Accelerated Proximal Gradient Methods for Convex Optimization

Paul Tseng
Mathematics, University of Washington
Seattle

Optimization Seminar, Univ. Washington
October 7, 2008
Talk Outline

• A Convex Optimization Problem
Talk Outline

- A Convex Optimization Problem
- Proximal Gradient Method
Talk Outline

- A Convex Optimization Problem
- Proximal Gradient Method
- Accelerated Proximal Gradient Method I
- Accelerated Proximal Gradient Method II
Talk Outline

- A Convex Optimization Problem
- Proximal Gradient Method
- Accelerated Proximal Gradient Method I
- Accelerated Proximal Gradient Method II
- Example: Matrix Game
Talk Outline

- A Convex Optimization Problem
- Proximal Gradient Method
- Accelerated Proximal Gradient Method I
- Accelerated Proximal Gradient Method II
- Example: Matrix Game
- Conclusions & Extensions
A Convex Optimization Problem

$$\min_{x \in \mathcal{E}} f^P(x) := f(x) + P(x)$$

$\mathcal{E}$ is a real linear space with norm $\| \cdot \|$.  
$\mathcal{E}^*$ is the dual space of cont. linear functionals on $\mathcal{E}$, with dual norm  
$\| x^* \|_* = \sup_{\| x \| \leq 1} \langle x^*, x \rangle$. 

$P : \mathcal{E} \to (-\infty, \infty]$ is proper, convex, lsc (and “simple”).

$f : \mathcal{E} \to \mathbb{R}$ is convex diff.  
$\| \nabla f(x) - \nabla f(y) \|_* \leq L \| x - y \| \ \forall x, y \in \text{dom}P \ (L \geq 0)$. 
A Convex Optimization Problem

\[
\min_{x \in \mathcal{E}} f^P(x) := f(x) + P(x)
\]

\(\mathcal{E}\) is a real linear space with norm \(\| \cdot \|\).

\(\mathcal{E}^*\) is the dual space of cont. linear functionals on \(\mathcal{E}\), with dual norm 
\[\|x^*\|_* = \sup_{\|x\| \leq 1} \langle x^*, x \rangle.\]

\(P : \mathcal{E} \to (-\infty, \infty]\) is proper, convex, lsc (and “simple”).

\(f : \mathcal{E} \to \mathbb{R}\) is convex diff. 
\[\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\| \quad \forall x, y \in \text{dom} P \quad (L \geq 0).\]

**Constrained case:** \(P = \delta_X\) with \(X \subseteq \mathcal{E}\) nonempty, closed, convex.

\[
\delta_X(x) = \begin{cases} 
0 & \text{if } x \in X \\
\infty & \text{else}
\end{cases}
\]
Examples:

- \( \mathcal{E} = \mathbb{R}^n, \ P(x) = \|x\|_1, \ f(x) = \|Ax - b\|_2^2 \) \hspace{1cm} \text{Basis Pursuit/Lasso}

- \( \mathcal{E} = \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_N}, \ P(x) = w_1 \|x_1\|_2 + \cdots + w_N \|x_N\|_2 \ (w_j > 0), \ f(x) = g(Ax) \text{ with } g(y) = \sum_{i=1}^{m} \ln(1 + e^{y_i}) - b_i y_i \) \hspace{1cm} \text{group Lasso}

- \( \mathcal{E} = \mathbb{R}^n, \ P \equiv \delta_X \text{ with } X = \{x \mid x \geq 0, x_1 + \cdots + x_n = 1\}, \ f(x) = g^*(Ax) \text{ with } g(y) = \begin{cases} \sum_{i=1}^{m} y_i \ln y_i & \text{if } y \geq 0, \ y_1 + \cdots + y_m = 1 \\ \infty & \text{else} \end{cases} \) \hspace{1cm} \text{matrix game}

- \( \mathcal{E} = S^n, \ P \equiv \delta_X \text{ with } X = \{x \mid |x_{ij}| \leq \rho \ \forall i, j\}, \ f(x) = g^*(x + s) \text{ with } g(y) = \begin{cases} -\ln \det y & \text{if } \alpha I \preceq y \preceq \beta I \ (\rho, \alpha, \beta > 0) \\ \infty & \text{else} \end{cases} \) \hspace{1cm} \text{covariance selection}
How to solve this (nonsmooth) convex optimization problem? In applications, \( m \) and \( n \) are large (\( m, n \geq 1000 \)), \( A \) may be dense.

