DATA8004: Optimization for Statistical Learning

Subject Lecturer: Man-Chung YUE

Lecture 7
Convex Optimization
Proximal gradient methods

Algorithm building block

Definition: Let $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper, closed and convex. We define the proximal operator (or proximal mapping) as

$$\mathrm{Prox}_g(x) := \operatorname*{Arg\,min}_{u \in \mathbb{R}^n} \left\{ \frac{1}{2}\|u - x\|_2^2 + g(u) \right\}.$$

Algorithm building block

Definition: Let $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper, closed and convex. We define the proximal operator (or proximal mapping) as

$$\mathrm{Prox}_g(x) := \operatorname*{Arg\,min}_{u \in \mathbb{R}^n} \left\{ \frac{1}{2}\|u - x\|_2^2 + g(u) \right\}.$$

Remark:

- Since g is proper, closed and convex, it follows that for any $x \in \mathbb{R}^n$, the function
$$u \mapsto \frac{1}{2}\|u - x\|_2^2 + g(u)$$
is proper, closed and strongly convex. Thus, it has a unique minimizer thanks to Theorem 4.12. Hence, $\mathrm{Prox}_g : \mathbb{R}^n \to \mathbb{R}^n$ is well defined.
- When $g = \delta_C$ for some nonempty closed convex set C, $\mathrm{Prox}_{\delta_C} = P_C$.

Example 1

Example: Let $C = \{x \in \mathbb{R}^n : \ \|x\|_1 \leq 1\}$. If $x \in C$, then $\mathrm{Prox}_{\delta_C}(x) = x$.

Example 1

Example: Let $C = \{x \in \mathbb{R}^n : \|x\|_1 \le 1\}$. If $x \in C$, then $\mathrm{Prox}_{\delta_C}(x) = x$.
Next, suppose $x \notin C$. Then $\|P_C(x)\|_1 = 1$ Why?.

Example 1

Example: Let $C = \{x \in \mathbb{R}^n : \ \|x\|_1 \leq 1\}$. If $x \in C$, then $\mathrm{Prox}_{\delta_C}(x) = x$. Next, suppose $x \notin C$. Then $\|P_C(x)\|_1 = 1$ Why?. Hence

$$\mathrm{Prox}_{\delta_C}(x) = \operatorname*{Arg\,min}_{\|u\|_1 \leq 1} \left\{ \frac{1}{2}\|u - x\|_2^2 \right\} = \operatorname*{Arg\,min}_{\|u\|_1 = 1} \left\{ \frac{1}{2}\|u\|_2^2 - u^T x \right\}.$$

Let $u = \alpha \circ v$, where $\alpha \in \{-1, 1\}^n$, $v \in \mathbb{R}^n_+$ and $\circ$ denotes entrywise product. Then

$$\min_{\|u\|_1 = 1} \left\{ \frac{1}{2}\|u\|_2^2 - u^T x \right\} = \min_{e^T v = 1, v \geq 0} \ \min_{\alpha \in \{-1,1\}^n} \left\{ \frac{1}{2}\|v\|_2^2 - v^T(\alpha \circ x) \right\}$$

$$= \min_{e^T v = 1, v \geq 0} \left\{ \frac{1}{2}\|v\|_2^2 - v^T|x| \right\} = \min_{e^T v = 1, v \geq 0} \left\{ \frac{1}{2}\|v - |x|\|_2^2 - \frac{1}{2}\|x\|_2^2 \right\}$$

with the min in α attained at $\alpha = \mathrm{sign}(x)$. Then $v = P_\Delta(|x|)$, where Δ is the unit simplex, and $|\cdot|$ is taken entrywise. Thus,

$$\mathrm{Prox}_{\delta_C}(x) = \mathrm{sign}(x) \circ P_\Delta(|x|), \quad \text{if } x \notin C.$$

See Slide 26 of Lecture 5 for P_Δ. Here, e denote the vector of ones.

