Large Scale Optimization for Machine Learning:
Lecture 6

Lecturer: Meisam Razaviyayn Scribe: Dongsheng Ding

Sep 12, 2017

1 Remark on the Newton's method

Consider the unconstrained optimization problem

$$\begin{aligned} &\underset{x}{\text{minimize}}\ f(\mathbf{x}) \\ &\text{subject to } \mathbf{x} \in \mathbb{R}^n. \end{aligned} \tag{1}$$

Claim 1 *Consider the Newton's method for (1):*

$$\boldsymbol{x}^{r+1} = \boldsymbol{x}^r - [\nabla^2 f(\boldsymbol{x}^r)]^{-1} \nabla f(\boldsymbol{x}^r) \tag{2}$$

where $\nabla^2 f(\boldsymbol{x}^r)$ is invertible and bounded away from zero. If it converges, then it converges super-linearly.

Proof Assume that (2) converges to $\mathbf{x}^\star$, according to the first order optimality, we have $\nabla f(\mathbf{x}^\star) = 0$. The convergence implies that $\mathbf{x}^r$ can be close to $\mathbf{x}^\star$ in any amount. So, for any $\mathbf{x}^r$ close enough to $\mathbf{x}^\star$, $\nabla f(\mathbf{x}^\star)$ can be expanded as follows

$$\begin{aligned} &\nabla f(\mathbf{x}^\star) = \nabla f(\mathbf{x}^r) + \nabla^2 f(\mathbf{x}^r)(\mathbf{x}^\star - \mathbf{x}^r) + o(\|\mathbf{x}^\star - \mathbf{x}^r\|) = 0 \\ \Leftrightarrow &0 = [\nabla^2 f(\mathbf{x}^r)]^{-1} \nabla f(\mathbf{x}^r) + \mathbf{x}^\star - \mathbf{x}^r + o(\|\mathbf{x}^\star - \mathbf{x}^r\|) \\ \Leftrightarrow &\mathbf{x}^r - \mathbf{x}^\star - [\nabla^2 f(\mathbf{x}^r)]^{-1} \nabla f(\mathbf{x}^r) = o(\|\mathbf{x}^\star - \mathbf{x}^r\|) \\ \Leftrightarrow &\mathbf{x}^{r+1} - \mathbf{x}^\star = o(\|\mathbf{x}^\star - \mathbf{x}^r\|) \end{aligned} \tag{3}$$

where the second line is because of invertible $\nabla^2 f(\mathbf{x}^r)$, and the fourth line uses the Newton's iteration.

Therefore, we have

$$\lim_{r\to\infty} \frac{\mathbf{x}^{r+1} - \mathbf{x}^\star}{\|\mathbf{x}^\star - \mathbf{x}^r\|} = 0 < 1. \tag{4}$$

Q.E.D Claim 1.

Notice that the above claim is under the condition that "if the algorithm converges". In general, it may happen a lot in practice that this naïve Newton algorithm may not even converge.

2 Recap on some basic definitions

Definition 1 *A function $f(\boldsymbol{x}) : \mathbb{R}^n \mapsto \mathbb{R}$ is said to be convex, if one of the following conditions holds:*

(i) $f(\boldsymbol{x})$ is continuously differentiable and $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle, \forall \boldsymbol{x}, \boldsymbol{y}$.

(ii) $f(\boldsymbol{x})$ is twice continuously differentiable and $\nabla^2 f(\boldsymbol{x}) \succeq 0, \forall \boldsymbol{x}$.

The two conditions above are clearly equivalent for twice continuously differentiable functions.

Definition 2 *A function $f(\boldsymbol{x}) : \mathbb{R}^n \mapsto \mathbb{R}$ is said to be strongly convex with a modulus σ, if one of the following conditions holds:*

(i) $f(\boldsymbol{x})$ is continuously differentiable and $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle + \frac{\sigma}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2, \forall \boldsymbol{x}, \boldsymbol{y}$.

(ii) $f(\boldsymbol{x})$ is twice continuously differentiable and $\nabla^2 f(\boldsymbol{x}) \succeq \sigma I, \forall \boldsymbol{x}$.

The above two conditions are clearly equivalent for twice continuously differentiable functions.

Definition 3 *A function $f(\boldsymbol{x}) : \mathbb{R}^n \mapsto \mathbb{R}$ is said to be Lipschitz continuous with a constant L, if one of the following conditions holds:*

(i) $f(\boldsymbol{x})$ is continuously differentiable and $\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|, \forall \boldsymbol{x}, \boldsymbol{y}$.

