Optimization Theory and Algorithm	**Lecture 14 - 06/11/2021**

Lecture 14

Lecturer:Xiangyu Chang *Scribe: Xiangyu Chang*

Edited by: Xiangyu Chang

1 Accelerate Gradient Descent

What is the fastest convergence speed of an optimization algorithm? We should know the lower bound

$$O(T^s) \leq f(\mathbf{x}^T) - f^* \leq O(T^s).$$

Then the optimal convergence speed is $O(T^s)$.

Theorem 1 *[Nesterov, 1998] Let $T \leq \frac{n-1}{2}$, $\beta > 0$. Then there exists a β-smooth convex quadratic f such that any black-box method stasifies*

$$\min_{1\leq t \leq T} f(\mathbf{x}^t) - f^* \geq \frac{3\beta\|\mathbf{x}^0 - \mathbf{x}^*\|^2}{32(1+T)^2}. \tag{1}$$

This means we have a chance to make an algorithm to achieve the convergence rate $O(T^{-2})$.

This is called the *accelerate (proximal) gradient descent* algorithm:

- Initial: $\mathbf{y}^1 = \mathbf{x}^0, a_1 = 1$ and $t = 1$.
- Step 1:
$$\mathbf{x}^t = \mathbf{y}^t - \frac{1}{\beta}\nabla f(\mathbf{y}^t) \text{ or } \mathbf{x}^t = prox_{g/\beta}(\mathbf{y}^t - \frac{1}{\beta}\nabla f(\mathbf{y}^t)). \tag{2}$$
- Step 2:
$$a_{t+1} = \frac{1+\sqrt{1+4a_t^2}}{2}. \tag{3}$$
- Step 3:
$$\mathbf{y}^{t+1} = \mathbf{x}^t + \frac{a_t - 1}{a_{t+1}}\underbrace{(\mathbf{x}^t - \mathbf{x}^{t-1})}_{momentum}. \tag{4}$$

Let us recall the inequality Theorem, take $\gamma = 1/\beta$ and change the position of $\mathbf{x}$ and $\mathbf{y}$. We obtain the following proposition.

Proposition 1

$$h(\mathbf{x}) \geq h(\mathbf{x}^+) + \beta\langle \mathbf{y} - \mathbf{x}^+, \mathbf{x} - \mathbf{y}\rangle + \frac{\beta}{2}\|\mathbf{y} - \mathbf{x}^+\|^2, \tag{5}$$

where $\mathbf{x}^+ = prox_{g/\beta}(\mathbf{y} - 1/\beta\nabla f(\mathbf{y})) = \arg\min_{\mathbf{x}} \left\{\frac{\beta}{2}\|\mathbf{x} - (\mathbf{y} - 1/\beta\nabla f(\mathbf{y}))\|^2 + g(\mathbf{x})\right\}$.

Lemma 1 *For any vector* $\mathbf{a}, \mathbf{b}$, *it has*

$$\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 + 2\langle \mathbf{a}, \mathbf{b}\rangle = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 + 2\|\mathbf{a}\|\|\mathbf{b}\|\cos < \mathbf{a}, \mathbf{b} >= \|\mathbf{a} + \mathbf{b}\|^2. \tag{6}$$

Lemma 2 *Let $\{c_t, b_t\}$ be positive sequences of reals satisfying $c_t - c_{t+1} \geq b_{t+1} - b_t$ for any $t \geq 1$, with $c_1 + b_1 \leq c, c > 0$, then $c_t \leq c, \forall t \geq 1$.*

Proof 1 *By induction.*

Lemma 3 *The sequences $\{\mathbf{x}^t, \mathbf{y}^t\}$ generated via FISTA with the constant step size $1/\beta$, then for every $t \geq 1$,*

$$a_t^2 v_t - a_{t+1}^2 v_{t+1} \geq \frac{\beta}{2}(\|\mathbf{u}^{t+1}\|^2 - \|\mathbf{u}^t\|^2), \tag{7}$$

where $v_t = h(\mathbf{x}^t) - h^*$ and $\mathbf{u}^t = a_t \mathbf{x}^t - (a_t - 1)\mathbf{x}^{t-1} - \mathbf{x}^*$.

