Smoothing for non-smooth optimization, lecture 2

Last time:

- Basics of convex analysis

Today:

- Standard gradient scheme (smooth): $O(1/\epsilon)$
  - unconstrained
  - simply-constrained
- Optimal gradient scheme (simply-constrained): $O(1/\sqrt{\epsilon})$
- Optimal subgradient scheme (non-smooth): $O(1/\epsilon^2)$

Unconstrained optimization

$$\min\{f(x) : x \in E\}$$

Notation

- $\bar{f} := \min\{f(x) : x \in E\}$
- $\bar{x} := \operatorname{argmin}\{f(x) : x \in E\}$

Assume (until we state otherwise)

- E: finite-dimensional Euclidean space
- $f : E \to \mathbb{R}$ is a smooth convex function
- ∇f is Lipschitz with constant L

Algorithm 1

- Pick $x_0 \in E$
- For $k = 0, 1, \ldots$
  $$x_{k+1} := x_k - h_k \nabla f(x_k)$$

Theorem 2 *If $h_j = \frac{1}{L}$ for $j = 1, 2, \ldots$, then*

$$f(x_k) - \bar{f} \leq \frac{2L(f(x_0) - \bar{f})\|x_0 - \bar{x}\|^2}{2L\|x_0 - \bar{x}\|^2 + k(f(x_0) - \bar{f})} \sim \frac{1}{k}$$

Immediate (but important) observations

- Speed of convergence to optimal value is $O(1/k)$
- Complexity to get within ϵ of optimal value is $O(1/\epsilon)$
- Gradient-step update at each iteration:
  $$x - \frac{1}{L}\nabla f(x) = \operatorname*{argmin}_y \left\{ \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2 \right\}.$$

$x \to x - \frac{1}{L}\nabla f(x)$

Proof. Since f is convex and ∇f is Lipschitz with constant L,

- $0 \le f(y) - f(x) - \langle \nabla f(x), y - x\rangle \le \frac{L}{2}\|y - x\|^2$ for all x, y
- $\frac{1}{L}\|\nabla f(y) - \nabla f(x)\|^2 \le \langle \nabla f(y) - \nabla f(x), y - x\rangle$ for all $x, y \in C$

Hence for $x \in E$

$$f\left(x - \frac{1}{L}\nabla f(x)\right) \le f(x) - \frac{1}{2L}\|\nabla f(x)\|^2,$$

$$f(x) - \bar{f} \le \langle \nabla f(x), x - \bar{x}\rangle \le \|\nabla f(x)\|\,\|x - \bar{x}\|,$$

and

$$\langle \nabla f(x), x - \bar{x}\rangle \ge \frac{1}{L}\|\nabla f(x)\|^2.$$

Thus

$$\|x - \frac{1}{L}\nabla f(x) - \bar{x}\|^2 \le \|x - \bar{x}\|^2 - \frac{1}{L^2}\|\nabla f(x)\|^2 \le \|x - \bar{x}\|^2.$$

Put $\Delta_k := f(x_k) - \bar{f}$, $r_k := \|x_k - \bar{x}\|$. We get

$$\Delta_{k+1} \le \Delta_k - \frac{1}{2L}\|\nabla f(x_k)\|^2 \le \Delta_k - \frac{\Delta_k^2}{2Lr_k^2} \le \Delta_k - \frac{\Delta_k^2}{2Lr_0^2}.$$

Therefore

$$\frac{1}{\Delta_{k+1}} \ge \frac{1}{\Delta_k} + \frac{\Delta_k}{2Lr_0^2\Delta_{k+1}} \ge \frac{1}{\Delta_k} + \frac{1}{2Lr_0^2}$$

Consequently

$$\frac{1}{\Delta_k} \ge \frac{1}{\Delta_0} + \frac{k}{2Lr_0^2},$$

that is,

$$\Delta_k \le \frac{2L\Delta_0 r_0^2}{2Lr_0^2 + k\Delta_0}$$

□

Simply-constrained optimization

$$\min\{f(x) : x \in Q\}$$

Assume $Q \subseteq E$ is a "simple" compact convex set.

