# 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 Opimization Problem

# Talk Outline

- A Convex Opimization Problem
- Proximal Gradient Method

# Talk Outline

- A Convex Opimization Problem
- Proximal Gradient Method
- Accelerated Proximal Gradient Method I
- Accelerated Proximal Gradient Method II

# Talk Outline

- A Convex Opimization Problem
- Proximal Gradient Method
- Accelerated Proximal Gradient Method I
- Accelerated Proximal Gradient Method II
- Example: Matrix Game

# Talk Outline

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

# A Convex Optimization Problem

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

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

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

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

$f : \mathcal{E} \to \Re$ is convex diff. $\|\nabla f(x) - \nabla f(y)\|_* \le L\|x - y\| \;\forall x, y \in \mathrm{dom}P$ ($L \ge 0$).

# 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 $\|x^*\|_* = \sup_{\|x\| \le 1} \langle x^*, x \rangle$.

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

$f : \mathcal{E} \to \Re$ is convex diff. $\|\nabla f(x) - \nabla f(y)\|_* \le L\|x - y\| \; \forall x, y \in \mathrm{dom} P$ ($L \ge 0$).

Constrained case: $P \equiv \delta_X$ 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:

- $\mathcal{E} = \Re^n$, $P(x) = \|x\|_1$, $f(x) = \|Ax - b\|_2^2$ Basis Pursuit/Lasso

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

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

- $\mathcal{E} = \mathcal{S}^n$, $P \equiv \delta_X$ with $X = \{x \mid |x_{ij}| \leq \rho\ \forall i, j\}$, $f(x) = g^*(x + s)$ with $g(y) = \begin{cases} -\ln\det y & \text{if } \alpha I \preceq y \preceq \beta I \\ \infty & \text{else} \end{cases}$ ($\rho, \alpha, \beta > 0$) 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

$$\begin{aligned} \ell(x;y) &:= f(y) + \langle \nabla f(y), x - y \rangle + P(x) \\ D(x,y) &:= h(x) - h(y) - \langle \nabla h(y), x - y \rangle, \end{aligned}$$

Bregman, ...

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

$$D(x,y) \geq \frac{1}{2}\|x - y\|^2 \qquad \forall\, x \in \mathrm{dom} P,\ y \in X_h.$$

# Proximal Gradient Method

Let

$$\begin{aligned} \ell(x;y) &:= f(y) + \langle \nabla f(y), x-y \rangle + P(x) \\ D(x,y) &:= h(x) - h(y) - \langle \nabla h(y), x-y \rangle, \end{aligned}$$ Bregman, ...

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

$$D(x,y) \geq \frac{1}{2}\|x-y\|^2 \qquad \forall\, x \in \mathrm{dom} P,\ y \in X_h.$$

For $k = 0, 1, \ldots$,

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

with $x_0 \in \mathrm{dom} P$. 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:

- $\mathcal{E} = \Re^n$, $\quad P(x) = \|x\|_1, \quad h(x) = \|x\|_2^2/2$.
- $\mathcal{E} = \Re^{n_1} \times \cdots \times \Re^{n_N}$, $\ P(x) = w_1\|x_1\|_2 + \cdots + w_N\|x_N\|_2$ ($w_j > 0$), $h(x) = \|x\|_2^2/2$.
- $\mathcal{E} = \Re^n$, $\quad P \equiv \delta_X$ with $X = \{x \mid x \geq 0, x_1 + \cdots + x_n = 1\}$, $h(x) = \sum_{j=1}^n x_j \ln x_j$.
- $\mathcal{E} = \mathcal{S}^n$, $\quad P \equiv \delta_X$ with $X = \{x \mid |x_{ij}| \leq \rho \ \forall i, j\}$, $\quad h(x) = \|x\|_F^2/2$.

Fact 1: $f^P(x) \geq \ell(x;y) \geq f^P(x) - \frac{L}{2}\|x-y\|^2 \quad \forall x, y \in \mathrm{dom}P$.