Proof 2 *Based on (5), let $\mathbf{x} = \mathbf{x}^t, \mathbf{y} = \mathbf{y}^{t+1}$, then $\mathbf{x}^+ = \mathbf{x}^{t+1}$. So,*

$$h(\mathbf{x}^t) \geq h(\mathbf{x}^{t+1}) + \beta\langle \mathbf{y}^{t+1} - \mathbf{x}^{t+1}, \mathbf{x}^t - \mathbf{y}^{t+1}\rangle + \frac{\beta}{2}\|\mathbf{y}^{t+1} - \mathbf{x}^{t+1}\|^2.$$

That is

$$h(\mathbf{x}^t) - h^* \geq h(\mathbf{x}^{t+1}) - h^* + \beta\langle \mathbf{y}^{t+1} - \mathbf{x}^{t+1}, \mathbf{x}^t - \mathbf{y}^{t+1}\rangle + \frac{\beta}{2}\|\mathbf{y}^{t+1} - \mathbf{x}^{t+1}\|^2,$$

and

$$\frac{2}{\beta}(v_t - v_{t+1}) \geq 2\langle \mathbf{y}^{t+1} - \mathbf{x}^{t+1}, \mathbf{x}^t - \mathbf{y}^{t+1}\rangle + \|\mathbf{y}^{t+1} - \mathbf{x}^{t+1}\|^2. \tag{8}$$

By the same way, let $\mathbf{x} = \mathbf{x}^*, \mathbf{y} = \mathbf{y}^{t+1}$ in (5), it has

$$-\frac{2}{\beta}v_{t+1} \geq 2\langle \mathbf{y}^{t+1} - \mathbf{x}^{t+1}, \mathbf{x}^* - \mathbf{y}^{t+1}\rangle + \|\mathbf{y}^{t+1} - \mathbf{x}^{t+1}\|^2. \tag{9}$$

Let Eq.(8)$\times(a_{t+1} - 1)$+Eq.(9), we have

$$\frac{2}{\beta}[(a_{t+1} - 1)v_t - a_{t+1}v_{t+1}] \geq 2\langle \mathbf{x}^{t+1} - \mathbf{y}^{t+1}, a_{t+1}\mathbf{y}^{t+1} - (a_{t+1} - 1)\mathbf{x}^t - \mathbf{x}^*\rangle + a_{t+1}\|\mathbf{y}^{t+1} - \mathbf{x}^{t+1}\|^2. \tag{10}$$

In addition, Step 2 of FASTA in (3) says that

$$a_{t+1}^2 - a_{t+1} = a_t^2. \tag{11}$$

Thus, Eq.(10)$\times a_{t+1}$ with Lemma 1 is

$$\begin{aligned}
\frac{2}{\beta}[a_t^2 v_t - a_{t+1}^2 v_{t+1}] &\geq 2\langle \mathbf{x}^{t+1} - \mathbf{y}^{t+1}, a_{t+1}^2\mathbf{y}^{t+1} - a_t^2\mathbf{x}^t - \mathbf{x}^*\rangle + a_{t+1}^2\|\mathbf{y}^{t+1} - \mathbf{x}^{t+1}\|^2 && (12)\\
&\geq \|a_{t+1}\mathbf{x}^{t+1} - (a_{t+1} - 1)\mathbf{x}^t - \mathbf{x}^*\|^2 + \|a_{t+1}\mathbf{y}^{t+1} - (a_{t+1} - 1)\mathbf{x}^t - \mathbf{x}^*\|^2 && (13)\\
&= \|\mathbf{u}^{t+1}\|^2 - \|\mathbf{u}^t\|^2. && (14)
\end{aligned}$$

Theorem 2 *[Beck and Teboulle, 2009] Let $\{\mathbf{x}^t, \mathbf{y}^t\}$ be generated by AGD or FISTA. Then for any $T \geq 1$,*

$$h(\mathbf{x}^T) - h^* \leq \frac{2\beta\|\mathbf{x}^0 - \mathbf{x}^*\|^2}{(1+T)^2}. \tag{15}$$