(ii) $f(\boldsymbol{x})$ is twice continuously differentiable and $\nabla^2 f(\boldsymbol{x}) \preceq LI, \forall \boldsymbol{x}$.

The above two conditions are clearly equivalent for twice continuously differentiable functions.

Definition 4 *If a function $f(\boldsymbol{x}) : \mathbb{R}^n \mapsto \mathbb{R}$ is Lipschitz continuous with a constant L, and strongly convex with a modulus σ, then its condition number can be defined as*

$$\kappa = \frac{L}{\sigma}. \tag{5}$$

3 Convergence of GD for strongly convex objective

The iteration complexity analysis of the gradient descent (GD) for strongly convex objective is introduced.

Lemma 1 *(Descent lemma) If a function $f(\boldsymbol{x}) : \mathbb{R}^n \mapsto \mathbb{R}$ is Lipschitz continuous with a constant L, then*

$$f(\boldsymbol{x}) \leq f(\boldsymbol{y}) + \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2, \forall \boldsymbol{x}, \boldsymbol{y} \tag{6}$$

Proof Let $g(t) = f(\mathbf{y} + t(\mathbf{x} - \mathbf{y}))$, we have

$$\begin{aligned}
f(\mathbf{x}) - f(\mathbf{y}) = g(1) - g(0) &= \int_0^1 g'(\tau) d\tau \\
&= \int_0^1 \nabla f(\mathbf{y} + \tau(\mathbf{x} - \mathbf{y}))^T (\mathbf{x} - \mathbf{y}) d\tau \\
&= \int_0^1 \nabla f(\mathbf{y})^T (\mathbf{x} - \mathbf{y}) d\tau + \int_0^1 \nabla \big(f(\mathbf{y} + \tau(\mathbf{x} - \mathbf{y})) - f(\mathbf{y}) \big)^T (\mathbf{x} - \mathbf{y}) d\tau \\
&\leq \nabla f(\mathbf{y})^T (\mathbf{x} - \mathbf{y}) + \int_0^1 \nabla \big\| f(\mathbf{y} + \tau(\mathbf{x} - \mathbf{y})) - f(\mathbf{y}) \big\| \|\mathbf{x} - \mathbf{y}\| d\tau \\
&\leq \nabla f(\mathbf{y})^T (\mathbf{x} - \mathbf{y}) + \int_0^1 L\tau \|\mathbf{x} - \mathbf{y}\|^2 d\tau \\
&\leq \nabla f(\mathbf{y})^T (\mathbf{x} - \mathbf{y}) + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2
\end{aligned} \tag{7}$$

where the fourth line follows the Cauchy-Schwartz inequality and the fifth line uses the Lipschitz continuity.

Q.E.D Descent lemma.

Claim 2 *Assume that $f(\boldsymbol{x})$ in (1) is Lipschitz continuous with a constant L, and strongly convex with a modulus σ. The gradient descent (GD):*

$$\boldsymbol{x}^{r+1} = \boldsymbol{x}^r - \alpha \nabla f(\boldsymbol{x}^r) \tag{8}$$

requires $O(\kappa log(D/\epsilon))$ iterations to achieve an ϵ-optimal solution, where $\epsilon > 0$ is the desired accuracy, i.e., $f(\mathbf{x}^r - \mathbf{x}^*) = \epsilon$; the parameter $D = f(\boldsymbol{x}^0) - f(\boldsymbol{x}^\star)$ is the initial distance and $\kappa = \frac{L}{\sigma}$ is the condition number.

Proof The proof consists of two parts: (i) Sufficient descent; (ii) Cost-to-go estimation.

We start from the sufficient descent. According to the Descent Lemma, we have

$$\begin{aligned} f(\mathbf{x}^{r+1}) &\leq f(\mathbf{x}^r) + \langle \nabla f(\mathbf{y}^r), \mathbf{x}^{r+1} - \mathbf{x}^r \rangle + \frac{L}{2}\|\mathbf{x}^{r+1} - \mathbf{x}^r\|^2 \\ &= f(\mathbf{x}^r) + \langle \nabla f(\mathbf{y}^r), -\alpha \nabla f(\mathbf{x}^r) \rangle + \frac{L}{2}\alpha^2 \|\nabla f(\mathbf{x}^r)\|^2 \\ &= f(\mathbf{x}^r) - \frac{2\alpha - L\alpha^2}{2}\|\nabla f(\mathbf{x}^r)\|^2 \end{aligned} \tag{9}$$

where the second line following by plugging the GD iteration.