2nd-order methods (Newton, interior-point)? Few iterations, but each iteration can be too expensive (e.g., \( O(n^3) \) ops).

1st-order methods (gradient)? Each iteration is cheap (by using suitable “prox function”), but often too many iterations. Accelerate convergence by interpolation Nesterov.
Proximal Gradient Method

Let

\[ \ell(x; y) := f(y) + \langle \nabla f(y), x - y \rangle + P(x) \]
\[ D(x, y) := h(x) - h(y) - \langle \nabla h(y), x - y \rangle, \]

with \( h : \mathcal{E} \to (-\infty, \infty] \) strictly convex, differentiable on \( X_h \supseteq \text{int}(\text{dom} P) \), and

\[ D(x, y) \geq \frac{1}{2} \| x - y \|^2 \quad \forall \, x \in \text{dom} P, \, y \in X_h. \]
**Proximal Gradient Method**

Let

\[
\ell(x; y) := f(y) + \langle \nabla f(y), x - y \rangle + P(x)
\]

\[
D(x, y) := h(x) - h(y) - \langle \nabla h(y), x - y \rangle,
\]

with \( h : \mathcal{E} \to (-\infty, \infty] \) strictly convex, differentiable on \( X_h \supseteq \text{int}(\text{dom}P) \), and

\[
D(x, y) \geq \frac{1}{2} \|x - y\|^2 \quad \forall x \in \text{dom}P, y \in X_h.
\]

For \( k = 0, 1, \ldots, \)

\[
x_{k+1} = \arg \min_x \{ \ell(x; x_k) + LD(x, x_k) \}
\]

with \( x_0 \in \text{dom}P \). Assume \( x_k \in X_h \ \forall k \).

Special cases: steepest descent, gradient-projection Goldstein, Levitin, Polyak, ..., mirror-descent Yudin, Nemirovski, iterative thresholding Daubechies et al., ...
For the earlier examples, $x_{k+1}$ has closed form when $h$ is chosen suitably:

- $\mathcal{E} = \mathbb{R}^n$, $P(x) = \|x\|_1$, $h(x) = \|x\|_2^2/2$.

- $\mathcal{E} = \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_N}$, $P(x) = w_1\|x_1\|_2 + \cdots + w_N\|x_N\|_2$ ($w_j > 0$), $h(x) = \|x\|_2^2/2$.

- $\mathcal{E} = \mathbb{R}^n$, $P \equiv \delta_X$ with $X = \{x \mid x \geq 0, x_1 + \cdots + x_n = 1\}$, $h(x) = \sum_{j=1}^{n} x_j \ln x_j$.

- $\mathcal{E} = S^n$, $P \equiv \delta_X$ with $X = \{x \mid |x_{ij}| \leq \rho \ \forall i, j\}$, $h(x) = \|x\|_F^2/2$. 
Fact 1: \[ f^P(x) \geq \ell(x; y) \geq f^P(x) - \frac{L}{2} \| x - y \|^2 \quad \forall x, y \in \text{dom}P. \]

Fact 2: For any proper convex lsc \( \psi : \mathcal{E} \to (-\infty, \infty] \) and \( z \in X_h \), let

\[ z_+ = \arg \min_{x} \{ \psi(x) + D(x, z) \}. \]

If \( z_+ \in X_h \), then

\[ \psi(z_+) + D(z_+, z) \leq \psi(x) + D(x, z) - D(x, z_+) \quad \forall x \in \text{dom}P. \]
Prop. 1: For any $x \in \text{dom} P$, 

$$\min\{e_1, \ldots, e_k\} \leq \frac{LD(x, x_0)}{k}, \quad k = 1, 2, \ldots$$

with $e_k := f^P(x_k) - f^P(x)$. 
Prop. 1: For any \( x \in \text{dom}P \),

\[
\min\{e_1, \ldots, e_k\} \leq \frac{LD(x,x_0)}{k}, \quad k = 1, 2, \ldots
\]

with \( e_k := f_P(x_k) - f_P(x) \).

