Nesterov's Acceleration Method

1 Convex Optimization Problem

In this lecture, we consider the general unconstrained convex optimization problem:

$$\min_{x\in\mathbb{R}^d} f(x).$$

In the last lecture, we have shown that for quadratic optimization problems, it is possible to improve the gradient descent method by using conjugate gradient method or the heavy-ball method.

Both methods employ the following form of recursion:

$$\begin{aligned} x_t =& x_{t-1} + p_{t-1} \\ p_t =& -\alpha_t \nabla f(x_t) + \beta_t p_{t-1}. \end{aligned}$$

We may refer to this class of methods as momentum methods, and it can be employed for general unconstrained optimization problems.

In CG, we set α_t and β_t automatically for quadratic functions. There are also generalizations to general functions. However, the automatic formula are difficult to generalize to stochastic methods. Alternatively, one can set α_t and $\bar{\beta}_t$ manually, leading to the heavy ball method. We have shown, it is possible to improve the convergence rate of the first order algorithms from $O(\kappa \log(1/\epsilon))$ to $O(\sqrt{\kappa}\log(1/\epsilon))$.

In this lecture, we consider a modification of the momentum methods by Nesterov, and show that the convergence rate of these methods can be proved for convex problems.

2 Nesterov's Acceleration

The momentum method can be written as the following form:

$$\begin{aligned} y_t &= x_{t-1} + \beta_t(x_{t-1} - x_{t-2}) \\ x_t &= y_t - \alpha_t \nabla f(x_{t-1}). \end{aligned}$$

Nesterov modified this equation as follows:

$$\begin{aligned} y_t &= x_{t-1} + \beta_t(x_{t-1} - x_{t-2}) \\ x_t &= y_t - \alpha_t \nabla f(y_t). \end{aligned}$$

With this modification, one can prove the global convergence of the resulting algorithm, which is listed in Algorithm 1.

Algorithm 1: Nesterov's Acceleration Method

Input: $f(x)$, x_0, $\alpha_1, \beta_1, \alpha_2, \beta_2, \ldots$

Output: x_T

Let $x_{-1} = x_0$

for $t = 1, \ldots, T$ **do**

  Let $y_t = x_{t-1} + \beta_t(x_{t-1} - x_{t-2})$

  Let $x_t = y_t - \alpha_t \nabla f(y_t)$

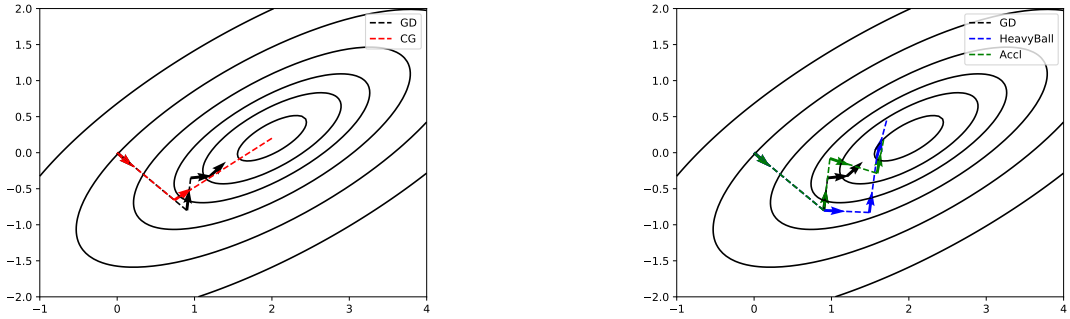
Return: x_T

Note that equivalently, we may write Nesterov's method as follows, with a different choice of parameters.

$$\begin{aligned} y_t &= x_{t-1} + \beta_t p_{t-1} \\ p_t &= \beta_t p_{t-1} - \alpha_t \nabla f(y_t) \\ x_t &= x_{t-1} + p_t. \end{aligned}$$

3 Motivations

We illustrates different methods in Figure 1. For gradient Descent, we can see that it has a zigzagged solution path. It tends to oscillate, and thus not efficient in terms of convergence. Conjugate gradient method, on the other hand, going through solution paths that are conjugate (A-orthogonal) directions. Therefore it goes more directly to the optimal solution. The solution paths for the accelerated methods are also smoother, which makes them more efficient in terms of convergence.