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

$$z_+ = \arg\min_x \left\{\psi(x) + D(x,z)\right\}.$$

If $z_+ \in X_h$, then

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

Prop. 1: For any $x \in \mathrm{dom}P$,

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

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

Prop. 1: For any $x \in \mathrm{dom} P$,

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

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

Proof:

$$\begin{aligned}
f^P(x_{k+1}) &\le \ell(x_{k+1}; x_k) + \frac{L}{2}\|x_{k+1} - x_k\|^2 && \text{Fact 1} \\
&\le \ell(x_{k+1}; x_k) + LD(x_{k+1}, x_k) \\
&\le \ell(x; x_k) + LD(x, x_k) - LD(x, x_{k+1}) && \text{Fact 2} \\
&\le f^P(x) + LD(x, x_k) - LD(x, x_{k+1}), && \text{Fact 1}
\end{aligned}$$

so

$$\begin{aligned}
0 \le LD(x, x_{k+1}) &\le LD(x, x_k) - e_{k+1} \\
&\le LD(x, x_0) - (e_1 + \cdots + e_{k+1}) \\
&\le LD(x, x_0) - (k+1)\min\{e_1, \ldots, e_{k+1}\}
\end{aligned}$$

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 $k = 0, 1, \ldots$,

$$\boxed{\begin{array}{lcl} y_k & = & (1-\theta_k)x_k + \theta_k z_k \\ z_{k+1} & = & \arg\min\limits_x \{\ell(x; y_k) + \theta_k L D(x, z_k)\} \\ x_{k+1} & = & (1-\theta_k)x_k + \theta_k z_{k+1} \\ \dfrac{1-\theta_{k+1}}{\theta_{k+1}^2} & \le & \dfrac{1}{\theta_k^2} \qquad (0 < \theta_{k+1} \le 1) \end{array}}$$

with $\theta_0 = 1$, $x_0, z_0 \in \mathrm{dom}P$ Nesterov, Auslender, Teboulle, Lan, Lu, Monteiro, ... Assume $z_k \in X_h$ $\forall k$.

For example, $\theta_k = \dfrac{2}{k+2}$ or $\theta_{k+1} = \dfrac{\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2}{2}$.

Prop. 2: For any $x \in \mathrm{dom} P$,

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

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

Prop. 2: For any $x \in \mathrm{dom} P$,

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

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

Proof:

$$\begin{aligned}
& f^P(x_{k+1}) \\
\leq\ & \ell(x_{k+1}; y_k) + \frac{L}{2}\|x_{k+1} - y_k\|^2 \qquad \text{Fact 1} \\
=\ & \ell((1-\theta_k)x_k + \theta_k z_{k+1}; y_k) + \frac{L}{2}\|(1-\theta_k)x_k + \theta_k z_{k+1} - y_k\|^2 \\
\leq\ & (1-\theta_k)\ell(x_k; y_k) + \theta_k \ell(z_{k+1}; y_k) + \frac{L}{2}\theta_k^2\|z_{k+1} - z_k\|^2 \\
\leq\ & (1-\theta_k)\ell(x_k; y_k) + \theta_k\left(\ell(z_{k+1}; y_k) + \theta_k LD(z_{k+1}, z_k)\right) \\
\leq\ & (1-\theta_k)\ell(x_k; y_k) + \theta_k\left(\ell(x; y_k) + \theta_k LD(x, z_k) - \theta_k LD(x, z_{k+1})\right) \quad \text{Fact 2} \\
\leq\ & (1-\theta_k)f^P(x_k) + \theta_k\left(f^P(x) + \theta_k LD(x, z_k) - \theta_k LD(x, z_{k+1})\right) \quad \text{Fact 1}
\end{aligned}$$

