Lecture 4: Optimal First Order Methods

- Unconstrained smooth convex minimization
- Analysis of classical methods in the degenerate & nondegenerate cases
- Oracle model of computation
- Optimal first order methods: degenerate/nondegenerate cases
- Lower and upper bounds on iteration complexity
- Dependence on the condition number

Unconstrained Convex Minimization: Degenerate Case

Let f be continuously differentiable with Lipschitz gradient, i.e.,

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|$$

L is the modulus of the Hessian (if exists): $\mathbf{0} \preceq \nabla^2 f(\boldsymbol{x}) \preceq L\boldsymbol{I}$.

Consider the gradient method $\boldsymbol{x}^{r+1} = \boldsymbol{x}^r - \alpha \nabla f(\boldsymbol{x}^r)$, with $0 < \alpha < 2/L$. Then

$$f(\boldsymbol{x}^r) - f(x^*) \leq \left(\frac{\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2}{\alpha(2 - \alpha L)}\right)\frac{1}{r}, \quad r \geq 1.$$

- $\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|$ is the distance from the optimal solution (set).
- Only sublinear convergence rate (in the absence of strong convexity).
- Rate is dimension-independent.

Analysis of Gradient Method: Degenerate Case

Step 1. Use definition and the Lipschitz condition

$$
\begin{aligned}
f(\boldsymbol{x}^{i+1}) &\leq f(\boldsymbol{x}^i) + \langle \nabla f(\boldsymbol{x}^i), \boldsymbol{x}^{i+1} - \boldsymbol{x}^i \rangle + \frac{L}{2}\|\boldsymbol{x}^{i+1} - \boldsymbol{x}^i\|^2 \\
&= f(\boldsymbol{x}^i) - \alpha\left(1 - \frac{\alpha L}{2}\right)\|\nabla f(\boldsymbol{x}^i)\|^2
\end{aligned}
$$

implying

$$
\sum_{i=0}^{r} \|\nabla f(\boldsymbol{x}^i)\|^2 \leq \frac{2}{\alpha(2-\alpha L)}\left(f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)\right) \leq \frac{L\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2}{\alpha(2-\alpha L)}
$$

Step 2. Use convexity of f to obtain

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

which implies

$$\begin{aligned}\sum_{i=0}^{r}(f(\boldsymbol{x}^i) - f(\boldsymbol{x}^*)) &\leq \frac{1}{2\alpha}\left[\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2 + \frac{\alpha^2 L\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2}{\alpha(2-\alpha L)}\right] \\ &= \frac{\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2}{\alpha(2-\alpha L)}\end{aligned}$$

Step 3. By monotonicity, we have

$$f(\boldsymbol{x}^r) - f(\boldsymbol{x}^*) \leq \left(\frac{\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2}{\alpha(2-\alpha L)}\right)\frac{1}{r+1}, \quad r \geq 1.$$

Choose $\alpha = 1/L$ yields

$$f(\boldsymbol{x}^r) - f(\boldsymbol{x}^*) \leq \frac{L\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2}{r+1}, \quad r \geq 1.$$

This upper bound is order-tight (i.e., can construct a quadratic f for which after r gradient descent steps the gap to minimum is of order $L\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|^2/r$).

Optimal First Order Methods?

- Let $P(D, L)$ denote the class of smooth unconstrained convex optimization problems with $\|\boldsymbol{x}^0 - \boldsymbol{x}^*\| \leq D$ and $\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|$.
- Consider the oracle model Ω for the first order algorithms:
  - ⋆ at iteration r, the algorithm takes any linear combination of $\boldsymbol{x}^0, \boldsymbol{x}^1, \ldots, \boldsymbol{x}^r$ and $\nabla f(\boldsymbol{x}^0), \nabla f(\boldsymbol{x}^1), \ldots, \nabla f(\boldsymbol{x}^r)$ to generate $\boldsymbol{x}^{r+1}$.
  - ⋆ given any $\boldsymbol{x}^r$, the oracle returns $\nabla f(\boldsymbol{x}^r)$
  - ⋆ the complexity of an algorithm $\mathcal{A} \in \Omega$ is

$$C_\epsilon(\mathcal{A}) = \sup_{f \in P(D,L)} \min\{r \mid f(\boldsymbol{x}^r) - f(\boldsymbol{x}^*) \leq \epsilon\}$$

- Bounds on the complexity (interesting only for problems with large dimensions):

$$O(1) \min\{n, \sqrt{LD^2\epsilon^{-1}}\} \leq \inf_{\mathcal{A} \in \Omega} C_\epsilon(\mathcal{A}) \leq \sqrt{4LD^2\epsilon^{-1}}$$

- Thus, the classical **gradient descent** method is not order-optimal!