Let $\alpha < \frac{2}{L}$, we have $f(\mathbf{x}^{r+1}) < f(\mathbf{x}^r)$. It means that the objective function is decreased sufficiently after the $r+1$-th iteration.

The cost-to-go estimation is based on the strong convexity as follows. At the optimal solution $\mathbf{x}^\star$, we have

$$\begin{aligned} f(\mathbf{x}^\star) &\geq f(\mathbf{x}^r) + \langle \nabla f(\mathbf{x}^r), \mathbf{x}^\star - \mathbf{x}^r \rangle + \frac{\sigma}{2}\|\mathbf{x}^\star - \mathbf{x}^r\|^2 \\ &\geq \min_{\mathbf{y}} \; f(\mathbf{x}^r) + \langle \nabla f(\mathbf{x}^r), \mathbf{y} \rangle + \frac{\sigma}{2}\|\mathbf{y}\|^2 \\ &= f(\mathbf{x}^r) + \langle \nabla f(\mathbf{x}^r), -\frac{1}{\sigma}\nabla f(\mathbf{x}^r) \rangle + \frac{\sigma}{2}\|\frac{1}{\sigma}\nabla f(\mathbf{x}^r)\|^2 \\ &= f(\mathbf{x}^r) - \frac{1}{2\sigma}\|\nabla f(\mathbf{x}^r)\|^2 \end{aligned} \tag{10}$$

where the third line follows solving the optimization in the second line with $\mathbf{y}^\star = -\frac{1}{\sigma}\nabla f(\mathbf{x}^r)$.

By combing (9) and (10), we have

$$
\begin{aligned}
f(\mathbf{x}^{r+1}) - f(\mathbf{x}^\star) &= f(\mathbf{x}^{r+1}) - f(\mathbf{x}^r) + f(\mathbf{x}^r) - f(\mathbf{x}^\star) \\
&\leq -\frac{2\alpha - L\alpha^2}{2}\|\nabla f(\mathbf{x}^r)\|^2 + f(\mathbf{x}^r) - f(\mathbf{x}^\star) \\
&\leq -\frac{2\alpha - L\alpha^2}{2} 2\sigma(f(\mathbf{x}^r) - f(\mathbf{x}^\star)) + f(\mathbf{x}^r) - f(\mathbf{x}^\star) \qquad (11) \\
&= [1 - \sigma(2\alpha - L\alpha^2)](f(\mathbf{x}^r) - f(\mathbf{x}^\star)) \\
&= (1 - \frac{1}{\kappa})(f(\mathbf{x}^r) - f(\mathbf{x}^\star))
\end{aligned}
$$

where the second line follows (9), the third line follow (10), and the last line chooses $\alpha^\star = \frac{1}{L}$ by minimizing the coefficient.

Let $e(x^r) = f(\mathbf{x}^r) - f(\mathbf{x}^\star)$, from (11), we have

$$
e(\mathbf{x}^r) \leq (1 - \frac{1}{\kappa})^r (f(\mathbf{x}^0) - f(\mathbf{x}^\star)) = (1 - \frac{1}{\kappa})^r D \qquad (12)
$$

To achieve the ϵ-optimal solution, it is sufficient to set $(1 - \frac{1}{\kappa})^r D \backsimeq \epsilon$, i.e., $r\log(1 - \frac{1}{\kappa}) \backsimeq \log(\frac{\epsilon}{D})$. Since $\log(1 - \frac{1}{\kappa}) \backsimeq -\frac{1}{\kappa}$, so $r \backsimeq \kappa\log(\frac{D}{\epsilon})$.

Q.E.D. Claim 2.

Remark 1 According to the Claim 2, the iteration complexity analysis follows two steps:

- Sufficient descent: $f(\mathbf{x}^{r+1}) - f(\mathbf{x}^r) \leq \cdots$ How much the improvement has achieved after the $r+1$-th iteration?
- Cost-to-go estimation: $f(\mathbf{x}^r) - f(\mathbf{x}^\star) \leq \cdots$ What is the distance of the current iterate to $\mathbf{x}^\star$?

