Smoothing for non-smooth optimization, lecture 3

Last time:

- Standard $O(1/\epsilon)$ & optimal $O(1/\sqrt{\epsilon})$ gradient schemes
- Optimal $O(1/\epsilon^2)$ subgradient scheme (non-smooth)

Today: *Smoothing*

- A "fixed" smoothing approach
- Excessive gap technique

Slides available online at

`http://www.andrew.cmu.edu/user/jfp/smooth/`

References for today's material

- Y. Nesterov, "Smooth minimization of non-smooth functions," Math Program. 103 (2005), 127–152.
- Y. Nesterov, "Excessive gap technique in non-smooth convex minimization," SIAM J. Opt. 16 (2005), 235–249.

Min-max problems with simple structure

Consider

$$\min_{x\in Q_1}\max_{y\in Q_2}\left\{\hat{f}(x)-\hat{\phi}(y)+\langle Ax,y\rangle\right\}.$$

Assume

- E_1, E_2 are finite dimensional Euclidean spaces, $A\in L(E_1,E_2)$
- $Q_i\subseteq E_i$ are simple compact convex sets.
- $\hat{f}$ and $\hat{\phi}$ are convex and differentiable with Lipschitz gradients

Fenchel duality yields

$$\min_{x\in Q_1}\max_{y\in Q_2}\left\{\hat{f}(x)-\hat{\phi}(y)+\langle Ax,y\rangle\right\}=\max_{y\in Q_2}\min_{x\in Q_1}\left\{\hat{f}(x)-\hat{\phi}(y)+\langle Ax,y\rangle\right\}.$$

Can write these problems as

$$\min\{f(x):x\in Q_1\}=\max\{\phi(y):y\in Q_2\}$$

for

$$f(x)=\hat{f}(x)+\max\left\{\langle Ax,y\rangle-\hat{\phi}(y):y\in Q_2\right\}$$

and

$$\phi(y)=-\hat{\phi}(y)+\min\left\{\langle Ax,y\rangle+\hat{f}(x):x\in Q_1\right\}$$

Notice:

- For all $x \in Q_1, y \in Q_2$ we have $f(x) \geq \phi(y)$
- $\bar{x} \in Q_1, \bar{y} \in Q_2$ are optimal solutions if and only if $f(\bar{x}) = \phi(\bar{y})$
- f, ϕ are convex and concave respectively but non-smooth

$f(x) := \hat{f}(x) + \max\{\langle Ax, y\rangle - \hat{\phi}(y) : y \in Q_2\}$

A "fixed" smoothing approach

Assume d_2 is a prox-function of the set Q_2. Given $\mu > 0$ consider

$$f_\mu(x) := \hat{f}(x) + \max\left\{\langle Ax, y\rangle - \hat{\phi}(y) - \mu d_2(y) : y \in Q_2\right\}.$$

Let $y_\mu(x)$ be the unique maximizer in this max-problem.

Theorem 1

(i) f_μ is differentiable with

$$\nabla f_\mu(x) = \nabla \hat{f}(x) + A^* y_\mu(x)$$

(ii) ∇f_μ is Lipschitz continuous with constant

$$L_{f_\mu} = L_{\hat{f}} + \frac{1}{\mu\rho_2}\|A\|.$$

Intuitively: $f_\mu \approx f$ for μ small.

Proposition 2 *Let $D_2 := \max\{d_2(y) : y \in Q_2\}$. Then for $\mu > 0$*

$$0 \leq f(x) - f_\mu(x) \leq \mu D_2.$$

Idea: To find an approximate solution to

$$\min_{x \in Q_1} f(x)$$

proceed as follows:

- Pick $\mu > 0$
- Apply optimal gradient scheme to $\min\{f_\mu(x) : x \in Q_1\}$

Recall optimal gradient scheme for $\min\{f_\mu(x) : x \in Q_1\}$:

Algorithm 3

- *Set* $x_0 := \operatorname{argmin}\{d_1(x) : x \in Q_1\}$, $u_0 := T_{Q_1}(x_0)$
- *For* $k = 0, 1, \ldots$
  $z_k := \operatorname{argmin}\left\{\frac{L_\mu}{\rho} d(x) + \sum_{i=0}^{k} \frac{i+1}{2}\left(f_\mu(x_i) + \langle \nabla f_\mu(x_i), x - x_i\rangle\right) : x \in Q_1\right\}$
  $x_{k+1} := \frac{2}{k+3} z_k + \frac{k+1}{k+3} u_k$
  $u_{k+1} := T_{Q_1}(x_{k+1})$

Theorem 1 is an immediate consequence of the following lemma.

