$$\frac{R^2 + L^2 t \eta^2}{2t\eta} = \frac{R^2}{2t\eta} + \frac{L^2\eta}{2} \ge \frac{LR}{\sqrt{t}} \tag{5}$$

where the inequality holds with equality when

$$\eta = \frac{R}{L\sqrt{t}} \tag{6}$$

The dependency of η on R and L is described in the remarks about Theorem assumptions. The dependency on t is also intuitive: if we have a larger time budget, we can be a little bit more careful and move slowly.

In general, we can also use a decaying learning rate. Specifically, as long as $\sum_s \eta_s \to \infty$ and $\sum_s \eta_s^2$ is bounded or approaches infinity at a slower rate than $\sum_s \eta_s$, (4) will give a reasonable bound. For example, take $\eta_s = R/(L\sqrt{s+1})$, since

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

Plug-in to (4), we get

$$\min_{0\leq s\leq t} f(x^s) - f(x^*) \leq LR\frac{1+\log t}{4\sqrt{t+1}-4} \lesssim \frac{LR\log t}{\sqrt{t}} \tag{7}$$

Comparing with the optimal bound we get with a fixed step size Theorem 3.2, we lose a factor of $\log t$, but our step size does not depend on the total number of iterations any more.

3.2 Gradient descent for smooth functions

3.3 Conditional gradient descent, aka Frank-Wolfe

3.4 Strong convexity

3.5 Lower bounds

3.6 Geometric descent

3.7 Nesterov's accelerated gradient descent

4 Almost dimension-free convex optimization in non-Euclidean spaces

5 Beyond the black-box model

6 Convex optimization and randomness

7 Projected Subgradient Descent for Convex Lipschitz Functions

7.1 Problem Setup

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

$$\mathcal{X} \subset \mathcal{B}_2(0; R) \tag{8}$$

and f is convex and G-Lipschitz on $\mathcal{X}$:

$$|f(x) - f(y)| \leq G\|x - y\|_2, \quad x, y \in \mathcal{X} \tag{9}$$

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

$$\underset{x \in \mathcal{X}}{\text{minimize}}\, f(x)$$

7.2 Algorithm and its Bounds

Algorithm 1 Projected Subgradient Descent

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

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

$$\min_{0 \le \tau \le T} f(x^\tau) - f(x^*) \le \frac{R^2 + G^2 \sum_{t=0}^{T-1} \eta_t^2}{2 \sum_{t=0}^{T-1} \eta_t} \tag{10}$$

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.

7.3 Intuitions and Analysis

7.3.1 Problem Assumptions

We consider the unconstrained case first, i.e. $\mathcal{X} = \mathbb{R}^n$. 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 G-Lipschitz.

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

7.3.2 Convergence Analysis

7.4 Choosing Step Sizes

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

$$(x - \Pi_{\mathcal{X}}(y))^\top (y - \Pi_{\mathcal{X}}(y)) \leq 0 \tag{11}$$

which also implies

$$\|x - \Pi_{\mathcal{X}}(y)\|^2 + \|y - \Pi_{\mathcal{X}}(y)\|^2 \leq \|y - x\|^2 \tag{12}$$

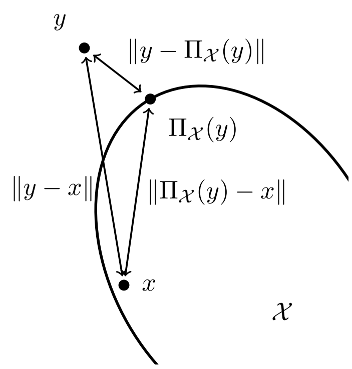

Figure 2: Illustration of convex projection. Figure source: Sébastien Bubeck, *Theory of Convex Optimization for Machine Learning*.

This lemma could be proved with *supporting hyperplane theorem* of convex sets. (12) implies

$$\|x - \Pi_{\mathcal{X}}\|^2 \leq \|y - x\|^2$$

So if we go back to our analysis in Section 7.3.2, the only thing we need to modify is (3). Since $x^* \in \mathcal{X}$,

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

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

8 Gradient Descent for Convex Smooth Functions

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).