This analysis roadmap is valid for most of gradient based optimization algorithms.

Example 1 Regression: In many cases, the incoherent scaling of the variables could cause ill-conditioning (a large condition number κ). This will increase the iteration complexity. How to control κ to avoid the ill-conditioning?

One simple way avoiding ill-conditioning for regression is to normalize the features into the same scaling.

4 Convergence of GD for general convex objective

Claim 3 *Assume that $f(\boldsymbol{x})$ in (1) is convex, and Lipschitz continuous with a constant L. The gradient descent (GD):*

$$\boldsymbol{x}^{r+1} = \boldsymbol{x}^r - \alpha \nabla f(\boldsymbol{x}^r) \tag{13}$$

requires $O(\frac{LD^2}{\epsilon})$ iterations to achieve an ϵ-optimal solution where $\epsilon > 0$ is the desired accuracy; $D = \|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|$ and L is the Lipschitz constant of the gradient of the objective function.

Proof The proof consists of two parts: (i) Sufficient descent; (ii) Cost-to-go estimation.

We start from the sufficient descent. Similar to (9), according to the Descent Lemma, we have

$$f(\mathbf{x}^{i+1}) \leq f(\mathbf{x}^i) - \frac{2\alpha - L\alpha^2}{2}\|\nabla f(\mathbf{x}^i)\|^2, \alpha < \frac{2}{L} \tag{14}$$

By summing (14) from $i = 0$ to $i = r$, we have

$$\begin{aligned}
\sum_{i=0}^{r} \|\nabla f(\mathbf{x}^i)\|^2 &= \frac{2}{\alpha(2-\alpha L)}(f(\mathbf{x}^0) - f(\mathbf{x}^{r+1})) \\
&\leq \frac{2}{\alpha(2-\alpha L)}(f(\mathbf{x}^0) - f(\mathbf{x}^\star)) \\
&\leq \frac{L\|\mathbf{x}^0 - \mathbf{x}^\star\|^2}{\alpha(2-\alpha L)}
\end{aligned} \tag{15}$$

where the second line follows the optimality of $\mathbf{x}^\star$ and the third line follows the Descent Lemma.

The cost-to-go estimation is based on the convexity as follows. At the optimal solution $\mathbf{x}^\star$, we have

$$\begin{aligned}
\|\mathbf{x}^{i+1} - \mathbf{x}^\star\|^2 &= \|\mathbf{x}^i - \alpha\nabla f(\mathbf{x}^i) - \mathbf{x}^\star\|^2 \\
&= \|\mathbf{x}^i - \mathbf{x}^\star\|^2 - 2\alpha\langle \nabla f(\mathbf{x}^i), \mathbf{x}^i - \mathbf{x}^\star\rangle + \alpha^2\|\nabla f(\mathbf{x}^i)\|^2 \\
&\leq \|\mathbf{x}^i - \mathbf{x}^\star\|^2 - 2\alpha(f(\mathbf{x}^i) - f(\mathbf{x}^\star)) + \alpha^2\|\nabla f(\mathbf{x}^i)\|^2
\end{aligned} \tag{16}$$

where the third line follows the convexity of $f(\mathbf{x})$ at $\mathbf{x}^i$.

By summing (16) from $i = 1$ to $i = r$, we have

$$\begin{aligned}\sum_{i=0}^{r}(f(\mathbf{x}^i) - f(\mathbf{x}^\star)) &\leq \frac{1}{2\alpha}\Big(\|\mathbf{x}^0 - \mathbf{x}^\star\|^2 - \|\mathbf{x}^{r+1} - \mathbf{x}^\star\|^2 + \alpha^2 \sum_{i=1}^{r} \|\nabla f(\mathbf{x}^i)\|^2\Big) \\ &\leq \frac{1}{2\alpha}\Big(\|\mathbf{x}^0 - \mathbf{x}^\star\|^2 + \alpha^2 \sum_{i=1}^{r} \|\nabla f(\mathbf{x}^i)\|^2\Big) \\ &\leq \frac{1}{2\alpha}\Big(\|\mathbf{x}^0 - \mathbf{x}^\star\|^2 + \frac{L\alpha^2\|\mathbf{x}^0 - \mathbf{x}^\star\|^2}{\alpha(2-\alpha L)}\Big) \\ &= \frac{\|\mathbf{x}^0 - \mathbf{x}^\star\|^2}{\alpha(2-\alpha L)}\end{aligned} \tag{17}$$

where the third line follows (15).