Lemma 4 *Assume $Q \subseteq E$ is a convex compact set and $\theta : Q \to \mathbb{R} \cup \{+\infty\}$ is closed and strongly convex with modulus ρ. Consider the conjugate*

$$\theta^*(s) = \max_{y \in Q} \{\langle s, y\rangle - \theta(y)\}.$$

Then

- *The maximizer $y(s)$ above is unique* ✓
- $\nabla\theta^*(s) = y(s)$ ✓
- *For all $s, u \in E$ we have $\|y(s) - y(u)\| \le \frac{1}{\rho}\|s - u\|$.* ✓

$\|\nabla\theta^*(s) - \nabla\theta^*(u)\|$

Theorem 5 *If*

$$\mu := \frac{2\|A\|}{N+1}\sqrt{\frac{D_1}{\rho_1\rho_2 D_2}}$$

then after N iterations the points

$$\widehat{x} = u_N,\ \widehat{y} := \sum_{i=0}^{N} \frac{2(i+1)}{(N+1)(N+2)} y_\mu(x_i)$$

satisfy

$$0 \le f(\widehat{x}) - \phi(\widehat{y}) \le \frac{4\|A\|}{N+1} \cdot \sqrt{\frac{D_1 D_2}{\rho_1\rho_2}} + \frac{4L_{\hat{f}} D_1}{\rho_1 (N+1)^2}.$$

Note: Need $O(1/\epsilon)$ to get an ϵ-approximate solution. It beats the lower complexity bound for black-box subgradient schemes.

Proof.

Step 1:

$$f_\mu(\widehat{x}) \le \frac{4L_{f_\mu} D_1}{\rho_1 (N+1)^2} + \frac{2}{(N+1)(N+2)} \min_{x \in Q_1} \left\{ \sum_{i=0}^{N} (i+1)\left(f_\mu(x_i) + \langle \nabla f_\mu(x_i), x - x_i\rangle\right) \right\}.$$

Step 2:

$$\min_{x \in Q_1} \left\{ \sum_{i=0}^{N} (i+1)\left(f_\mu(x_i) + \langle \nabla f_\mu(x_i), x - x_i\rangle\right) \right\} \le \frac{(N+1)(N+2)}{2} \phi(\widehat{y}).$$

Thus

$$f(\widehat{x}) \le f_\mu(\widehat{x}) + \mu D_2 \le \frac{4L_{f_\mu} D_1}{\rho_1 (N+1)^2} + \phi(\widehat{y}) + \mu D_2.$$

So

$$0 \le f(\widehat{x}) - \phi(\widehat{y}) \le \frac{4L_{\hat{f}} D_1}{\rho_1 (N+1)^2} + \frac{4\|A\|^2 D_1}{\mu\rho_1\rho_2 (N+1)^2} + \mu D_2.$$

The minimum of the last expression (as a function of μ) is

$$\frac{4\|A\|}{N+1} \cdot \sqrt{\frac{D_1 D_2}{\rho_1\rho_2}} + \frac{4L_{\hat{f}} D_1}{\rho_1 (N+1)^2},$$

and is attained at $\mu = \frac{2\|A\|}{N+1}\sqrt{\frac{D_1}{\rho_1\rho_2 D_2}}$.

Details of Step 1: By Theorem 6 (lecture 2)

Details of Step 2:

By construction, for $x \in Q_1$

$$f_\mu(x) = \hat{f}(x) + \langle Ax, y_\mu(x)\rangle - \hat{\phi}(y_\mu(x)) - \mu d_2(y_\mu(x)),$$

and by Theorem 1

$$\langle \nabla f_\mu(x) - \nabla \hat{f}(x), x\rangle = \langle A^* y_\mu(x), x\rangle = \langle Ax, y_\mu(x)\rangle.$$

Hence

$$\begin{aligned} f_\mu(x) - \hat{f}(x) - \langle \nabla f_\mu(x) - \nabla \hat{f}(x), x\rangle &= -\hat{\phi}(y_\mu(x)) - \mu d_2(y_\mu(x)) \\ &\leq -\hat{\phi}(y_\mu(x)). \end{aligned}$$

Thus

$$\begin{aligned} &\sum_{i=0}^{N}(i+1)\left(f_\mu(x_i) + \langle \nabla f_\mu(x_i), x - x_i\rangle\right) \\ &\quad\leq \sum_{i=0}^{N}(i+1)\left(f_\mu(x_i) - \hat{f}(x_i) + \langle \nabla f_\mu(x_i) - \nabla \hat{f}(x_i), x - x_i\rangle\right) \\ &\quad\quad + \frac{(N+1)(N+2)}{2}\hat{f}(x) \\ &\quad\leq -\sum_{i=0}^{N}(i+1)\hat{\phi}(y_\mu(x_i)) + \frac{(N+1)(N+2)}{2}\left(\hat{f}(x) + \langle A^*\hat{y}, x\rangle\right) \\ &\quad\leq \frac{(N+1)(N+2)}{2}\left(-\hat{\phi}(\hat{y}) + \hat{f}(x) + \langle A^*\hat{y}, x\rangle\right) \end{aligned}$$

□

Remarks

- The approach is mostly "primal".
- The smoothing parameter μ needs to be fixed up front.

