# Lecture 11 — Pre-conditioned Gradient Descent

*Lecturer: Yunwei Ren* *Scribed by Zhidan Li*

## Contents



## 1 Overview

In this lecture, we put our sight on two pre-conditioned versions of the gradient descent: Newton's method and adaptive gradient descent.

The second-order Taylor series expansion is often used in this lecture: for a function $f \in C^2(\mathbb{R}^d)$, for $x, y \in \mathbb{R}^d$, when $x, y$ is 'near',

$$f(y) \approx f(x) + (\nabla f(x))^\top (y - x) + \frac{1}{2}(y - x)^\top \nabla^2 f(x)(y - x). \tag{1}$$

## 2 Newton's Method

Recall the gradient descent algorithm. At each iteration, we choose locally minimize

$$g(x) = f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{2\eta}\|x - x_t\|^2.$$

If $f$ is $L$-smooth, together with the descent lemma

$$f(x_{t+1}) \le f(x_t) - \frac{\eta}{2}\|\nabla f(x_t)\|^2,$$

we can establish the convergence rate of gradient descent. However, when $f \in C^2$, such an algorithm does not truly make use of the pre-conditioned information of quadratic terms.

For a function $f \in C^2(\mathbb{R}^d)$, from (1), when $x$ is near $x_t$, it holds that

$$f(x) \approx f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{2}(x - x_t)^\top \nabla^2 f(x_t)(x - x_t) =: h(x)$$

To seek the minimizer of $h$, we take its gradient

$$\nabla h(x) = \nabla f(x_t) + \nabla^2 f(x_t)(x - x_t).$$

Let $\nabla h(x_{t+1}) = 0$, and we solve that

$$x_{t+1} = x_t - [\nabla^2 f(x_t)]^{-1} \nabla f(x_t). \tag{2}$$

Often we use the following modified version

$$x_{t+1} = x_t - \eta [\nabla^2 f(x_t)]^{-1} \nabla f(x_t) \tag{3}$$

where $\eta \in (0, 1]$ is a chosen parameter.

### 2.1 Convergence rate of Newton's method

Now we analyze the convergence rate of Newton's method (3).

**Proposition 1** (local convergence of Newton's method)**.** *Assume that $x_*$ is the strict minimizer of $f$ in the sense that $\nabla f(x_*) = 0$ and $\nabla^2 f(x_*) \succeq \rho I_d$ for some $\rho > 0$. Assume $f$ is $L$-Hessian Lipschitz (with respect to spectral norm). Then if $\|x_0 - x_*\| \leq \frac{\rho}{2L}$, with $\eta = 1$, it holds that*

$$\|x_{t+1} - x_*\| \leq \frac{2L}{\rho} \|x_t - x_*\|^2.$$

*Proof.* From the calculus fact,

$$\nabla f(x_t) - \nabla f(x_*) = \int_0^1 \nabla^2 f(x_* + s(x_t - x_*))(x_t - x_*) \, ds.$$

Plugging it into (3) with $\eta = 1$, we obtain

$$\begin{aligned}
\|x_{t+1} - x_*\| &= \left\| x_t - x_* - [\nabla^2 f(x_t)]^{-1} \nabla f(x_t) \right\| \\
&= \left\| x_t - x_* - [\nabla^2 f(x_t)]^{-1} \int_0^1 \nabla^2 f(x_* + s(x_t - x_*))(x_t - x_*) \, ds \right\| \\
&= \left\| [\nabla^2 f(x_t)]^{-1} \int_0^1 (\nabla^2 f(x_t) - \nabla^2 f(x_* + s(x_t - x_*))(x_t - x_*) \, ds \right\| \\
&\leq \left\| [\nabla^2 f(x_t)]^{-1} \right\| \cdot \|x_t - x_*\| \int_0^1 \left\| \nabla^2 f(x_t) - \nabla^2 f(x_* + s(x_t - x_*)) \right\| ds \\
&\overset{(i)}{\leq} \left\| [\nabla^2 f(x_t)]^{-1} \right\| \cdot \|x_t - x_*\|^2 \int_0^1 Ls \, ds \\
&= \frac{L}{2} \left\| [\nabla^2 f(x_t)]^{-1} \right\| \cdot \|x_t - x_*\|^2
\end{aligned}$$

where $(i)$ comes from the assumption that $f$ is $L$-Hessian Lipschitz. Then it remains to bound $\left\| [\nabla^2 f(x_t)]^{-1} \right\|$. Since the update rule is a descent process (it's not hard to verify it), then for every $x_t$, $\|x_t - x_*\| \leq \|x_0 - x_*\| \leq \frac{\rho}{2L}$.

$$\nabla^2 f(x_t) \succeq \nabla^2 f(x_*) - L\|x_t - x_*\| I_d \succeq \frac{\rho}{2} I_d.$$

This means $\left\|[\nabla^2 f(x_t)]^{-1}\right\| \leq \frac{2}{\rho}$. Hence we conclude

$$\|x_{t+1} - x_*\| \leq \frac{2L}{\rho}\|x_t - x_*\|^2.$$

□

*Remark* 2. The local convergence of Newton's method is quadratic, which is better than the gradient descent (linear system).

## 3 Adaptive Gradient Descent

The note for *Lecture 5, COS 597G: Toward Theoretical Understanding of Deep Learning, Fall 2018*, lectured by *Sanjeev Arora* is precise and clear enough for this part. I really recommend you to refer this note together with the slide.