An Optimal First Order Method

Nesterov (1983) proposed an order-optimal first order method.
Define two sequences $\{\boldsymbol{x}^r\}$ and $\{\boldsymbol{y}^r\}$ (test points) satisfying, for all $r \geq 1$,

$$A_r f(\boldsymbol{x}^r) \leq \min_{\boldsymbol{y}} \left\{ \frac{L}{2} \|\boldsymbol{y} - \boldsymbol{x}^0\|^2 + \sum_{i=0}^{r} a_i \left(f(\boldsymbol{y}^i) + \langle \nabla f(\boldsymbol{y}^i), \boldsymbol{y} - \boldsymbol{y}^i \rangle \right) \right\}, \tag{1}$$

where $A_r = \sum_{i=0}^{r} a_i$, with $a_i \geq 0$. Denote the minimizer of (1) by $\boldsymbol{z}^r$ (explicit formula?). Define $\theta^r = a_{r+1}/A_{r+1}$ and update using

$$\begin{cases} \boldsymbol{y}^{r+1} = (1 - \theta^r)\boldsymbol{x}^r + \theta^r \boldsymbol{z}^r, \\ \boldsymbol{x}^{r+1} = \boldsymbol{y}^{r+1} - \frac{1}{L} \nabla f(\boldsymbol{y}^{r+1}), \end{cases} \quad r \geq 1. \tag{2}$$

Iteration Complexity: For any continuously differentiable $f \in P(L, D)$, if (1) holds for all r, then after r steps,

$$f(\boldsymbol{x}^r) - f(\boldsymbol{x}^*) \leq \frac{LD^2}{2A_r}, \quad r \geq 1.$$

Prove using (1) (set $\boldsymbol{y} = \boldsymbol{x}^*$ and use convexity).

Choose Optimal Parameters

Claim: if $a_0 \in (0, 1]$, and $a_r^2 \le A_r$ for $r \ge 1$, then by induction (1) holds.

- If $a_r^2 = A_r$, then

$$a_r^2 - a_r = a_{r-1}^2 \quad \Rightarrow \quad a_r = \frac{1}{2}\left(1 + \sqrt{4a_{r-1}^2 + 1}\right)$$

- A specific choice: $a_r = \frac{(r+1)}{2}$, then $A_r = \frac{(r+1)(r+2)}{4}$ and $\theta^r = \frac{2}{(r+3)}$.

- The iteration bound becomes

$$f(\boldsymbol{x}^r) - f(\boldsymbol{x}^*) \le \frac{2LD^2}{(r+2)(r+1)}, \quad r \ge 1.$$

which is order-optimal.

A Recursive Description

Define a scalar sequence $\{a_r\}$ satisfying

$$a_r = \frac{1}{2}\left(1 + \sqrt{1 + 4a_{r-1}^2}\right), \text{ with } a_0 = 0.$$

Then $a_r \geq (r+1)/2$ for all $r \geq 1$. Let

$$t^r := (a_r - 1)/a_{r+1}, \text{ for } r \geq 1.$$

An optimal first order method (Nesterov)

1. Initialization: $\boldsymbol{x}^0 = \boldsymbol{x}^1 = \boldsymbol{0}$.
2. Iteration $r \geq 1$: first generate a test point using extrapolation $\boldsymbol{y}^{r+1} = (1 + t^r)\boldsymbol{x}^r - t^r \boldsymbol{x}^{r-1}$. Then let $\boldsymbol{x}^{r+1} = \boldsymbol{y}^{r+1} - \frac{1}{L}\nabla f(\boldsymbol{y}^{r+1})$.