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*

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\|, \quad \forall x, y \in \mathbb{R}^d \tag{13}$$

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

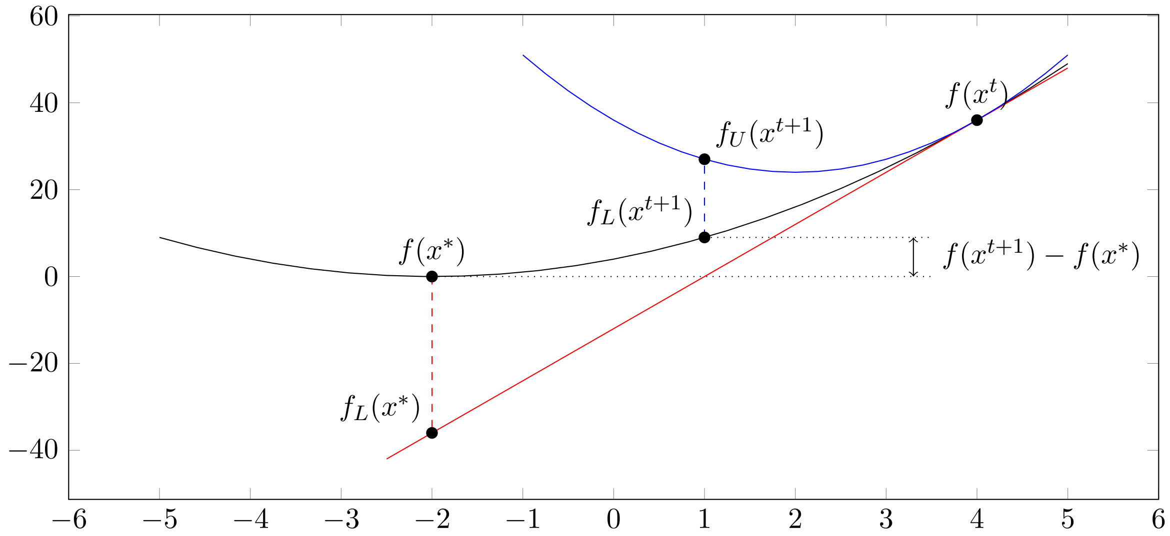

Figure 3: Illustration of a convex β-smooth function $f(x)$, with its lower bound $f^L(x) = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle$ and upper bound $f^U(x) = f(x_0) + \langle \nabla f(x_0), (x - x_0) \rangle + 0.5\beta \|x - x_0\|^2$. The lower bound makes sure $f(x^t)$ is not too far away from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

Theorem 2. *Solving Problem 2 using gradient descent with step size $\eta_t = 1/\beta$ for T iterations gives*

$$f(x^T) - f(x^*) \leq \frac{2\beta \|x_1 - x^*\|^2}{T+3} \tag{14}$$

That means to get an approximation error of ε, we will need to run the algorithm for $O(1/\varepsilon)$ iterations.

Note we get much faster convergence rate than Problem 1 by making more assumptions on f. The parameter β control the smoothness of f: smaller β means the gradient of f changes more slowly, and $\eta_t = 1/\beta$ means we could make aggressive movement in each iteration.

8.1 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 $\forall g \in \partial f(x)$

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

this leads to (2). 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:

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

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 $\forall x, y \in \mathbb{R}^n$*

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

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

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

□

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

$$0 \le f(y) - f(x) - \langle \nabla f(x), y - x \rangle \le \frac{\beta}{2}\|y - x\|^2 \tag{15}$$

See again Fig. 3 for an illustration.

8.2 Convergence Analysis and Tighter Sandwiching

To analyze the convergence, let $\Delta_t = f(x^t) - f(x^*)$ be the suboptimality gap at x^t. Let $y = x^*$, and $x = x^t$ in (15), and using the lower bound, we can upper bound Δ_t by

$$\Delta_t = f(x^t) - f(x^*) \le -\langle \nabla f(x^t), x^* - x^t \rangle \le \|\nabla f(x^t)\| \|x^* - x^t\| \le R\|\nabla f(x^t)\| \tag{16}$$

where we define

$$R = \max_{1 \le t \le T} \|x^* - x^t\| \tag{17}$$

On the other hand, using the upper bound in (15) by letting $y = x^{t+1}$ and $x = x^t$, we could lower bound $\Delta_t - \Delta_{t+1}$ by

$$\begin{aligned}
\Delta_t - \Delta_{t+1} = f(x^t) - f(x^{t+1}) &\ge -\langle \nabla f(x^t), x^{t+1} - x^t \rangle - \frac{\beta}{2}\|x^{t+1} - x^t\|^2 \\
&= \left(\eta_t - \frac{\beta \eta_t^2}{2} \right) \|\nabla f(x^t)\|^2
\end{aligned}$$