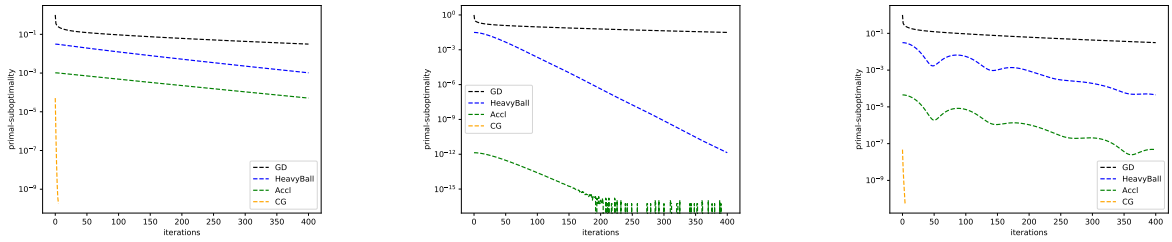
(a) Gradient Descent and CG

(b) Gradient Descent, Heavy Ball, and Acceleration

Figure 1: Gradient Descent versus Other Methods

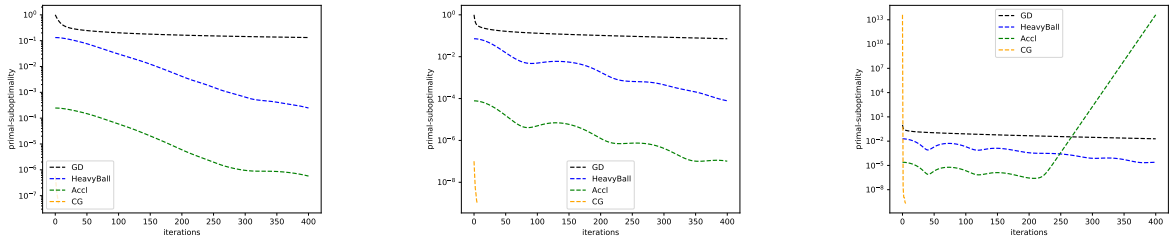
Next, we try to study the sensitivity of α and β parameters, using a quadratic optimization problem with condition number $\kappa \approx 1000$. Figure 2 compares the convergence of different methods, with the learning rate is set as $\alpha = 1/L$. We vary β and examine its effect. In Figure 2a, $\beta = 1 - 10/\sqrt{\kappa} \approx 0.7$. In Figure 2b, $\beta = 1 - 2/\sqrt{\kappa} \approx 0.94$. In Figure 2c, $\beta = 1 - 0.5/\sqrt{\kappa} \approx 0.98$. Figure 3 compares the convergence of different methods with $\beta = 1 - 0.5/\sqrt{\kappa} \approx 0.98$. We vary α and examine its effect. In Figure 3a, $\alpha = 0.1/L$. In Figure 3b, $\alpha = 0.5/L$. In Figure 3c, $\alpha = 1.5/L$.

We can see that parameter tuning is important. For Nesterov's method, the best result can be achieved with an appropriate setting of α and β, as in Figure 2b. Large β causes oscillation. It is thus safer to choose a not so large β. Small α alleviates oscillation, but slows down convergence. Including β makes the algorithm less sensitive to the choice of learning rate α.