Remarks: fixed step size; non-monotone. Both can be easily corrected.

Recursion

Denote $\boldsymbol{g}^r = \nabla f(\boldsymbol{y}^r)/L$ for $r \geq 1$ and $\boldsymbol{g}^0 \equiv 0$. Then $\boldsymbol{z}^r = -\sum_{i=0}^{r} a_i \boldsymbol{g}^i$, and

$$\begin{aligned} \boldsymbol{y}^{r+1} &= (1+t^r)\boldsymbol{x}^r - t^r \boldsymbol{x}^{r-1}, \\ \boldsymbol{x}^{r+1} &= \boldsymbol{y}^{r+1} - \boldsymbol{g}^{r+1}, \quad \text{with } t^r = (a_r - 1)/a_{r+1}. \end{aligned}$$

A simple recursion:

$$\left[a_{r+1}\boldsymbol{y}^{r+1} - (a_{r+1}-1)\boldsymbol{x}^r\right] = \left[a_r \boldsymbol{y}^r - (a_r - 1)\boldsymbol{x}^{r-1}\right] - a_r \boldsymbol{g}^r$$

implying (for $r \geq 1$)

$$\boldsymbol{y}^{r+1} = (a_{r+1}-1)(\boldsymbol{x}^r - \boldsymbol{y}^{r+1}) - \sum_{i=0}^{r} a_i \boldsymbol{g}^i = (a_{r+1}-1)(\boldsymbol{x}^r - \boldsymbol{y}^{r+1}) + \boldsymbol{z}^r. \quad (3)$$

or

$$\boldsymbol{y}^{r+1} = (1 - a_{r+1}^{-1})\boldsymbol{x}^r + (a_{r+1}^{-1})\boldsymbol{z}^r$$

which corresponds to the nonrecursive version (2) with $\theta_r = a_{r+1}/A_{r+1} = a_{r+1}^{-1}$.

Iteration Complexity Analysis

Denote $e^r = f(\boldsymbol{x}^r) - f(\boldsymbol{x}^*)$. Use Taylor expansion of $f(\boldsymbol{x}^{r+1})$ at $\boldsymbol{y}^{r+1}$ and the definition of $\boldsymbol{x}^{r+1}$

$$f(\boldsymbol{x}^{r+1}) - f(\boldsymbol{x}) \le L\langle \boldsymbol{g}^{r+1}, \boldsymbol{y}^{r+1} - \boldsymbol{x}\rangle - \frac{L}{2}\|\boldsymbol{g}^{r+1})\|^2$$

Choose $\boldsymbol{x} = \boldsymbol{x}^r$ to obtain a "**sufficient decrease**" estimate

$$e^{r+1} - e^r = f(\boldsymbol{x}^{r+1}) - f(\boldsymbol{x}^r) \le L\langle \boldsymbol{g}^{r+1}, \boldsymbol{y}^{r+1} - \boldsymbol{x}^r\rangle - \frac{L}{2}\|\boldsymbol{g}^{r+1}\|^2. \tag{4}$$

Also, choose $\boldsymbol{x} = \boldsymbol{x}^*$ and use (3) to obtain a estimate of the "**cost-to-go**"

$$\begin{aligned} e^{r+1} &\le L\langle \boldsymbol{g}^{r+1}, \boldsymbol{y}^{r+1} - \boldsymbol{x}^*\rangle - \frac{L}{2}\|\boldsymbol{g}^{r+1}\|^2 \\ &= L\langle \boldsymbol{g}^{r+1}, \boldsymbol{z}^r - \boldsymbol{x}^*\rangle + L(a_{r+1} - 1)\langle \boldsymbol{g}^{r+1}, \boldsymbol{x}^r - \boldsymbol{y}^{r+1}\rangle - \frac{L}{2}\|\boldsymbol{g}^{r+1}\|^2 \quad (5) \end{aligned}$$

Multiply (4) by $(a_{r+1}-1)$ and add it to (5) to obtain

$$a_{r+1}e^{r+1}-(a_{r+1}-1)e^r \le L\langle \boldsymbol{g}^{r+1}, \boldsymbol{z}^r-\boldsymbol{x}^*\rangle - \frac{L}{2}a_{r+1}\|\boldsymbol{g}^{r+1}\|^2$$