Naturally, we want to maximize the lower bound, so the step size is chosen to be $\eta_t = 1/\beta$. In this case,

$$\Delta_t - \Delta_{t+1} \ge \frac{1}{2\beta}\|\nabla f(x^t)\|^2 \tag{18}$$

Note (16) and (18) are linked together by $\|\nabla f(x^t)\|$: if $\|\nabla f(x^t)\|$ is large, then moving one step makes a lot of progress in $\Delta_t - \Delta_{t+1}$; on the other hand, if $\|\nabla f(x^t)\|$ is small, Δ_t is also small, meaning we are close to the optimal. Combining the two equations, we get

$$\Delta_t - \Delta_{t+1} \ge \frac{1}{2\beta}\|\nabla f(x^t)\|^2 \ge \frac{1}{2\beta R^2}\Delta_t^2 \tag{19}$$

Note the right hand side is non-negative, so $\Delta_t \geq \Delta_{t+1}$. To solve this recursion, divide both side by $\Delta_t \Delta_{t+1}$:

$$\frac{1}{\Delta_{t+1}} - \frac{1}{\Delta_t} \geq \frac{1}{2\beta R^2} \frac{\Delta_t}{\Delta_{t+1}} \geq \frac{1}{2\beta R^2} \tag{20}$$

Sum the recursion for $t = 2, \ldots, T$, we get

$$\frac{1}{\Delta_T} \geq \frac{T-1}{2\beta R^2} + \frac{1}{\Delta_1} \geq \frac{T+3}{2\beta R^2}$$

where the last inequality is because Δ_1 can be controlled by the upper bound in (15). Let $x = x^*$ and $y = x^1$, notice $\nabla f(x^*) = 0$,

$$\Delta_1 = f(x^1) - f(x^*) \leq \frac{\beta}{2}\|x^1 - x^*\|^2 \leq \frac{\beta R^2}{2}$$

At this point, we almost proved Theorem 2, except that we have to control R. In the following, we will show that $\|x^t - x^*\|$ is actually decreasing at each iteration, and bound R by $\|x^1 - x^*\|$. Recall in the case of subgradient descent, we use the linear lower bound of the function by the subgradient to construct the inequality in (3). That inequality shows that when η_t is small enough, because the quadratic term decays faster, we get $\|x^{t+1} - x^*\|^2 \leq \|x^t - x^*\|^2$. However, in the case here, since $\eta_t = 1/\beta$, especially when β is small, we could be moving with very large step size. So the argument is no longer useful here.

To properly bound R, we will need to get a better lower bound of f than in (15). Actually, combining convexity and β-smoothness, the lower bound in (15) 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 $\forall x, y \in \mathbb{R}^n$*

$$\frac{1}{2\beta}\|\nabla f(x) - \nabla f(y)\|^2 \leq f(y) - f(x) - \langle \nabla f(x), y - x\rangle \leq \frac{\beta}{2}\|y - x\|^2 \tag{21}$$

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 (15)

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

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. $\square$

With the improved lower bound of f in (21), we can now bound R by showing that

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

Therefore, $R \leq \|x^1 - x^*\|$, which conclude the proof of Theorem 2. Actually, in the proof above, we only need a property of the gradient called *co-coercivity*.

Definition 2 (Co-coercive mapping). *A mapping $F : \mathbb{R}^n \to \mathbb{R}^n$ is co-coercive with parameter C if $\forall x, y \in \mathbb{R}^d$,*

$$\langle F(x) - F(y), x - y \rangle \geq C\|F(x) - F(y)\|^2 \tag{22}$$

By switching the role of x and y in the lower bound of f in (21), it is easy to show the following property on the co-coercivity of $\nabla f(x)$.

Lemma 4 (Co-coercivity of gradient of convex smooth functions). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex β-smooth function, then $\nabla f(x)$ is co-coercive mapping with parameter $1/\beta$. That is*

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{1}{\beta}\|\nabla f(x) - \nabla f(y)\|^2 \tag{23}$$

We will re-visit some variants of this property in our later adventures of convergence analysis.

9 Projected Gradient Descent for Convex Smooth Functions with Constraints

For Problem 1, generalization from unconstrained case to constrained case is straightforward, because we only rely on deriving an upper bound of $\|(x^t - \eta_t g_t) - x^*\|$, as in (3). By the contraction property of convex projection, as stated in Lemma 1, we naturally get

$$\|\Pi_{\mathcal{X}}(x^t - \eta_t g_t) - x^*\| \leq \|(x^t - \eta_t g_t) - x^*\|$$

and everything follows directly. In the smooth case, things become a bit more complicated. Especially in the argument that we are making a lot of progress in one step: the contraction might reduce the aggressive movement we made, and the effective progress might get discounted. So we state this as a separate problem.