Overwrite previous notation

- $\bar{f} := \min\{f(x) : x \in Q\}$
- $\bar{x} := \operatorname{argmin}\{f(x) : x \in Q\}$

Key ingredient: Consider the *gradient mapping*

$$T_Q(x) := \operatorname*{argmin}_y \left\{\langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2 : y \in Q\right\}$$

Algorithm 3

- Pick $x_0 \in E$
- For $k = 0, 1, \ldots$
  $x_{k+1} := T_Q(x_k)$

Theorem 4

$$f(x_k) - \bar{f} \le \frac{(f(x_0) - \bar{f})(2L\|x_0 - \bar{x}\|^2 + f(x_0) - \bar{f})}{2L\|x_0 - \bar{x}\|^2 + f(x_0) - \bar{f} + k}.$$

Similar speed of convergence to that in the unconstrained case.

Proof. (similar to that of Theorem 2.)

Main steps: Put $g(x) := L(x - T_Q(x))$ (sort of $\nabla f(x)$). Then

$$f(T_Q(x)) \le f(x) - \frac{1}{2L}\|g(x)\|^2,$$

$$f(T_Q(x)) - \bar{f} \le \langle g(x), x - \bar{x}\rangle \le \|g(x)\|\|x - \bar{x}\|,$$

and

$$\langle g(x), x - \bar{x}\rangle \ge \frac{1}{2L}\|g(x)\|^2.$$

Hence

$$\Delta_{k+1} \le \Delta_k - \frac{1}{2L}\|g(x_k)\|^2 \le \Delta_k - \frac{\Delta_{k+1}^2}{2Lr_0^2},$$

and therefore,

$$\Delta_k \le \frac{\Delta_0(2Lr_0^2 + \Delta_0)}{2Lr_0^2 + \Delta_0 + k}.$$

□

The main steps above are consequences of the following lemma.

Lemma 5 *Assume $x \in E, T_Q(x), g(x)$ are as above. Then for any $y \in Q$*

$$f(y) \ge f(T_Q(x)) + \langle g(x), y - x\rangle + \frac{1}{2L}\|g(x)\|^2.$$

(for $y = x \Rightarrow f(x) \ge f(T_Q(x)) + \frac{1}{2L}\|g(x)\|^2$

Optimal gradient scheme

Assume d is a *prox-function* of Q:

- d is strongly convex on Q with modulus $\rho > 0$
- $\min\{d(x) : x \in Q\} = 0$
- $\min\{d(x) - \langle g, x\rangle : x \in Q\}$ is easily computable

Idea: construct $\{x_k\}, \{y_k\} \subseteq Q$ such that

$$\frac{(k+1)(k+2)}{4} f(y_k) \le \min_x \left\{ \frac{L}{\rho} d(x) + \sum_{i=0}^{k} \frac{i+1}{2} \left(f(x_i) + \langle \nabla f(x_i), x - x_i\rangle\right) : x \in Q \right\} \quad (1)$$

Theorem 6 *If $\{x_k\}, \{y_k\}$ satisfy (1) then*

$$f(y_k) - \bar{f} \le \frac{4Ld(\bar{x})}{\rho(k+1)(k+2)}$$

Proof. Since $f(x_i) + \langle \nabla f(x_i), x - x_i\rangle \le f(x)$ for $i = 1, \ldots, k$ from (1) we get

$$\begin{aligned}\frac{(k+1)(k+2)}{4} f(y_k) &\le \min\left\{ \frac{L}{\rho} d(x) + \frac{(k+1)(k+2)}{4} f(x) : x \in Q \right\} \\ &\le \frac{L}{\rho} d(\bar{x}) + \frac{(k+1)(k+2)}{4} f(\bar{x})\end{aligned}$$

□

Remarks

- If we can generate $\{x_k\}, \{y_k\}$ that satisfy (1), then we get an algorithm with speed of convergence $O(1/k^2)$
- Complexity to get within ϵ-approximate solution: $O(1/\sqrt{\epsilon})$
- This would be an *optimal* complexity result: for any "black-box" gradient-based algorithm there are smooth problems that require $\Omega(1/\sqrt{\epsilon})$ iterations to find an ϵ-approximate solution.