Proof 3 *According Lemma 2, let $c_t = \frac{2}{\beta}a_t^2 v_t$, $b_t = \|\mathbf{u}^t\|^2$, and $c = \|\mathbf{y}^1 - \mathbf{x}^*\|^2 = \|\mathbf{x}^0 - \mathbf{x}^*\|^2$. Then Lemma 3 implies $c_t - c_{t+1} \geq b_{t+1} - b_t$.*

Furthermore, let $\mathbf{x} = \mathbf{x}^*, \mathbf{y} = \mathbf{y}^1$ in (5), then

$$\begin{aligned}
h^* - h(\mathbf{x}^1) &\geq \frac{\beta}{2}\|\mathbf{x}^1 - \mathbf{y}^1\|^2 + \beta\langle \mathbf{y}^1 - \mathbf{x}^*, \mathbf{x}^1 - \mathbf{y}^1\rangle\\
&= \frac{\beta}{2}(\|\mathbf{x}^1 - \mathbf{x}^*\|^2 - \|\mathbf{y}^1 - \mathbf{x}^*\|^2).
\end{aligned}$$

This indicates $c_1 + b_1 \le c$, where $c_1 = \frac{2}{\beta} v_1$ and $b_1 = \|\mathbf{x}^1 - \mathbf{x}^*\|^2$. Thus, for any $T > 0$, it has

$$v_T \le \frac{\beta \|\mathbf{x}^0 - \mathbf{x}^*\|^2}{2a_T^2}. \tag{16}$$

By the induction method, we have justify that $a_t \ge \frac{t+1}{2}, \forall t \ge 1$. So,

$$h(\mathbf{x}^T) - h^* \le \frac{2\beta \|\mathbf{x}^0 - \mathbf{x}^*\|^2}{(1+T)^2}.$$

2 Newton-Raphson Method

2.0.1 Motivation

Think about what GD is? Let us consider the first order Taylor approximation of $f(\mathbf{x} + \mathbf{d})$ around at $\mathbf{x}$ is

$$f(\mathbf{x} + \mathbf{d}) \approx f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{d} \rangle.$$

We need $f(\mathbf{x} + \mathbf{d}) \le f(\mathbf{x})$, so the quantity of $\langle \nabla f(\mathbf{x}), \mathbf{d} \rangle$ should be as negative as possible, then

$$\mathbf{d}_{\mathbf{x}}^* = \arg\min\{\langle \nabla f(\mathbf{x}), \mathbf{d} \rangle, \|d\| \le 1\}. \tag{17}$$

Based on Cauchy inequality, it has $\mathbf{d}_{\mathbf{x}}^* = -\frac{\nabla f(\mathbf{x})}{\|f(\mathbf{x})\|}$. We generalized this idea to the second-order Taylor approximation of $f(\mathbf{x} + \mathbf{d})$ around $\mathbf{x}$ is

$$f(\mathbf{x} + \mathbf{d}) \approx f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{d} \rangle + \frac{1}{2} \mathbf{d}^\top \nabla^2 f(\mathbf{x}) \mathbf{d}.$$

So the quantity of $\langle \nabla f(\mathbf{x}), \mathbf{d} \rangle + \frac{1}{2} \mathbf{d}^\top \nabla^2 f(\mathbf{x}) \mathbf{d}$ should be as negative as possible, then

$$\mathbf{d}_{\mathbf{x}}^* = \arg\min\{\langle \nabla f(\mathbf{x}), \mathbf{d} \rangle + \frac{1}{2} \mathbf{d}^\top \nabla^2 f(\mathbf{x}) \mathbf{d}\}. \tag{18}$$

Thus, $\mathbf{d}^*$ should be the solution of the following *Newton Equation*,

$$\nabla^2 f(\mathbf{x}) \mathbf{d} = -\nabla f(\mathbf{x}). \tag{19}$$

If $\nabla^2 f(\mathbf{x}) \succ 0$, then $\mathbf{d}_{\mathbf{x}}^* = -(\nabla^2 f(\mathbf{x}))^{-1} \nabla f(\mathbf{x})$ is called *Newton direction.*