Proof:

\[
\begin{align*}
    f_P(x_{k+1}) & \leq \ell(x_{k+1}; x_k) + \frac{L}{2}\|x_{k+1} - x_k\|^2 \quad \text{Fact 1} \\
        & \leq \ell(x_{k+1}; x_k) + LD(x_{k+1}, x_k) \\
        & \leq \ell(x; x_k) + LD(x, x_k) - LD(x, x_{k+1}) \quad \text{Fact 2} \\
        & \leq f_P(x) + LD(x, x_k) - LD(x, x_{k+1}), \quad \text{Fact 1}
\end{align*}
\]

so

\[
\begin{align*}
    0 & \leq LD(x, x_{k+1}) \leq LD(x, x_k) - e_{k+1} \\
        & \leq LD(x, x_0) - (e_1 + \cdots + e_{k+1}) \\
        & \leq LD(x, x_0) - (k + 1) \min\{e_1, \ldots, e_{k+1}\}
\end{align*}
\]
We will improve the global convergence rate by interpolation.

**Idea:** At iteration $k$, use a stepsize of $O(k/L)$ instead of $1/L$ and backtrack towards $x_k$. 
Accelerated Proximal Gradient Method I

For \( k = 0, 1, \ldots, \)

\[
\begin{align*}
y_k &= (1 - \theta_k)x_k + \theta_k z_k \\
z_{k+1} &= \arg \min_x \{ \ell(x; y_k) + \theta_k L D(x, z_k) \} \\
x_{k+1} &= (1 - \theta_k)x_k + \theta_k z_{k+1} \\
\frac{1 - \theta_{k+1}}{\theta_{k+1}^2} &\leq \frac{1}{\theta_k^2} \quad (0 < \theta_{k+1} \leq 1)
\end{align*}
\]

with \( \theta_0 = 1, x_0, z_0 \in \text{dom} P \) \cite{nesterov2018lectures, auslender2013first, teboulle2015expanded, lan2015 repetitive, lu2015efficient, monteiro2011first}. Assume \( z_k \in X_h \ \forall k. \)

For example, \( \theta_k = \frac{2}{k + 2} \) or \( \theta_{k+1} = \frac{\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2}{2} \).
Prop. 2: For any $x \in \text{dom} P$,

$$\min\{e_1, \ldots, e_k\} \leq LD(x, z_0) \theta^2_k, \quad k = 1, 2, \ldots$$

with $e_k := f^P(x_k) - f^P(x)$. 
Prop. 2: For any $x \in \text{dom} P$,

$$\min\{e_1, \ldots, e_k\} \leq LD(x, z_0)\theta_k^2, \quad k = 1, 2, \ldots$$

with $e_k := f^P(x_k) - f^P(x)$.

Proof:

\[
\begin{align*}
f^P(x_{k+1}) & \leq \ell(x_{k+1}; y_k) + \frac{L}{2}\|x_{k+1} - y_k\|^2 \quad \text{Fact 1} \\
& = \ell((1 - \theta_k)x_k + \theta_k z_{k+1}; y_k) + \frac{L}{2}\|(1 - \theta_k)x_k + \theta_k z_{k+1} - y_k\|^2 \\
& \leq (1 - \theta_k)\ell(x_k; y_k) + \theta_k\ell(z_{k+1}; y_k) + \frac{L}{2}\theta_k^2\|z_{k+1} - z_k\|^2 \\
& \leq (1 - \theta_k)\ell(x_k; y_k) + \theta_k (\ell(z_{k+1}; y_k) + \theta_k LD(z_{k+1}, z_k)) \\
& \leq (1 - \theta_k)\ell(x_k; y_k) + \theta_k (\ell(x; y_k) + \theta_k LD(x, z_k) - \theta_k LD(x, z_{k+1})) \quad \text{Fact 2} \\
& \leq (1 - \theta_k)f^P(x_k) + \theta_k (f^P(x) + \theta_k LD(x, z_k) - \theta_k LD(x, z_{k+1})) \quad \text{Fact 1}
\end{align*}
\]
so, subtracting by $f^P(x)$ and then dividing by $\theta_k^2$, we have