Example 2

Example: Let $C = \{x \in \mathbb{R}^n : \ \|x\|_2 \leq 1\}$. Then $\text{Prox}_{\delta_C} = P_C$ and hence

$$\text{Prox}_{\delta_C}(x) = \begin{cases} x & \text{if } \|x\|_2 \leq 1, \\ \frac{x}{\|x\|_2} & \text{otherwise.} \end{cases}$$

Similarly, if $C = \{x \in \mathbb{R}^n : \ \|x\|_\infty \leq 1\}$, then for all i,

$$[\text{Prox}_{\delta_C}(x)]_i = \max\{-1, \min\{1, x_i\}\}.$$

Example 3

Example: We consider $\mathrm{Prox}_{\mu\|\cdot\|_1}$, where $\mu > 0$. Note that

$$\mathrm{Prox}_{\mu\|\cdot\|_1}(x) = \operatorname*{Arg\,min}_{u\in\mathbb{R}^n} \left\{ \sum_{i=1}^{n} \left(\frac{1}{2}(u_i - x_i)^2 + \mu|u_i| \right) \right\}$$

Example 3

Example: We consider $\operatorname{Prox}_{\mu\|\cdot\|_1}$, where $\mu > 0$. Note that

$$\operatorname{Prox}_{\mu\|\cdot\|_1}(x) = \operatorname*{Arg\,min}_{u\in\mathbb{R}^n}\left\{\sum_{i=1}^{n}\left(\frac{1}{2}(u_i - x_i)^2 + \mu|u_i|\right)\right\}$$

Consider for any $s \in \mathbb{R}$ the problem $\min_{t\in\mathbb{R}}\left\{\frac{1}{2}(t-s)^2 + \mu|t|\right\}$. Let $t = \alpha v$ with $\alpha \in \{-1, 1\}$ and $v \geq 0$, then

$$\min_{t\in\mathbb{R}}\left\{\frac{1}{2}(t-s)^2 + \mu|t|\right\} = \min_{v\geq 0}\min_{\alpha\in\{-1,1\}}\left\{\frac{1}{2}[v^2 - 2\alpha vs + s^2] + \mu v\right\}$$

$$= \min_{v\geq 0}\left\{\frac{1}{2}v^2 - v|s| + \mu v + \frac{1}{2}s^2\right\} = \min_{v\geq 0}\left\{\frac{1}{2}v^2 - (|s| - \mu)v + \frac{1}{2}s^2\right\},$$

where the min in α is attained at $\alpha = \operatorname{sign}(s)$.

Example 3

Example: We consider $\mathrm{Prox}_{\mu\|\cdot\|_1}$, where $\mu > 0$. Note that

$$\mathrm{Prox}_{\mu\|\cdot\|_1}(x) = \operatorname*{Arg\,min}_{u\in\mathbb{R}^n} \left\{ \sum_{i=1}^{n} \left(\frac{1}{2}(u_i - x_i)^2 + \mu|u_i| \right) \right\}$$

Consider for any $s \in \mathbb{R}$ the problem $\min_{t\in\mathbb{R}} \left\{ \frac{1}{2}(t-s)^2 + \mu|t| \right\}$. Let $t = \alpha v$ with $\alpha \in \{-1, 1\}$ and $v \geq 0$, then

$$\min_{t\in\mathbb{R}} \left\{ \frac{1}{2}(t-s)^2 + \mu|t| \right\} = \min_{v\geq 0} \min_{\alpha\in\{-1,1\}} \left\{ \frac{1}{2}[v^2 - 2\alpha v s + s^2] + \mu v \right\}$$

$$= \min_{v\geq 0} \left\{ \frac{1}{2}v^2 - v|s| + \mu v + \frac{1}{2}s^2 \right\} = \min_{v\geq 0} \left\{ \frac{1}{2}v^2 - (|s| - \mu)v + \frac{1}{2}s^2 \right\},$$

where the min in α is attained at $\alpha = \mathrm{sign}(s)$. Thus, the minimizer is $t_* = \mathrm{sign}(s)\max\{|s| - \mu, 0\}$. Hence, for all i,

$$\left[\mathrm{Prox}_{\mu\|\cdot\|_1}(x)\right]_i = \mathrm{sign}(x_i) \cdot \max\{|x_i| - \mu, 0\}.$$