2.0.2 Algorithm

The Newton-Raphson Algorithm is

$$\begin{aligned} &\mathbf{d}^t = -(\nabla^2 f(\mathbf{x}^t))^{-1} \nabla f(\mathbf{x}^t), \\ &\mathbf{x}^{t+1} = \mathbf{x}^t + \mathbf{d}^t. \end{aligned}$$

Actually, in numerical analysis, Newton's method, also known as the Newton–Raphson method, named after Isaac Newton and Joseph Raphson, is a *root-finding algorithm* which produces successively better approximations to the roots (or zeroes) of a real-valued function. In the optimization community, which root is to find by Newton-Raphson algorithm?

Let us consider the unconstrained optimization problem, and its optimality condition says that the local minimum satisfies $\nabla f(\mathbf{x}) = 0$. So, we need to solve the equation $g(\mathbf{x}) := \nabla f(\mathbf{x}) = 0$. How to do? See Figure 1. That is,

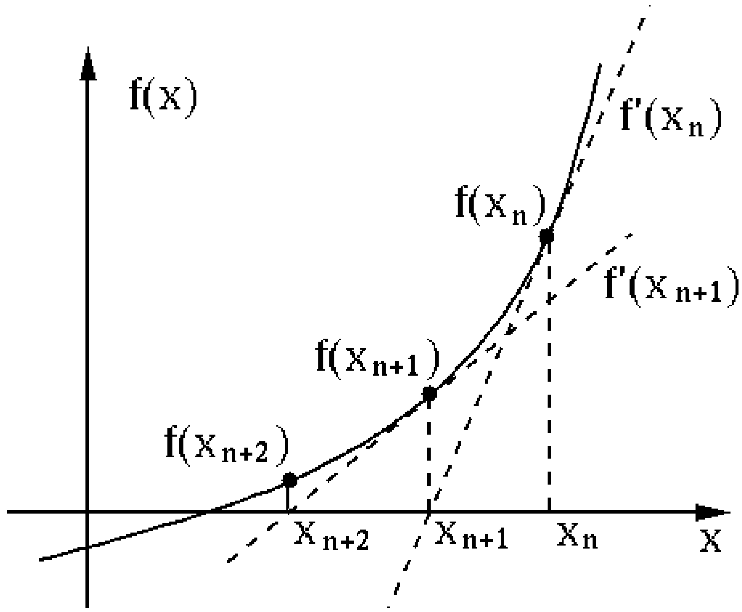

Figure 1: Newton-Raphson algorithm

$$g(\mathbf{x}^t) + \langle \nabla g(\mathbf{x}^t), \mathbf{x} - \mathbf{x}^t \rangle = 0, \text{(Secant Equation)},$$
$$\nabla f(\mathbf{x}^t) + \langle \nabla^2 f(\mathbf{x}^t), \mathbf{x} - \mathbf{x}^t \rangle = 0.$$

So, $\mathbf{x}^{t+1} = \mathbf{x}^t - (\nabla^2 f(\mathbf{x}^t))^{-1} \nabla f(\mathbf{x}^t)$.

Example 1 *Go back to LS problem,*

$$\min_{\mathbf{x}} \frac{1}{2} \|A\mathbf{x} - \mathbf{b}\|^2.$$

The NR algorithm is

$$\begin{aligned}
\mathbf{x}^{t+1} &= \mathbf{x}^t - (\nabla^2 f(\mathbf{x}^t))^{-1} \nabla f(\mathbf{x}^t) \\
&= \mathbf{x}^t - (A^\top A)^{-1} A^\top (A\mathbf{x}^t - \mathbf{b}) \\
&= \mathbf{x}^t - \mathbf{x}^t + (A^\top A)^{-1} A^\top \mathbf{b} \\
&= (A^\top A)^{-1} A^\top \mathbf{b} := \mathbf{x}^*.
\end{aligned}$$