(a) $\beta \approx 0.7$ (b) $\beta \approx 0.94$ (c) $\beta \approx 0.98$

Figure 2: Convergence Comparisons with Fixed α



(a) $\alpha = 0.1/L$ (b) $\alpha \approx 0.4/L$ (c) $\alpha \approx 1.4/L$

Figure 3: Convergence Comparisons with Fixed β

4 Convergence Analysis

We will prove the following result for Nesterov's acceleration method for strongly convex functions. It implies a convergence rate of

$$O(\sqrt{\kappa}\log(1/\epsilon)),$$

where $\kappa = L/\lambda$.

Theorem 1 *Assume $f(x)$ is L-smooth and λ-strongly convex. Let $\eta \leq 1/L$ and $\theta = \sqrt{\eta\lambda}$. Let $\alpha_t = \eta \leq 1/L$ and $\beta_t = \beta = (1-\theta)/(1+\theta)$. Then*

$$f(x_t) \leq f(x_*) + (1-\theta)^t \left[f(x_0) - f(x_*) + \frac{\lambda}{2}\|z - x_0\|_2^2 \right]$$

.

The proof relies on the notation of estimation sequence introduced by Nesterov.

Definition 1 *A pair of sequences* $\{(\phi_t(x), \lambda_t \geq 0\}$ *is called an estimation sequence of function* $f(x)$*, if for any* $x \in \mathbb{R}^d$ *and all* $t \geq 0$*:*

$$\phi_t(x) \leq (1 - \lambda_t) f(x) + \lambda_t \phi_0(x).$$

We will need the following two lemmas under the conditions of the theorem.

Lemma 1 *Let* $x^+ = y - \eta \nabla f(y)$*. We define*

$$\phi(z; y) = f(x^+) - \frac{1}{2\eta}\|x^+ - y\|_2^2 + \frac{1}{\eta}(y - x^+)^\top (z - x^+) + \frac{\lambda}{2}\|z - y\|_2^2.$$

Then the following inequality holds:

$$\phi(z; y) \leq f(z).$$

Therefore if we define recursively

$$\phi_t(z) = (1 - \theta)\phi_{t-1}(z) + \theta \phi(z; y_t)$$

with

$$\phi_0(z) = f(x_0) + \frac{\lambda}{2}\|z - x_0\|_2^2,$$

then $\{\phi_t, (1 - \theta)^t\}$ *is an estimation sequence.*

Proof We have

$$\begin{aligned}
f(z) \geq & f(y) + \nabla f(y)^\top (z - y) + \frac{\lambda}{2}\|z - y\|_2^2 \\
= & f(y) + \nabla f(y)^\top (x^+ - y) + \nabla f(y)^\top (z - x^+) + \frac{\lambda}{2}\|z - y\|_2^2 \\
\geq & f(x^+) - \frac{1}{2\eta}\|x^+ - y\|_2^2 + \frac{1}{\eta}(y - x^+)^\top (z - x^+) + \frac{\lambda}{2}\|z - y\|_2^2 \\
= & \phi(z; y).
\end{aligned}$$

The first inequality uses the strong convexity. The second inequality uses the smoothness with $1/\eta \geq L$, and replaces $\nabla f(y)$ by $\eta^{-1}(y - x^+)$. This proves the first bound in the theorem.

Since the following hold trivially at $t = 0$:

$$\phi_0(z) \leq (1 - (1 - \theta)^0) f(x) + (1 - \theta)^0 \phi_0(x).$$

and thus we can assume by induction that at $t - 1$: $\phi_{t-1}(x) \leq (1 - (1 - \theta)^{t-1}) f(x) + (1 - \theta)^t \phi_0(x)$. Then

$$\begin{aligned}
\phi_t(x) = & (1 - \theta)\phi_{t-1}(x) + \theta \phi(x; y_t) \\
\leq & (1 - \theta)[(1 - (1 - \theta)^{t-1}) f(x) + (1 - \theta)^{t-1} \phi_0(x)] + \theta f(x) \\
= & (1 - (1 - \theta)^t) f(x) + (1 - \theta)^t \phi_0(x).
\end{aligned}$$

