Comp6211e: Optimization for Machine Learning

Tong Zhang

Lecture 6: Nesterov's Acceleration Method

Convex Optimization

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

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

First Order Optimization

We consider the 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.

Reformulation

The momentum (Heavy Ball) 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's Method

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 for convex functions.

Algorithm

Algorithm 1: Nesterov's Acceleration Method

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

Output: x_T

1 Let $x_{-1} = x_0$

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

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

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

Return: x_T

Equivalent Formulation

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

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

This can be compared to the heavy ball formulation.

Motivation: GD versus CG

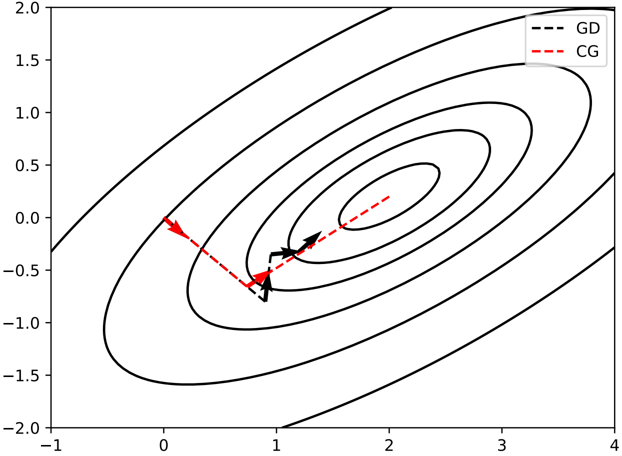


Figure: Gradient Descent and CG

Motivation: GD versus Accelerated Methods

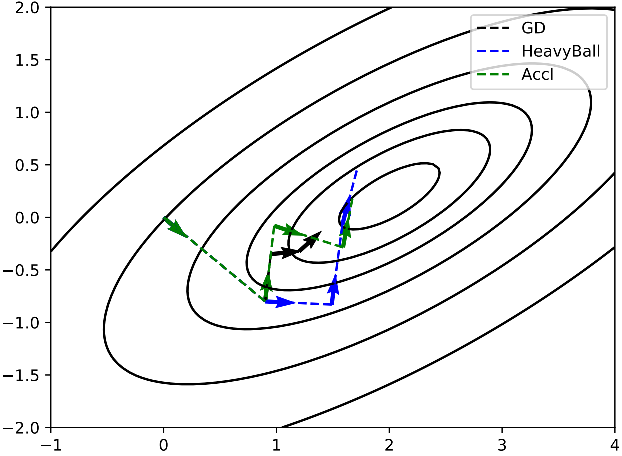
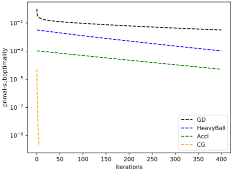
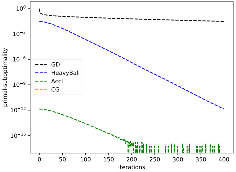


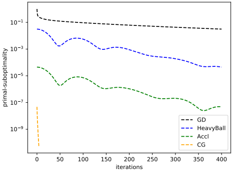
Figure: Gradient Descent, Heavy Ball, and Acceleration



(a) $\beta \approx 0.7$



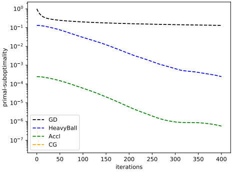
(b) $\beta \approx 0.94$



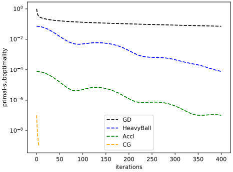
(c) $\beta \approx 0.98$

Figure: Convergence Comparisons with Fixed α

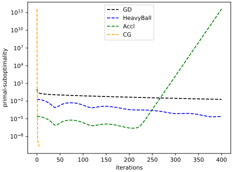
Sensitivity to α



(a) $\alpha = 0.1/L$



(b) $\alpha \approx 0.4/L$



(c) $\alpha \approx 1.4/L$

Figure: Convergence Comparisons with Fixed β

Convergence Analysis

Theorem

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

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

.

Proof: Estimation Sequence

Definition

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 $t \geq 0$:

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

Convergence Analysis with Estimation Sequence

If for an estimation sequence, we have the following property (upper bound of $f(x_t)$)

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

then

$$f(x_t) \leq (1 - \lambda_t) f(x_*) + \lambda_t \phi_0(x_*).$$

Estimation Sequence Lemma

Lemma

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 of Second part

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

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

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

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

Upper Bound

Lemma

We have

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

Summary

We have studied accelerated first order methods with momentum terms and tuning parameters α and β.

- α is like learning rate;
- β is decaying term for the aggregated gradients.

In practice, careful tuning of α and β are important.

- Right setting of β can significantly improve convergence, but inappropriate setting can lead to oscillation
- The sensitivity to α is less severe with appropriate $\beta > 0$, because the gradients are aggregated.