so, subtracting by $f^P(x)$ and then dividing by $\theta_k^2$, we have

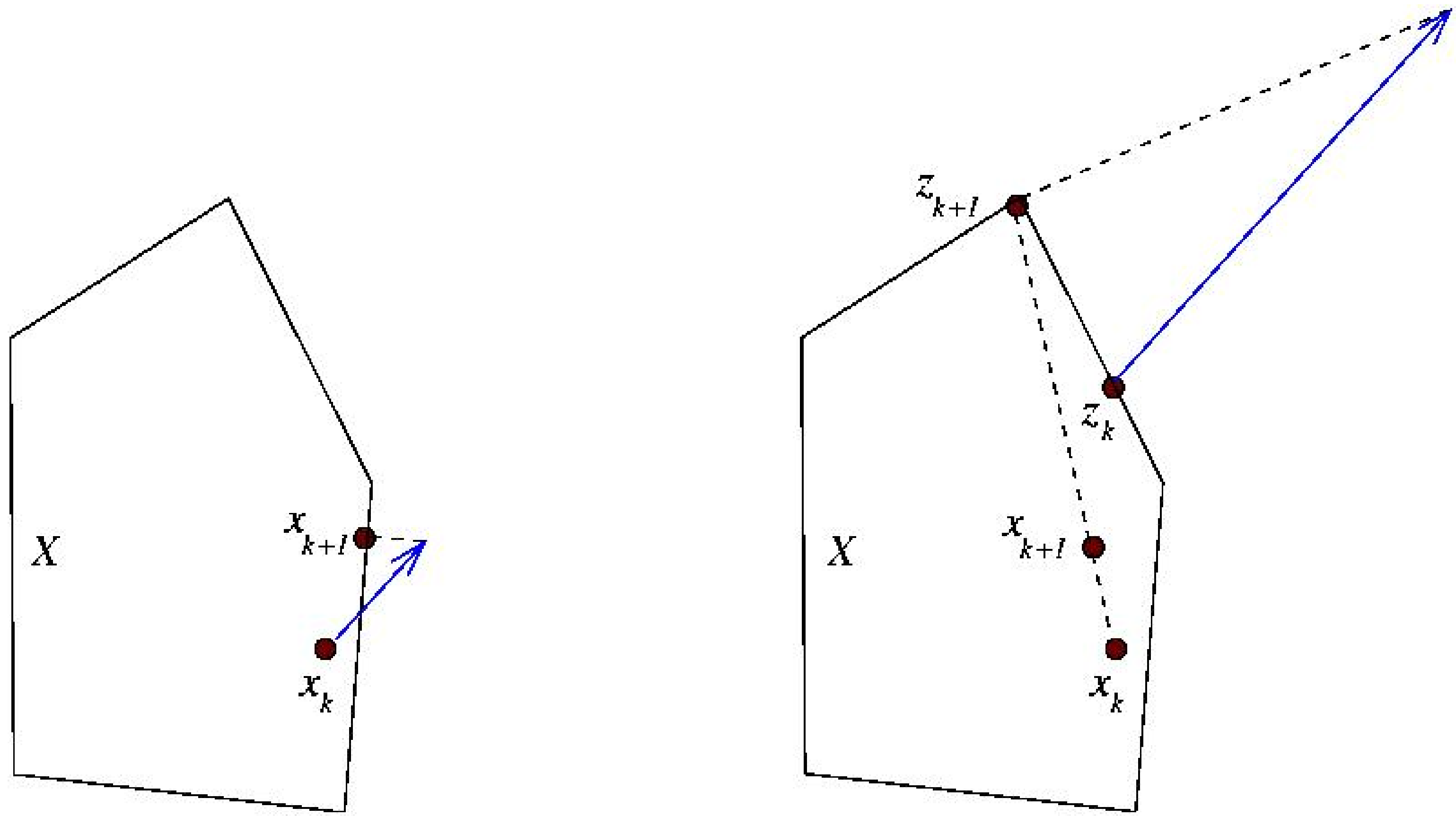
$$\frac{1}{\theta_k^2} e_{k+1} \leq \frac{1-\theta_k}{\theta_k^2} e_k + LD(x; z_k) - LD(x; z_{k+1})$$

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 $P \equiv \delta_X$.