$$
\frac{1}{\theta_k^2} e_{k+1} \leq \frac{1 - \theta_k}{\theta_k^2} e_k + LD(x; z_k) - LD(x; z_{k+1})
$$

etc.

Thus, global convergence rate improves from $O(1/k)$ to $O(1/k^2)$ with little extra work per iteration!
Comparing PGM with APGM I:

Assume $P \equiv \delta_X$. 
Can also replace $\ell(x; y_k)$ by a certain weighted sum of $\ell(x; y_0), \ell(x; y_1), \ldots, \ell(x; y_k)$.

Then...
Accelerated Proximal Gradient Method II

For \( k = 0, 1, \ldots \),

\[
\begin{align*}
y_k &= (1 - \theta_k)x_k + \theta_k z_k \\
z_{k+1} &= \text{arg min}_{x} \left\{ \sum_{i=0}^{k} \frac{\ell(x; y_i)}{\vartheta_i} + Lh(x) \right\} \\
x_{k+1} &= (1 - \theta_k)x_k + \theta_k z_{k+1} \\
\frac{1 - \theta_{k+1}}{\theta_{k+1}\vartheta_{k+1}} &= \frac{1}{\theta_k\vartheta_k} \quad (\vartheta_{k+1} \geq \theta_{k+1} > 0)
\end{align*}
\]

with \( \vartheta_0 \geq \theta_0 = 1 \), \( x_0 \in \text{dom}P \), and \( z_0 = \text{arg min}_{x \in \text{dom}P} h(x) \) Nesterov, d'Aspremont et al., Lu, ... Assume \( z_k \in X_h \forall k \).

For example, \( \vartheta_k = \frac{2}{k+1}, \theta_k = \frac{2}{k+2} \) or \( \vartheta_{k+1} = \theta_{k+1} = \frac{\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2}{2} \).
Prop. 3: For any $x \in \text{dom} P$, 

$$\min\{e_1, \ldots, e_k\} \leq L(h(x) - h(z_0))\theta_{k-1}\vartheta_{k-1}, \quad k = 1, 2, \ldots$$

with $e_k := f^P(x_k) - f^P(x)$. 
Prop. 3: For any $x \in \text{dom} P$,

$$
\min\{e_1, \ldots, e_k\} \leq L(h(x) - h(z_0))\theta_{k-1}\vartheta_{k-1}, \quad k = 1, 2, \ldots
$$

with $e_k := f^P(x_k) - f^P(x)$.

Proof replaces Fact 2 with:

Fact 3: For any proper convex lsc $\psi : \mathcal{E} \to (-\infty, \infty]$, let

$$
z = \arg \min_x \{\psi(x) + h(x)\}.
$$

If $z \in X_h$, then

$$
\psi(z) + h(z) \leq \psi(x) + h(x) - D(x, z) \quad \forall x \in \text{dom} P.
$$

Advantage? Possibly better performance on compressed sensing and certain conic programs. Lu
Example: Matrix Game

\[
\min_{x \in X} \max_{v \in V} \langle v, Ax \rangle
\]

with \(X\) and \(V\) unit simplices in \(\mathbb{R}^n\) and \(\mathbb{R}^m\), and \(A \in \mathbb{R}^{m \times n}\). Generate \(A_{ij} \sim U[-1, 1]\) with probab. \(p\); otherwise \(A_{ij} = 0\). Nesterov, Nemirovski

Set \(P = \delta_X\) and \(f(x) = g^*(Ax/\mu)\), with \(\mu = \frac{\epsilon}{2 \ln m} (\epsilon > 0)\) and

\[
g(v) = \begin{cases} 
\sum_{i=1}^{m} v_i \ln v_i & \text{if } v \in V \\
\infty & \text{else}
\end{cases}
\]