Multiplying both sides by a_{r+1} and noting $\boldsymbol{z}^{r+1}=\boldsymbol{z}^r-a_{r+1}\boldsymbol{g}^{r+1}$ gives

$$a_{r+1}^2e^{r+1}-a_{r+1}(a_{r+1}-1)e^r \le -L\langle a_{r+1}\boldsymbol{g}^{r+1}, \boldsymbol{x}^*\rangle - \frac{L}{2}(\|\boldsymbol{z}^{r+1}\|^2-\|\boldsymbol{z}^r\|^2)$$

If $a_{r+1}(a_{r+1}-1)=a_r^2$ (which is equivalent to $a_r^2=A_r$), then

$$a_{r+1}^2e^{r+1}-a_r^2e^r \le -L\langle a_{r+1}\boldsymbol{g}^{r+1}, \boldsymbol{x}^*\rangle - \frac{L}{2}(\|\boldsymbol{z}^{r+1}\|^2-\|\boldsymbol{z}^r\|^2), \quad r\ge 1.$$

Summing over r and using $\boldsymbol{z}^{r+1}=-\sum_{i=0}^{r+1}a_i\boldsymbol{g}^i$, $\boldsymbol{z}^1=\boldsymbol{0}$, gives

$$a_{r+1}^2e^{r+1}-a_0^2e^0 \le L\langle \boldsymbol{z}^{r+1}, \boldsymbol{x}^*\rangle - \frac{L}{2}\|\boldsymbol{z}^{r+1}\|^2 \le \frac{L\|\boldsymbol{x}^*\|^2}{2},$$

implying $e^{r+1}\le (L\|\boldsymbol{x}^*\|^2)/(2a_{r+1}^2)=LD^2/(2A_{r+1})$.

Optimal First Order Methods for Strongly Convex Problems

Suppose f is strongly convex s.t. $f(\boldsymbol{x}) - f(\boldsymbol{x}^*) \geq \sigma\|\boldsymbol{x} - \boldsymbol{x}^*\|^2$. The condition number $\kappa = L/\sigma$. An ϵ-relative optimal solution $\boldsymbol{x}^r$ satisfies

$$f(\boldsymbol{x}^r) - f(\boldsymbol{x}^*) \leq \epsilon(f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)).$$

Running Nesterov's method with restart can yield an ϵ-relative optimal solution with an iteration complexity of

$$O(1)\sqrt{\kappa}\ln(1/\epsilon). \tag{6}$$

Strategy: Start from $\boldsymbol{x}^0$, run Nesterov's method for $i = \sqrt{2\kappa}$ iterations. Set $\boldsymbol{x}^0 = \boldsymbol{x}^i$ and restart, etc.

Each round has $\sqrt{2\kappa}$ iterations. After the r-th round, we have

$$f(\boldsymbol{x}^{ir}) - f(\boldsymbol{x}^*) \leq \frac{L\|\boldsymbol{x}^{i(r-1)} - \boldsymbol{x}^*\|^2}{i^2} \leq \frac{1}{2}(f(\boldsymbol{x}^{i(r-1)}) - f(\boldsymbol{x}^*)).$$

This implies (6).

Impact of Condition Number

- For strongly convex problems, Nesterov's method (with multi-start) has a complexity that is the same as any linearly convergent method (e.g., **gradient descent**), with a factor of $\sqrt{\kappa}$ improvement.

- In practice, $\ln(1/\epsilon)$ is small (less than 20), but κ can be large (e.g., 10^3 – 10^6). A removal of a $\sqrt{\kappa}$ factor is significant.

- **Lower bound:**

$$\inf_{\mathcal{A}\in\Omega} C_\epsilon(\mathcal{A}) \geq O(1)\min\{n, \sqrt{\kappa}\ln(1/2\epsilon)\}$$

  So for strongly convex problems Nesterov's method is order-optimal with respect to κ.

- Nesterov's method, without restart, is linearly convergent, with iteration complexity $O(1)\sqrt{\kappa}\ln(1/\epsilon)$. [The sequence $\{a_r\}$ depends on κ.]