2.1 Convergence

Theorem 3 *Suppose that $f \in C^2$ and $\nabla f(\mathbf{x}^*) = 0$ and $\nabla^2 f(\mathbf{x}^*) \succ 0$. In addition, there exsits an neighborhood of $\mathbf{x}^*$, $\mathcal{N}_\delta(\mathbf{x}^*)$ such that*

$$\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\|, \forall \mathbf{x}, \mathbf{y} \in \mathcal{N}_\delta(\mathbf{x}^*), \tag{20}$$

then

(1) $\lim_{t \to \infty} \mathbf{x}^t = \mathbf{x}^*$ *where* $\{\mathbf{x}^t\}_{t=1}^\infty$ *is generated by Newton-Raphson iteration algorithm.*

(2) there exists a constant c such that

$$\|\mathbf{x}^{t+1} - \mathbf{x}^*\| \leq c \|\mathbf{x}^t - \mathbf{x}^*\|^2.$$

(3) there exists a constant c' such that

$$\|\nabla f(\mathbf{x}^{t+1})\| \leq c'\|\nabla f(\mathbf{x}^t)\|^2.$$

Proof 4 *According to the fact*

$$\nabla f(\mathbf{x}^t) - \nabla f(\mathbf{x}^*) = \int_0^1 \nabla^2 f(\mathbf{x}^t + s(\mathbf{x}^* - \mathbf{x}^t))(\mathbf{x}^t - \mathbf{x}^*)ds,$$

it has

$$\begin{aligned}
\mathbf{x}^{t+1} - \mathbf{x}^* &= \mathbf{x}^t - (\nabla^2 f(\mathbf{x}^t))^{-1}\nabla f(\mathbf{x}^t) - \mathbf{x}^* \\
&= (\nabla^2 f(\mathbf{x}^t))^{-1}(\nabla^2 f(\mathbf{x}^t)(\mathbf{x}^t - \mathbf{x}^*) - \nabla f(\mathbf{x}^t)) \\
&= (\nabla^2 f(\mathbf{x}^t))^{-1}(\nabla^2 f(\mathbf{x}^t)(\mathbf{x}^t - \mathbf{x}^*) - (\nabla f(\mathbf{x}^t) - \nabla f(\mathbf{x}^*))) \\
&= (\nabla^2 f(\mathbf{x}^t))^{-1}\int_0^1 (\nabla^2 f(\mathbf{x}^t) - \nabla^2 f(\mathbf{x}^t + s(\mathbf{x}^* - \mathbf{x}^t)))(\mathbf{x}^t - \mathbf{x}^*)ds.
\end{aligned}$$

By the continuity of $\nabla^2 f$, there exist a constant r such that for any $\mathbf{x} \in (f)$ satisfies $\|\mathbf{x} - \mathbf{x}^*\| \leq r$, then $\|(\nabla^2 f(\mathbf{x}))^{-1}\| \leq 2\|(\nabla^2 f(\mathbf{x}^*))^{-1}\|$. Thus, when $\|\mathbf{x}^0 - \mathbf{x}^*\| \leq \min\{\delta, r, \frac{1}{2L\|\nabla^2 f(\mathbf{x}^*))^{-1}\|}\}$, then

$$\begin{aligned}
\|\mathbf{x}^{t+1} - \mathbf{x}^*\| &\leq \|(\nabla^2 f(\mathbf{x}^t))^{-1}\| \int_0^1 \|\nabla^2 f(\mathbf{x}^t) - \nabla^2 f(\mathbf{x}^t + s(\mathbf{x}^* - \mathbf{x}^t))\|\|\mathbf{x}^t - \mathbf{x}^*\|ds \\
&\leq 2\|(\nabla^2 f(\mathbf{x}^*))^{-1}\| \int_0^1 sL\|\mathbf{x}^t - \mathbf{x}^*\|^2 ds \\
&= L\|(\nabla^2 f(\mathbf{x}^*))^{-1}\|\mathbf{x}^t - \mathbf{x}^*\|^2.
\end{aligned}$$