Excessive gap technique

Idea: Smooth *both* f and ϕ.

Given $\mu_1, \mu_2 > 0$ consider

$$f_{\mu_2}(x) := \hat{f}(x) + \max\left\{\langle Ax, y\rangle - \hat{\phi}(y) - \mu_2 d_2(y) : y \in Q_2\right\}.$$

and

$$\phi_{\mu_1}(y) := -\hat{\phi}(y) + \min\left\{\langle Ax, y\rangle + \hat{f}(x) + \mu_1 d_1(x) : x \in Q_1\right\}.$$

Recall: for all $x \in Q_1, y \in Q_2$

$$f(x) \geq \phi(y).$$

Since $f_{\mu_2}(x) \le f(x)$ and $\phi_{\mu_1}(y) \ge \phi(y)$, it is conceivable to have the following *excessive gap condition*

$$f_{\mu_2}(x) \le \phi_{\mu_1}(y) \qquad \text{(EGC)}$$

Intuitively: $f \approx f_{\mu_2}$ and $\phi \approx \phi_{\mu_1}$ for μ_1, μ_2 small.

Proposition 6 *Assume $\mu_1, \mu_2 > 0$ and $x \in Q_1, y \in Q_2$ satisfy* (EGC). *Then*

$$0 \le f(x) - \phi(y) \le \mu_1 D_1 + \mu_2 D_2$$

In particular $f(x), \phi(y)$ are within $\mu_1 D_1 + \mu_2 D_2$ of the optimal value.

Idea: To find an approximate solution to

$$\min_{x \in Q_1} f(x) = \max_{y \in Q_2} \phi(y)$$

Proceed as follows: generate a sequence $(\mu_{1,k}, \mu_{2,k}, x_k, y_k)$ such that

- Each $(\mu_1^k, \mu_2^k, x_k, y_k)$ satisfies (EGC)
- $\mu_1^k, \mu_2^k \downarrow 0$.

Need:

- initial point $(\mu_1^0, \mu_2^0, x_0, y_0)$
- update that preserves (EGC) while reducing μ_1, μ_2

Bregman projection

Assume d is a differentiable prox-function of $Q \subseteq E$.

For $x, z \in Q$ define the *Bregman distance* between x and z as

$$\xi(z, x) := d(x) - d(z) - \langle \nabla d(z), x - z \rangle$$

Notice: Since d is strongly convex with modulus ρ,

$$\xi(z, x) \ge \frac{\rho}{2} \|x - z\|^2$$

Define the *Bregman projection* of $g \in E$ onto Q as

$$V(z, g) := \operatorname{argmin} \{\langle g, x - z \rangle + \xi(z, x)\}$$

Back to f_{μ_2}, ϕ_{μ_1}: As before, we have

$$\nabla f_{\mu_2}(x) = \nabla \hat{f}(x) + A^* y_{\mu_2}(x), \ \nabla \phi_{\mu_1}(y) = -\nabla \hat{\phi}(y) + A x_{\mu_1}(y),$$

Furthermore, $\nabla f, \nabla \phi$ are Lipschitz continuous with constants

$$L_{f_{\mu_2}} = L_{\hat{f}} + \frac{1}{\mu_2 \rho_2} \|A\|, \ L_{\phi_{\mu_1}} = L_{\hat{\phi}} + \frac{1}{\mu_1 \rho_1} \|A\|.$$

For simplicity assume $L_{\hat{f}} = L_{\hat{\phi}} = 0$ in the sequel.

Initial point:

Theorem 7 *Assume* $\mu_2 > 0$ *and* $\widehat{x} := \operatorname{argmin}\{d_1(x) : x \in Q_1\}$. *Set* $\mu_1 = \frac{\|A\|^2}{\mu_2 \rho_1 \rho_2}$ *and*

$$x = V_1\left(\widehat{x}, \frac{1}{\mu_1}\nabla f_{\mu_2}(\widehat{x})\right), \quad y = y_{\mu_2}(\widehat{x}).$$

Then (μ_1, μ_2, x, y) *satisfies* (EGC).