Finally, we combine the results in (14) and (17). It is obvious from (14) that $f(\mathbf{x}^r)$ decreases monotonically:

$$f(\mathbf{x}^r) - f(\mathbf{x}^\star) \leq f(\mathbf{x}^i) - f(\mathbf{x}^\star), i = 0, 1, \cdots, r \tag{18}$$

By summing them in (18), we have

$$\begin{aligned} f(\mathbf{x}^r) - f(\mathbf{x}^\star) &\leq \frac{1}{r+1}\sum_{i=0}^{r}(f(\mathbf{x}^i) - f(\mathbf{x}^\star)) \\ &\leq \frac{\|\mathbf{x}^0 - \mathbf{x}^\star\|^2}{\alpha(2-\alpha L)}\frac{1}{r+1} \\ &= \frac{L\|\mathbf{x}^0 - \mathbf{x}^\star\|^2}{r+1}\end{aligned} \tag{19}$$

where the second line follows (17) and the last line picks the tight upper bound at $\alpha^\star = \frac{1}{L}$.

Let $e(\mathbf{x}^r) = f(\mathbf{x}^r) - f(\mathbf{x}^\star)$, to achieve the ϵ-optimal solution, it is sufficient to have $\frac{L\|\mathbf{x}^0-\mathbf{x}^\star\|^2}{r+1} = \frac{LD^2}{r+1} \backsimeq \epsilon$. When r is large enough, $r = O(\frac{LD^2}{\epsilon})$.

Q.E.D. Claim 3.

Remark 2 The GD has a sub-linear convergence rate for general convex case as $\epsilon = O(\frac{1}{r})$. The upper-bound rate is order tight and also dimension-independent.

5 Optimal first order methods

The iteration complexity of GD for strongly convex case is

$$r = O\Big(\kappa \log(\frac{1}{\epsilon})\Big) \tag{20}$$

where κ is the condition number, and ϵ is the accuracy of the ϵ-optimal solution.

The iteration complexity of GD for general convex case is

$$r = O\Big(\frac{L}{\epsilon}\Big) \tag{21}$$

where κ is the condition number, L is the Lipschitz constant, and ϵ is the accuracy of the ϵ-optimal solution.

Can we improve the these bounds for these first order methods? What are optimal first order methods? Before answering them, it is necessary to setup the definitions of classes of convex optimization problems, and first order algorithms.

Definition 5 *A class of smooth unconstrained convex optimization problems: $\mathcal{P}(D, L)$, is defined satisfying:*

- $||\boldsymbol{x}^0 - \boldsymbol{x}^\star|| \leq D$;
- $||\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})|| \leq L||\boldsymbol{x} - \boldsymbol{y}||$.

where $D > 0$ and L is the Lipschitz constant of $f(\boldsymbol{x})$.

Definition 6 *The oracle model of first order algorithms: Ω, is defined satisfying:*

- *at iteration r, an algorithm $\mathcal{A} \in \Omega$ uses a linear combination of $\boldsymbol{x}^0, \cdots, \boldsymbol{x}^r$ and $\nabla f(\boldsymbol{x}^0), \cdots, \nabla f(\boldsymbol{x}^r)$ to generate $\boldsymbol{x}^{r+1}$;*
- *given $\boldsymbol{x}^r$, the oracle Ω returns $\nabla f(\boldsymbol{x}^r)$;*
- *the complexity of an algorithm $\mathcal{A} \in \Omega$ is*

$$\mathcal{C}_\epsilon(\mathcal{A}) = \sup_{f \in \mathcal{P}(D,L)} \min \Big\{ r | f(\boldsymbol{x}^r) - f(\boldsymbol{x}^\star) \leq \epsilon \Big\}. \tag{22}$$

Claim 4 *Given a class of smooth unconstrained convex optimization problems: $\mathcal{P}(D, L)$, and the oracle Ω for first order algorithms, the iteration complexity $\mathcal{C}_\epsilon(\mathcal{A})$ satisfies*

$$\sqrt{\frac{3}{32}} \min\left\{n, \sqrt{LD^2\epsilon^{-1}}\right\} \leq \inf_{\mathcal{A}\in\Omega} \mathcal{C}_\epsilon(\mathcal{A}) \leq \sqrt{4LD^2\epsilon^{-1}}. \tag{23}$$

The upper bound was proved by Yu. Nesterov in 1983. Clearly, the GD is not order optimal. It should be noted that the upper bound is also dimension-independent. This is interesting, especially for high dimensional problems.