Algorithm 7

- *Set* $x_0 := \operatorname{argmin}\{d(x) : x \in Q\}$, $y_0 := T_Q(x_0)$
- *For* $k = 0, 1, \ldots$
  $z_k := \operatorname{argmin}\left\{\frac{L}{\rho}d(x) + \sum_{i=0}^{k} \frac{i+1}{2}\left(f(x_i) + \langle \nabla f(x_i), x - x_i\rangle\right) : x \in Q\right\}$
  $x_{k+1} := \frac{2}{k+3}z_k + \frac{k+1}{k+3}y_k$
  $y_{k+1} := T_Q(x_{k+1})$

Theorem 8 *The sequences* $\{x_k\}, \{y_k\}$ *generated by Algorithm 7 satisfy* (1).

Proof. Will use the following claim:

Claim: If d is strongly convex on Q with modulus ρ and $x_0 = \operatorname{argmin}\{d(x) : x \in Q\}$ then

$$d(x) \geq d(x_0) + \frac{\rho}{2}\|x - x_0\|^2, \quad \text{for all } \ y \in Q.$$

By construction, for $x \in Q$ we have

$$f(T_Q(x)) \leq f(x) + \min_y \left\{\langle \nabla f(x), y - x\rangle + \frac{1}{2}L\|y - x\|^2 : y \in Q\right\}.$$

Since $d(y) \geq \frac{\rho}{2}\|y - x_0\|^2$, we have

$$f(T_Q(x_0)) \leq \min_y \left\{f(x_0) + \langle \nabla f(x_0), y - x_0\rangle + \frac{L}{\rho}d(y) : y \in Q\right\}.$$

Thus (1) holds for $k = 0$. Next proceed by induction on k.

Put $A_k := \frac{(k+1)(k+2)}{4}$, $\alpha_k := \frac{k+1}{2}$, $\tau_k := \frac{2}{k+3}$, and

$$\psi_k := \min_x \left\{\frac{L}{\rho}d(x) + \sum_{i=0}^{k} \frac{i+1}{2}\left(f(x_i) + \langle \nabla f(x_i), x - x_i\rangle\right) : x \in Q\right\}.$$

Want to show that if $A_k f(y_k) \leq \psi_k$ then $A_{k+1} f(y_{k+1}) \leq \psi_{k+1}$.

Proceed in two main steps.

Step 1:

$$\psi_{k+1} \geq A_{k+1}\left(f(x_{k+1}) + \min_{x\in Q}\left\{\frac{\tau_k^2}{2}L\|x - z_k\|^2 + \tau_k\langle\nabla f(x_{k+1}), x - z_k\rangle\right\}\right).$$

Step 2:

$$\begin{aligned}
&\min_x\left\{\frac{\tau_k^2}{2}L\|x - z_k\|^2 + \tau_k\langle\nabla f(x_{k+1}), x - z_k\rangle : x \in Q\right\}\\
&\quad\geq \min_y\left\{\frac{1}{2}L\|y - z_k\|^2 + \langle\nabla f(x_{k+1}), y - z_k\rangle : y \in Q\right\}\\
&\quad\geq f(y_{k+1}) - f(x_{k+1}).
\end{aligned}$$

Details of step 1:

$$\begin{aligned}
\psi_{k+1} &\geq \min_x\left\{\psi_k + \frac{L}{2}\|x - z_k\|^2 + \alpha_{k+1}\left(f(x_{k+1}) + \langle\nabla f(x_{k+1}), x - x_{k+1}\rangle\right) : x \in Q\right\}\\
&\geq A_{k+1}\left(f(x_{k+1}) + \min_x\left\{\frac{\tau_k^2}{2}L\|x - z_k\|^2 + \tau_k\langle\nabla f(x_{k+1}), x - z_k\rangle\right\} : x \in Q\right).
\end{aligned}$$

The first inequality holds because d is strongly convex on Q with modulus ρ.