This finishes the proof. ■

Lemma 2 *We have*

$$f(x_t) \leq \phi_t(v_t) = \min_z \phi_t(z).$$

Proof We have

$$\phi_t(z) = \phi_t(v_t) + \frac{\lambda}{2}\|z - v_t\|_2^2.$$

Thus

$$\phi_t(z) = (1-\theta)\phi_{t-1}(v_{t-1}) + \theta\phi(z; y_t) + (1-\theta)\frac{\lambda}{2}\|z - v_{t-1}\|_2^2.$$

We have

$$\theta\frac{1}{\eta}(y_t - x_t) + \theta\lambda(v_t - y_t) + (1-\theta)\lambda(v_t - v_{t-1}) = 0. \tag{1}$$

By induction, it can be shown from this equation that

$$v_t = \theta^{-1}(x_t + (\theta - 1)x_{t-1}). \tag{2}$$

Moreover, from (1), we obtain

$$x_t - y_t = \theta[\theta(v_t - y_t) + (1-\theta)(v_t - v_{t-1})].$$

It also implies that

$$-\|x_t - y_t\|_2^2 = -\theta^2\theta\|v_t - y_t\|_2^2 - \theta^2(1-\theta)\|v_t - v_{t-1}\|_2^2 + \theta^2 \cdot \theta(1-\theta)\|v_{t-1} - y_t\|_2^2. \tag{3}$$

It follows by induction that

$$\begin{aligned}
\phi_t(v_t) =& (1-\theta)\left[\phi_{t-1}(v_{t-1}) + \frac{\lambda}{2}\|v_t - v_{t-1}\|_2^2\right] + \theta\phi(v_t; y_t) \\
\geq& (1-\theta)\left[f(x_{t-1}) + \frac{\lambda}{2}\|v_t - v_{t-1}\|_2^2\right] + \theta\phi(v_t; y_t) \\
\geq& (1-\theta)\left[\phi(x_{t-1}; y_t) + \frac{\lambda}{2}\|v_t - v_{t-1}\|_2^2\right] + \theta\phi(v_t; y_t) \\
=& f(x_t) - \frac{1}{2\eta}\|y_t - x_t\|_2^2 + \frac{1}{\eta}(y_t - x_t)^\top((1-\theta)x_{t-1} + \theta v_t - x_t) \\
&+ (1-\theta)\frac{\lambda}{2}\|x_{t-1} - y_t\|_2^2 + \theta\frac{\lambda}{2}\|v_t - y_t\|_2^2 + (1-\theta)\frac{\lambda}{2}\|v_t - v_{t-1}\|_2^2 \\
=& f(x_t) - \frac{1}{2\eta}\|y_t - x_t\|_2^2 + (1-\theta)\frac{\lambda}{2}\|x_{t-1} - y_t\|_2^2 + \theta\frac{\lambda}{2}\|v_t - y_t\|_2^2 + (1-\theta)\frac{\lambda}{2}\|v_t - v_{t-1}\|_2^2 \\
=& f(x_t) + (1-\theta)\frac{\lambda}{2}\|x_{t-1} - y_t\|_2^2 + \theta(1-\theta)\frac{\lambda}{2}\|v_{t-1} - y_t\|_2^2.
\end{aligned}$$

This finishes the induction. In the above derivation, the first equality is definition. The first inequality uses the induction hypothesis. The second inequality uses Lemma 1. The second equality is by definition of $\phi(x; y)$. The third equality uses (2). The final equality uses (3). ■

Proof [of Theorem 1] Let $\lambda_t = (1-\theta)^t$. We have

$$f(x_t) \leq \min_{x\in\mathbb{R}^d} \phi_t(x) \leq \min_{x\in\mathbb{R}^d}[(1-\lambda_t)f(x) + \lambda_t\phi_0(x)] \leq (1-\lambda_t)f(x_*) + \lambda_t\phi_0(x_*).$$

This implies the desired result. ■