PGM

APGM I

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 $k = 0, 1, \ldots,$

$$\boxed{\begin{aligned} y_k &= (1-\theta_k)x_k + \theta_k z_k \\ z_{k+1} &= \arg\min_x \left\{ \sum_{i=0}^{k} \frac{\ell(x; y_i)}{\vartheta_i} + Lh(x) \right\} \\ x_{k+1} &= (1-\theta_k)x_k + \theta_k z_{k+1} \\ \frac{1-\theta_{k+1}}{\theta_{k+1}\vartheta_{k+1}} &= \frac{1}{\theta_k \vartheta_k} \qquad (\vartheta_{k+1} \geq \theta_{k+1} > 0) \end{aligned}}$$

with $\vartheta_0 \geq \theta_0 = 1$, $x_0 \in \mathrm{dom} P$, and $z_0 = \arg\min_{x \in \mathrm{dom} P} h(x)$ Nesterov, d'Aspremont et al., Lu, ...

Assume $z_k \in X_h \ \forall k$.

For example, $\vartheta_k = \dfrac{2}{k+1}$, $\theta_k = \dfrac{2}{k+2}$ or $\vartheta_{k+1} = \theta_{k+1} = \dfrac{\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2}{2}$.

Prop. 3: For any $x \in \mathrm{dom} P$,

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

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

Prop. 3: For any $x \in \mathrm{dom}P$,

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

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

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

Advantage? Possibly better performance on compressed sensing and certain conic programs. Lu

# Example: Matrix Game

$$\boxed{\min_{x \in X} \max_{v \in V} \langle v, Ax \rangle}$$

with $X$ and $V$ unit simplices in $\Re^n$ and $\Re^m$, and $A \in \Re^{m \times n}$. 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

$$g(v) = \begin{cases} \sum_{i=1}^m v_i \ln v_i & \text{if } v \in V \\ \infty & \text{else} \end{cases} \qquad (L = \tfrac{1}{\mu},\ \| \cdot \| = 1\text{-norm})$$

# Example: Matrix Game

$$\boxed{\min_{x \in X} \max_{v \in V} \langle v, Ax \rangle}$$

with $X$ and $V$ unit simplices in $\Re^n$ and $\Re^m$, and $A \in \Re^{m \times n}$. 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

$$g(v) = \begin{cases} \sum_{i=1}^m v_i \ln v_i & \text{if } v \in V \\ \infty & \text{else} \end{cases} \qquad (L = \tfrac{1}{\mu},\ \|\cdot\| = 1\text{-norm})$$

- 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

$$\max_i (Ax_k)_i - \min_j (A^* v_k)_j \le \epsilon$$

with $v_k \in V$ a weighted sum of dual vectors associated with $x_0, x_1, \ldots, x_k$.

| | | PGM | APGM I | APGM II |
|---|---|---|---|---|
| $n/m/p$ | $\epsilon$ | k/cpu (sec) | k/cpu (sec) | k/cpu (sec) |
| 1000/100/.01 | .001 | 1082480/1500 | 3325/5 | 10510/9 |
| | .0001 | – | 20635/23 | 61865/45 |
| 10000/100/.01 | .001 | – | 10005/142 | 10005/128 |
| 10000/100/.1 | .001 | – | 10005/201 | 10005/185 |
| 10000/1000/.01 | .001 | – | 10005/202 | 10005/191 |
| 10000/1000/.1 | .001 | – | 10005/706 | 10005/695 |

Table 1: Performance of PGM, APGM I & II for different $n$, $m$, sparsity $p$, and soln accuracy $\epsilon$.

# Conclusions & Extensions

1. Accelerated prox gradient method is promising in theory and practice. Applicable to convex-concave optimization by using smoothing Nesterov. Further extension to add cutting planes.

# Conclusions & Extensions

1. Accelerated prox gradient method is promising in theory and practice. Applicable to convex-concave optimization by using smoothing Nesterov. Further extension to add cutting planes.

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

# Conclusions & Extensions

1. Accelerated prox gradient method is promising in theory and practice. Applicable to convex-concave optimization by using smoothing Nesterov. Further extension to add cutting planes.

2. Application to matrix completion, where $\mathcal{E} = \Re^{m \times n}$ 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

1. Accelerated prox gradient method is promising in theory and practice. Applicable to convex-concave optimization by using smoothing Nesterov. Further extension to add cutting planes.

2. Application to matrix completion, where $\mathcal{E} = \Re^{m \times n}$ 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