Problem 3. *Give a closed convex set $\mathcal{X}$, and a convex, β-smooth function $f : \mathcal{X} \to \mathbb{R}$. Find a minimizer of f on $\mathcal{X}$.*

9.1 Intuition and Convergence Analysis

In the unconstrained case, we use bound $\Delta_{t+1} = f(x^{t+1}) - f(x^*)$ (see Fig. 3) by roughly

$$\begin{aligned} f(x^{t+1}) - f(x^*) &= f(x^t) - f(x^*) - (f(x^t) - f(x^{t+1})) \\ &\le \langle \nabla f(x^t), x^t - x^* \rangle - \frac{1}{2\beta}\|\nabla f(x^t)\|^2 \end{aligned}$$

where the upper bound of the red term is provided in (16) and the lower bound of the blue term in (18). And the quantity here is characterized by $\|\nabla f(x^t)\|$. In the constrained case, we will do a similar thing, but with a different charactering quantity.

Remember the bounds for both the red term and blue term are directly from (15). In the constrained case,

$$x^{t+1} = \Pi_{\mathcal{X}}(x^t - \eta_t \nabla f(x^t)) \tag{24}$$

By the property of convex projection in Lemma 1, for any $y \in \mathcal{X}$,

$$\langle x^{t+1} - (x^t - \eta_t \nabla f(x^t)), x^{t+1} - y \rangle \le 0 \quad \Leftrightarrow \quad \langle \nabla f(x^t), x^{t+1} - y \rangle \le \frac{1}{\eta_t}\langle x^t - x^{t+1}, x^{t+1} - y \rangle$$

With this tool, we carry the out the analysis as before, by applying (15),

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

In particular, if we still choose the step size $\eta_t = 1/\beta$ as in the unconstrained case, we get

$$\begin{aligned} f(x^{t+1}) - f(x^*) &\le \langle \beta(x^t - x^{t+1}), x^t - x^* \rangle - \frac{\beta}{2}\|x^{t+1} - x^t\|^2 \\ &= \langle g_{\mathcal{X}}(x^t), x^t - x^* \rangle - \frac{1}{2\beta}\|g_{\mathcal{X}}(x^t)\|^2 \end{aligned} \tag{25}$$

where we have defined the quantity

$$g_{\mathcal{X}}(x^t) = \beta(x^t - x^{t+1}) \tag{26}$$

Note in the unconstrained case, $\beta(x^t - x^{t+1}) = \nabla f(x^t)$, and its norm characterized our bounds. It turns out that in the constrained case, this quantity plays the same role. Actually, the derivation above does not use the optimality of x^*, so replacing x^* with any $y \in \mathcal{X}$, we get the following lemma.

Lemma 5 (Three-point Inequality). *Let* $x, y \in \mathcal{X}$ *and* $x^+ = \Pi_{\mathcal{X}}\left(x - \frac{1}{\beta}\nabla f(x)\right)$*, and* $g_{\mathcal{X}}(x) = \beta(x - x^+)$*. Then*

$$f(x^+) - f(y) \leq \langle g_{\mathcal{X}}(x), x - y \rangle - \frac{1}{2\beta}\|g_{\mathcal{X}}(x)\|^2 \tag{27}$$

By plugging in $x = x^t$, $y = x^*$, we immediately get the upper bound

$$\Delta_{t+1} = f(x^{t+1}) - f(x^*) \leq \langle g_{\mathcal{X}}(x^t), x^t - x^* \rangle \leq \|g_{\mathcal{X}}(x^t)\|\|x^t - x^*\| \tag{28}$$

Similarly, plugging in $x = y = x^t$, we get the lower bound

$$\Delta_t - \Delta_{t+1} = f(x^t) - f(x^{t+1}) \geq \frac{1}{2\beta}\|g_{\mathcal{X}}(x^t)\|^2 \tag{29}$$

The two inequalities are now linked by $\|g_{\mathcal{X}}(x^t)\|$, combing them, we get

$$\Delta_t - \Delta_{t+1} \geq \frac{1}{2\beta}\frac{\Delta_{t+1}^2}{\|x^t - x^*\|^2} \geq \frac{\Delta_{t+1}^2}{2\beta R^2}$$

where again we let $R = \max_{1 \leq t \leq T} \|x^t - x^*\|$.

References

Bubeck, S. (2015). Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning 8*(3-4), 231–357.