Example 4

Example: We consider $\mathrm{Prox}_{\mu\|\cdot\|_2}$, where $\mu > 0$. Note that

$$\mathrm{Prox}_{\mu\|\cdot\|_2}(x) = \operatorname*{Arg\,min}_{u\in\mathbb{R}^n}\left\{\frac{1}{2}\|u\|_2^2 - u^T x + \mu\|u\|_2\right\}$$

Example 4

Example: We consider $\mathrm{Prox}_{\mu\|\cdot\|_2}$, where $\mu > 0$. Note that

$$\mathrm{Prox}_{\mu\|\cdot\|_2}(x) = \operatorname*{Arg\,min}_{u\in\mathbb{R}^n} \left\{ \frac{1}{2}\|u\|_2^2 - u^T x + \mu\|u\|_2 \right\}$$

Let $u = rv$ with $r \geq 0$ and $\|v\|_2 = 1$, then

$$\min_{u\in\mathbb{R}^n} \left\{ \frac{1}{2}\|u\|_2^2 - u^T x + \mu\|u\|_2 \right\} = \min_{r\geq 0} \min_{\|v\|_2=1} \left\{ \frac{1}{2}r^2 - rv^T x + \mu r \right\}$$

$$= \min_{r\geq 0} \left\{ \frac{1}{2}r^2 - r\|x\|_2 + \mu r \right\} = \min_{r\geq 0} \left\{ \frac{1}{2}r^2 - (\|x\|_2 - \mu)r \right\},$$

where the min in v is attained at $v = \mathrm{Sgn}(x)$, where

$$\mathrm{Sgn}(x) = \begin{cases} e/\|e\|_2 & \text{if } x = 0, \\ x/\|x\|_2 & \text{otherwise.} \end{cases}$$

Example 4

Example: We consider $\mathrm{Prox}_{\mu\|\cdot\|_2}$, where $\mu > 0$. Note that

$$\mathrm{Prox}_{\mu\|\cdot\|_2}(x) = \operatorname*{Arg\,min}_{u \in \mathbb{R}^n} \left\{ \frac{1}{2}\|u\|_2^2 - u^T x + \mu\|u\|_2 \right\}$$

Let $u = rv$ with $r \geq 0$ and $\|v\|_2 = 1$, then

$$\min_{u \in \mathbb{R}^n} \left\{ \frac{1}{2}\|u\|_2^2 - u^T x + \mu\|u\|_2 \right\} = \min_{r \geq 0} \min_{\|v\|_2 = 1} \left\{ \frac{1}{2} r^2 - r v^T x + \mu r \right\}$$

$$= \min_{r \geq 0} \left\{ \frac{1}{2} r^2 - r\|x\|_2 + \mu r \right\} = \min_{r \geq 0} \left\{ \frac{1}{2} r^2 - (\|x\|_2 - \mu) r \right\},$$

where the min in v is attained at $v = \mathrm{Sgn}(x)$, where

$$\mathrm{Sgn}(x) = \begin{cases} e/\|e\|_2 & \text{if } x = 0, \\ x/\|x\|_2 & \text{otherwise.} \end{cases}$$

Thus, $\mathrm{Prox}_{\mu\|\cdot\|_2}(x) = \max\{\|x\|_2 - \mu, 0\} \cdot \mathrm{Sgn}(x)$.

Nonexpansiveness

Proposition 7.1 (Nonexpansiveness)
Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper, closed and convex. Then it holds that

$$\|\mathrm{Prox}_f(x) - \mathrm{Prox}_f(y)\|_2 \leq \|x - y\|_2$$

for all x and $y \in \mathbb{R}^n$.