Before showing the Nesterov accelerated GD, one motivation of looking at the optimal first order method is based on another formulation of the GD as follows.

$$\begin{aligned} \mathbf{x}^{r+1} &= \mathbf{x}^r - \alpha\nabla f(\mathbf{x}^r) \\ &= \arg\min_{\mathbf{y}} \; f(\mathbf{x}^r) + \langle \nabla f(\mathbf{x}^r), \mathbf{y} - \mathbf{x}^r\rangle + \frac{1}{2\alpha}\|\mathbf{y} - \mathbf{x}^r\|^2 \end{aligned} \tag{24}$$

where the optimal $\mathbf{y}^\star$ of the second line is exactly the GD iteration.

It is interesting that the GD iteration $\mathbf{x}^{r+1}$ is equivalent to minimizing a local quadratic approximation of $f(\mathbf{x})$ at $\mathbf{x}^r$ as shown in (24). If given an oracle Ω, this can be extended to a local quadratic approximation of $f(\mathbf{x})$ at all previous iterations $\mathbf{x}^i, i = 0, \cdots, r$.

The Nesterov accelerated GD can be explained as follows. Two testing sequences $\{\mathbf{x}^r\}, \{\mathbf{y}^r\}$ are chosen smartly such that

$$A_r f(\mathbf{x}^r) \leq \min_{\mathbf{y}} \sum_{i=0}^{r} a_i\Big(f(\mathbf{y}^i) + \langle \nabla f(\mathbf{y}^i), \mathbf{y} - \mathbf{y}^i\rangle\Big) + \frac{L}{2}\|\mathbf{y} - \mathbf{x}^0\|^2 \tag{25}$$

where $A_r = \sum_{i=0}^{r} a_i$ and $a_i \geq 0$.

The minimizer in (25) is denoted by

$$\mathbf{z}^r = \mathbf{x}^0 - \frac{1}{L}\sum_{i=0}^{r} a_i \nabla f(\mathbf{y}^i). \tag{26}$$

Claim 5 *For any continuously function $f(\boldsymbol{x}) \in \mathcal{P}(D, L)$, if (25) holds for*

all r, then

$$f(\boldsymbol{x}^r) - f(\boldsymbol{x}^\star) \leq \frac{LD^2}{2A_r}. \tag{27}$$

where $D = ||\boldsymbol{x}^0 - \boldsymbol{x}^\star||$.

Proof By plugging $\mathbf{y} = \mathbf{x}^\star$ into (25), we have

$$\begin{aligned} A_r f(\mathbf{x}^r) &\leq \sum_{i=0}^{r} a_i(f(\mathbf{y}^i) + \langle \nabla f(\mathbf{y}^i), \mathbf{x}^\star - \mathbf{y}^i \rangle) + \frac{L}{2}||\mathbf{x}^\star - \mathbf{x}^0||^2 \\ &\leq f(\mathbf{x}^\star) \sum_{i=0}^{r} a_i + \frac{L}{2}||\mathbf{x}^0 - \mathbf{x}^\star||^2 \\ &= A_r f(\mathbf{x}^\star) + \frac{L}{2}||\mathbf{x}^0 - \mathbf{x}^\star||^2 \end{aligned} \tag{28}$$

where the second line follows the convexity of $f(\mathbf{x})$ at $\mathbf{y}^i$.

Therefore, we have $f(\mathbf{x}^r) - f(\mathbf{x}^\star) \leq \frac{LD^2}{2A_r}$, where $D = ||\mathbf{x}^0 - \mathbf{x}^\star||$.

Q.E.D. Claim 5.

A general version of the Nesterov accelerated GD is

$$\begin{aligned} \mathbf{y}^{r+1} &= (1 - \theta^r)\mathbf{x}^r + \theta^r \mathbf{z}^r \\ \mathbf{x}^{r+1} &= \mathbf{y}^{r+1} - \frac{1}{L}\nabla f(\mathbf{y}^{r+1}) \end{aligned} \tag{29}$$

where $\theta^r = \frac{a_{r+1}}{A_{r+1}}$.

When $a_0 \in [0, 1]$ and $a_r^2 \leq A_r, r \geq 1$, it can be shown that (29) satisfies (25) and this is not covered in our lecture.