The second inequality follows from $A_{k+1}\tau_k = \alpha_{k+1} \leq 1/\tau_k$ and

$$\begin{aligned}
&\psi_k + \alpha_{k+1}\left(f(x_{k+1}) + \langle\nabla f(x_{k+1}), x - x_{k+1}\rangle\right)\\
&\geq A_k f(y_k) + \alpha_{k+1}\left(f(x_{k+1}) + \langle\nabla f(x_{k+1}), x - x_{k+1}\rangle\right)\\
&\geq A_k\left(f(x_{k+1}) + \langle\nabla f(x_{k+1}), y_k - x_{k+1}\rangle\right)\\
&\quad\alpha_{k+1}\left(f(x_{k+1}) + \langle\nabla f(x_{k+1}), x - x_{k+1}\rangle\right)\\
&= A_{k+1}f(x_{k+1}) + \alpha_{k+1}\langle\nabla f(x_{k+1}), x - z_k\rangle.
\end{aligned}$$

use

$\psi_k \geq A_k f(y_k)$

Details of step 2:

$$\begin{aligned}
&\min_x\left\{\frac{\tau_k^2}{2}L\|x - z_k\|^2 + \tau_k\langle\nabla f(x_{k+1}), x - z_k\rangle : x \in Q\right\}\\
&\quad= \min_y\left\{\frac{1}{2}L\|y - x_{k+1}\|^2 + \langle\nabla f(x_{k+1}), y - x_{k+1}\rangle : y \in \tau_k Q + (1-\tau_k)y_k\right\}\\
&\quad\geq \min_y\left\{\frac{1}{2}L\|y - x_{k+1}\|^2 + \langle\nabla f(x_{k+1}), y - x_{k+1}\rangle : y \in Q\right\}\\
&\quad\geq f(y_{k+1}) - f(x_{k+1}).
\end{aligned}$$

The first equality follows from

$$\begin{aligned}
\tau_k(x - z_k) &= \tau_k x + (1-\tau_k)y_k - (1-\tau_k)y_k - \tau_k z_k\\
&= \tau_k x + (1-\tau_k)y_k + x_{k+1}.
\end{aligned}$$

□

Q is convex

$\tau_k Q + (1-\tau_k)y_k \subseteq Q$

$= \frac{L}{2}\|y_{k+1} - x_{k+1}\|^2 + \langle\nabla f(x_{k+1}), y_{k+1} - x_{k+1}\rangle$

Simply-constrained non-smooth optimization

$$\min\{f(x) : x \in Q\}$$

Assume

- $Q \subseteq E$ is a "simple" compact convex set
- f is convex and Lipschitz continuous with constant M
- f is not necessarily smooth
- At each $x \in Q$ can compute some $g \in \partial f(x)$

Subgradient scheme

Notation: Let $\pi_Q : E \to Q$ be the projection map

$$x \mapsto \text{argmin}\,\{\|y - x\| : y \in Q\}$$

Algorithm 9

- Pick $x_0 \in Q$
- For $k = 0, 1, \ldots$
  Compute $g_k \in \partial f(x_k)$
  $x_{k+1} := \pi_Q\left(x_k - h_k \frac{g_k}{\|g_k\|}\right)$

Theorem 10 *If $h_k = h > 0$ then*

$$f(x_k) - \bar{f} \leq \frac{M\left(\|x_0 - \bar{x}\|^2 + \sum_{j=0}^{k} h_j^2\right)}{\sum_{j=0}^{k} h_j}.$$

Let $D := \max\{\|y - x\| : x, y \in Q\}$.

Corollary 11 *If $h_k = \frac{D}{\sqrt{N+1}}$ for $k = 0, 1, \ldots, N$ then*

$$f(x_N) - \bar{f} \leq \frac{MD}{\sqrt{N+1}}$$

Remarks

- Complexity to get within ϵ of optimal value is $O(1/\epsilon^2)$
- This is an *optimal* complexity result: for any "black-box" subgradient-based algorithm there are non-smooth problems that require $\Omega(1/\epsilon^2)$ iterations to find an ϵ-approximate solution.

References for today's material

- Y. Nesterov, "Introductory Lectures on Convex Optimization," Kluwer Academic Publishers, 2004.
- Y. Nesterov, "Smooth minimization of non-smooth functions," Math Program. 103 (2005), 127–152.