Update:

Theorem 8 *Assume* (μ_1, μ_2, x, y) *satisfies* (EGC) *and* $\tau \in (0, 1)$ *is such that*

$$\frac{\tau^2}{1-\tau} \leq \frac{\mu_1 \mu_2 \rho_1 \rho_2}{\|A\|^2}.$$

Set

- $\widehat{x} := (1-\tau)x + \tau x_{\mu_1}(y)$
- $y^+ := (1-\tau)y + \tau y_{\mu_2}(\widehat{x})$
- $\tilde{x} := V_1\left(x_{\mu_1}(y), \frac{\tau}{(1-\tau)\mu_1}\nabla f_{\mu_2}(\widehat{x})\right)$
- $x^+ := (1-\tau)x + \tau\tilde{x}$
- $\mu_1^+ := (1-\tau)\mu_1$

Then $(\mu_1^+, \mu_2, x^+, y^+)$ *satisfies* (EGC).

EGT algorithm

Ingredients: subroutines `initial` and `shrink`

`initial`(A, d_1, d_2)

1. $\mu_1^0 := 2\|A\|\sqrt{\frac{D_2}{\rho_1\rho_2 D_1}}$, $\mu_2^0 := \|A\|\sqrt{\frac{D_1}{\rho_1\rho_2 D_2}}$
2. $\widehat{\mathbf{x}} := \operatorname{argmin}\{d_1(\mathbf{x}) : \mathbf{x} \in Q_1\}$
3. $\mathbf{x}^0 := V_1\left(\widehat{\mathbf{x}}, \frac{2}{\mu_1^0}\nabla f_{\mu_2}(\widehat{\mathbf{x}})\right)$
4. $\mathbf{y}^0 := y_{\mu_2}(\widehat{\mathbf{x}})$
5. `return` $(\mu_1^0, \mu_2^0, \mathbf{x}^0, \mathbf{y}^0)$

`shrink`$(A, \mu_1, \mu_2, \tau, \mathbf{x}, \mathbf{y}, d_1, d_2)$

1. $\widehat{x} := (1-\tau)x + \tau x_{\mu_1}(y)$
2. $y^+ := (1-\tau)y + \tau y_{\mu_2}(\widehat{x})$
3. $\tilde{x} := V_1\left(x_{\mu_1}(y), \frac{\tau}{(1-\tau)\mu_1}\nabla f_{\mu_2}(\widehat{x})\right)$
4. $x^+ := (1-\tau)x + \tau\tilde{x}$
5. $\mu_1^+ := (1-\tau)\mu_1$
6. `return` $(\mu_1^+, \mathbf{x}^+, \mathbf{y}^+)$

Algorithm 9

1. $(\mu_1^0, \mu_2^0, \mathbf{x}^0, \mathbf{y}^0) = \texttt{initial}(A, d_1, d_2)$
2. `For` $k = 0, 1, \ldots$
   (a) $\tau := \frac{2}{k+3}$
   (b) `If` k `is even: // shrink` μ_1
   $(\mu_1^{k+1}, \mathbf{x}^{k+1}, \mathbf{y}^{k+1}) := \texttt{shrink}(A, \mu_1^k, \mu_2^k, \tau, \mathbf{x}^k, \mathbf{y}^k, d_1, d_2)$
   $\mu_2^{k+1} := \mu_2^k$
   (c) `If` k `is odd: // shrink` μ_2
   $(\mu_2^{k+1}, \mathbf{y}^{k+1}, \mathbf{x}^{k+1}) := \texttt{shrink}(-A^*, \mu_2^k, \mu_1^k, \tau, \mathbf{y}^k, \mathbf{x}^k, d_2, d_1)$
   $\mu_1^{k+1} := \mu_1^k$

Theorem 10 *Each* $(\mu_1^k, \mu_2^k, \mathbf{x}^k, \mathbf{y}^k)$ *satisfies* (EGC) *and*

$$0 \le f(\mathbf{x}^k) - \phi(\mathbf{y}^k) \le \frac{4\|A\|}{k+1}\sqrt{\frac{D_1 D_2}{\rho_1 \rho_2}}$$

Proof. *This follows from*

$$\mu_1^k = \|A\|\frac{2}{k+1}\sqrt{\frac{D_2}{\rho_1 \rho_2 D_1}},\ \mu_2^k = \|A\|\frac{2}{k+2}\sqrt{\frac{D_1}{\rho_1 \rho_2 D_2}},\ k \text{ even} \quad (1)$$

and

$$\mu_1^k = \|A\|\frac{2}{k+2}\sqrt{\frac{D_2}{\rho_1 \rho_2 D_1}},\ \mu_2^k = \|A\|\frac{2}{k+1}\sqrt{\frac{D_1}{\rho_1 \rho_2 D_2}},\ k \text{ odd} \quad (2)$$

References for today's material

- Y. Nesterov, "Smooth minimization of non-smooth functions," Math Program. 103 (2005), 127–152.
- Y. Nesterov, "Excessive gap technique in non-smooth convex minimization," SIAM J. Opt. 16 (2005), 235–249.