For the gradient,

$$\begin{aligned}
\|\nabla f(\mathbf{x}^{t+1})\| &= \|\nabla f(\mathbf{x}^{t+1}) - \nabla f(\mathbf{x}^t) - \nabla^2 f(\mathbf{x}^t)\mathbf{d}^t\| \\
&= \|\int_0^1 (\nabla^2 f(\mathbf{x}^t + s\mathbf{d}^t) - \nabla^2 f(\mathbf{x}^t))\mathbf{d}^t ds\| \\
&\leq \frac{L}{2}\|\mathbf{d}^t\|^2 \leq \frac{L}{2}\|(\nabla^2 f(\mathbf{x}^t))^{-1}\|^2\|\nabla f(\mathbf{x}^t)\|^2 \\
&\leq 2L\|(\nabla^2 f(\mathbf{x}^*))^{-1}\|^2\|\nabla f(\mathbf{x}^t)\|^2.
\end{aligned}$$

Remark 1 *(1) $\mathbf{x}^t \to \mathbf{x}^*$ is extremely fast, quadratic convergence rate.*

(2) We need a very good $\mathbf{x}^0$.

(3) $\nabla^2 f(\mathbf{x}^*) \succ 0$.

(4) Every step we need compute a Newton equation. When n is really big, we cannot afford the computational complexity.

(5) $f(\mathbf{x}^{t+1}) \leq f(\mathbf{x}^t)$??? The algorithm is not stable (not decreasing).

In Example 1, we have seen that Newton-Raphson is extremely fast for strongly convex LS problem. Whenever the initial point is close to $\mathbf{x}^*$ or not, one can archive the global minimum by one step. In this part, we will discuss the convergence property of NR-algorithm with line search for general α-strongly convex and β-smooth objective function.

NR-Algorithm with line search as follows:

Algorithm 1 Newton-Raphson Algorithm with Line Search

1: **Input:** Given a initial starting point $\mathbf{x}^0 \in dom(f)$, a tolerance ϵ and $t = 0$. Let $\lambda_t^2 = \nabla f(\mathbf{x}^t)^\top (\nabla^2 f(\mathbf{x}^t))^{-1} \nabla f(\mathbf{x}^t)$ be the *Newton decrement* at $\mathbf{x}^t$.
2: **while** $\lambda_t^2/2 \geq \epsilon$ **do**
3: Backtracking line search a step size s_t such that

$$f(\mathbf{x}^t + s_t \mathbf{d}^t) \leq f(\mathbf{x}^t) + c s_t \nabla f(\mathbf{x}^t)^\top \mathbf{d}^t,$$

where $0 < c < 1$ and $\mathbf{d}^t = (\nabla^2 f(\mathbf{x}^t))^{-1} \nabla f(\mathbf{x}^t)$,
4: $\mathbf{x}^{t+1} = \mathbf{x}^t + s_t \mathbf{d}^t$,
5: $t := t + 1$.
6: **end while**
7: **Output:** $\mathbf{x}^T$, where T is the last step index.

Theorem 4 *Suppose that f is α-strongly convex and β-smooth function, and*

$$\|\nabla^2 f(\mathbf{y}) - \nabla^2 f(\mathbf{y})\| \leq L\|\mathbf{y} - \mathbf{x}\|, \forall \mathbf{x}, \mathbf{y} \in dom(f).$$

Then, there exists numbers η and γ with $0 < \eta \leq \alpha/\gamma$ and $\gamma > 0$ such that the following arguments hold:

(1) If $\|\nabla f(\mathbf{x}^t)\| \geq \eta$,

$$f(\mathbf{x}^{t+1}) - f(\mathbf{x}^t) \leq -\gamma; \tag{21}$$

(2) If $\|\nabla f(\mathbf{x}^t)\| \leq \eta$, then the backtracking line search selects $s_t = 1$, and

$$\frac{L}{2\alpha^2}\|\nabla f(\mathbf{x}^{t+1})\| \leq \left(\frac{L}{2\alpha^2}\|\nabla f(\mathbf{x}^t)\|^2\right). \tag{22}$$

Proof 5 *Please see Page 489 of [Boyd et al., 2004].*

References

[Beck and Teboulle, 2009] Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202.

[Boyd et al., 2004] Boyd, S., Boyd, S. P., and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

[Nesterov, 1998] Nesterov, Y. (1998). Introductory lectures on convex programming volume i: Basic course. *Lecture notes*, 3(4):5.