Nonexpansiveness

Proposition 7.1 (Nonexpansiveness)
Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper, closed and convex. Then it holds that

$$\|\mathrm{Prox}_f(x) - \mathrm{Prox}_f(y)\|_2 \leq \|x - y\|_2$$

for all x and $y \in \mathbb{R}^n$.

Proof: Write $u = \mathrm{Prox}_f(x)$ and $v = \mathrm{Prox}_f(y)$. Then we see from Theorem 5.5 and Proposition 4.4(ii) that

$$x - u \in \partial f(u) \quad \text{and} \quad y - v \in \partial f(v).$$

These together with Proposition 4.4(iii) give

$$(u - v)^T[(x - u) - (y - v)] \geq 0.$$

Rearrange terms and apply the Cauchy-Schwartz inequality.

Problem setting

We consider the following optimization problem:

$$\underset{x\in\mathbb{R}^n}{\text{Minimize}} \quad f(x) + g(x), \tag{1}$$

where

- f is convex with Lipschitz continuous gradient. Specifically, there exists $L > 0$ such that

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2, \quad \forall x, y \in \mathbb{R}^n.$$

- g is proper, closed and convex.
- We assume that $\operatorname{Arg\,min}(f + g) \neq \emptyset$.
- We also assume that the proximal operator of g can be computed efficiently.

Example

Example: (LASSO / Compressed sensing)

Let $A \in \mathbb{R}^{m \times n}$ with $m \ll n$, $b \in \mathbb{R}^m$ and $\mu > 0$. It is common to consider the following models for sparse recovery:

$$\underset{x \in \mathbb{R}^n}{\text{Minimize}} \quad \frac{1}{2}\|Ax - b\|_2^2 + \mu\|x\|_1$$

or

$$\begin{array}{rl} \underset{x \in \mathbb{R}^n}{\text{Minimize}} & \frac{1}{2}\|Ax - b\|_2^2 \\ \text{Subject to} & \|x\|_1 \le \mu. \end{array}$$

See [Tibshirani '96] and [Foucart-Rauhut '13].

Proximal gradient algorithm

Consider (1) and recall that L is a Lipschitz continuity modulus of ∇f.

Proximal gradient algorithm: Let $x^0 \in \operatorname{dom} g$. For $k = 0, 1, \ldots$,

$$x^{k+1} = \operatorname{Prox}_{\frac{1}{L}g}\left(x^k - \frac{1}{L}\nabla f(x^k)\right).$$

Proximal gradient algorithm

Consider (1) and recall that L is a Lipschitz continuity modulus of ∇f.

Proximal gradient algorithm: Let $x^0 \in \operatorname{dom} g$. For $k = 0, 1, \ldots$,

$$x^{k+1} = \operatorname{Prox}_{\frac{1}{L}g}\left(x^k - \frac{1}{L}\nabla f(x^k)\right).$$

Remark:

- Note that

$$x^{k+1} = \operatorname*{Arg\,min}_{x \in \mathbb{R}^n}\left\{\nabla f(x^k)^T(x - x^k) + \frac{L}{2}\|x - x^k\|_2^2 + g(x)\right\} \quad (2)$$

- This algorithm is globally convergent. Specifically, $\{x^k\}$ converges to a global minimizer of $f + g$ from any $x^0 \in \operatorname{dom} g$. See [Lions-Mercier '79], [Tseng '91] and references therein.
- In this lecture, we follow the lines of analysis in [Tseng unpub] and [Tseng '10] to derive iteration complexity.