\((L = \frac{1}{\mu}, \| \cdot \| = 1\text{-norm})\)
Example: Matrix Game

\[
\min_{x \in X} \max_{v \in V} \langle v, Ax \rangle
\]

with \( X \) and \( V \) unit simplices in \( \mathbb{R}^n \) and \( \mathbb{R}^m \), and \( A \in \mathbb{R}^{m \times n} \). Generate \( A_{ij} \sim U[-1, 1] \) with probab. \( p \); otherwise \( A_{ij} = 0 \). Nesterov, Nemirovski

Set \( P \equiv \delta_X \) and \( f(x) = g^*(Ax/\mu) \), with \( \mu = \frac{\epsilon}{2 \ln m} (\epsilon > 0) \) and

\[
g(v) = \begin{cases} 
\sum_{i=1}^{m} v_i \ln v_i & \text{if } v \in V \\
\infty & \text{else}
\end{cases}
\]

\((L = \frac{1}{\mu}, \| \cdot \| = 1\)-norm\)

- Implement PGM, APGM I & II in Matlab, with \( h(x) = \sum_{j=1}^{n} x_j \ln x_j \) and \( L^{\text{init}} = \frac{1}{8\mu} \). Matrix-vector mult. by \( A, A^* \) per iter.

- Initialize \( x_0 = z_0 = (\frac{1}{n}, \ldots, \frac{1}{n}) \). Terminate when

\[
\max_i (Ax_k)_i - \min_j (A^*v_k)_j \leq \epsilon
\]

with \( v_k \in V \) a weighted sum of dual vectors associated with \( x_0, x_1, \ldots, x_k \).
### Table 1: Performance of PGM, APGM I & II for different $n$, $m$, sparsity $p$, and soln accuracy $\epsilon$.

<table>
<thead>
<tr>
<th>$n/m/p$</th>
<th>$\epsilon$</th>
<th>PGM</th>
<th>APGM I</th>
<th>APGM II</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000/100/.01</td>
<td>.001</td>
<td>1082480/1500</td>
<td>3325/5</td>
<td>10510/9</td>
</tr>
<tr>
<td></td>
<td>.0001</td>
<td>–</td>
<td>20635/23</td>
<td>61865/45</td>
</tr>
<tr>
<td>10000/100/.01</td>
<td>.001</td>
<td>–</td>
<td>10005/142</td>
<td>10005/128</td>
</tr>
<tr>
<td>10000/100/.1</td>
<td>.001</td>
<td>–</td>
<td>10005/201</td>
<td>10005/185</td>
</tr>
<tr>
<td>10000/1000/.01</td>
<td>.001</td>
<td>–</td>
<td>10005/202</td>
<td>10005/191</td>
</tr>
<tr>
<td>10000/1000/.1</td>
<td>.001</td>
<td>–</td>
<td>10005/706</td>
<td>10005/695</td>
</tr>
</tbody>
</table>
Conclusions & Extensions

Conclusions & Extensions


2. Application to matrix completion, where \( \mathcal{E} = \mathbb{R}^{m \times n} \) and \( P(x) = \|\sigma(x)\|_1 \)? Or to total-variation image restoration (joint work with Steve Wright)?
Conclusions & Extensions


2. Application to matrix completion, where $\mathcal{E} = \mathbb{R}^{m \times n}$ and $P(x) = \|\sigma(x)\|_1$? Or to total-variation image restoration (joint work with Steve Wright)?

3. Extending the interpolation technique to incremental gradient methods and coordinate-wise gradient methods?
Conclusions & Extensions


2. Application to matrix completion, where $\mathcal{E} = \mathbb{R}^{m \times n}$ and $P(x) = \|\sigma(x)\|_1$? Or to total-variation image restoration (joint work with Steve Wright)?

3. Extending the interpolation technique to incremental gradient methods and coordinate-wise gradient methods?

The END 📖