Taylor's inequality

Theorem 7.1 (Taylor's inequality)
Let $f \in C^1(\mathbb{R}^n)$ and suppose that there exists $L > 0$ so that

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2, \quad \forall x, y \in \mathbb{R}^n.$$

Then for all x and $y \in \mathbb{R}^n$, it holds that

$$f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{L}{2}\|y - x\|_2^2.$$

Taylor's inequality

Theorem 7.1 (Taylor's inequality)
Let $f \in C^1(\mathbb{R}^n)$ and suppose that there exists $L > 0$ so that

$$\|\nabla f(x) - \nabla f(y)\|_2 \le L\|x - y\|_2, \quad \forall x, y \in \mathbb{R}^n.$$

Then for all x and $y \in \mathbb{R}^n$, it holds that

$$f(y) \le f(x) + \nabla f(x)^T(y - x) + \frac{L}{2}\|y - x\|_2^2.$$

Proof sketch: For any $x, y \in \mathbb{R}^n$ and define $\psi(t) := f(x + t(y - x))$. Then $\psi(0) = f(x)$ and $\psi'(s) = (y - x)^T \nabla f(x + s(y - x))$. Hence

$$\begin{aligned}
\psi(1) = \psi(0) + \int_0^1 \psi'(s)ds &= \psi(0) + \psi'(0) + \int_0^1 (\psi'(s) - \psi'(0))ds \\
&= \psi(0) + \psi'(0) + \int_0^1 (y - x)^T[\nabla f(x + s(y - x)) - \nabla f(x)]ds \\
&\le \psi(0) + \psi'(0) + L\int_0^1 s\|y - x\|^2 ds.
\end{aligned}$$

Convergence of PG

Theorem 7.2 (PG complexity)
Consider (1) and let $\{x^k\}$ be generated by the proximal gradient algorithm on Slide 9. Then for all $k \geq 1$, it holds that

$$F(x^k) - F(\bar{x}) \leq \frac{L}{2k}\|x^0 - \bar{x}\|_2^2,$$

where $F := f + g$ and $\bar{x}$ is any element in $\operatorname{Arg\,min} F$.

Convergence of PG

Theorem 7.2 (PG complexity)
Consider (1) and let $\{x^k\}$ be generated by the proximal gradient algorithm on Slide 9. Then for all $k \geq 1$, it holds that

$$F(x^k) - F(\bar{x}) \leq \frac{L}{2k}\|x^0 - \bar{x}\|_2^2,$$

where $F := f + g$ and $\bar{x}$ is any element in $\operatorname{Arg\,min} F$.

Proof: Let $x \in \operatorname{dom} g$. Note from Taylor's inequality that we have for all $k \geq 0$

$$F(x^{k+1}) \leq f(x^k) + \nabla f(x^k)^T(x^{k+1} - x^k) + \frac{L}{2}\|x^{k+1} - x^k\|_2^2 + g(x^{k+1})$$

$$\leq f(x^k) + \nabla f(x^k)^T(x - x^k) + \frac{L}{2}\|x - x^k\|_2^2 + g(x) - \frac{L}{2}\|x^{k+1} - x\|_2^2,$$

where the 2nd inequality follows from (2) and Theorem 4.12.

Convergence of PG cont.

Proof of Theorem 7.2 cont.: Setting $x = x^k$, we get

$$F(x^{k+1}) \le f(x^k) + g(x^k) - \frac{L}{2}\|x^{k+1} - x^k\|_2^2 \le F(x^k)$$

showing that $\{F(x^k)\}$ is nonincreasing.

Now, pick any $\bar{x} \in \operatorname{Arg\,min} F$ and let $x = \bar{x}$, then

$$\begin{aligned} F(x^{k+1}) &\le f(x^k) + \nabla f(x^k)^T(\bar{x} - x^k) + \frac{L}{2}\|\bar{x} - x^k\|_2^2 + g(\bar{x}) - \frac{L}{2}\|x^{k+1} - \bar{x}\|_2^2 \\ &\le f(\bar{x}) + g(\bar{x}) + \frac{L}{2}\|\bar{x} - x^k\|_2^2 - \frac{L}{2}\|x^{k+1} - \bar{x}\|_2^2. \end{aligned}$$

Convergence of PG cont.

Proof of Theorem 7.2 cont.: Setting $x = x^k$, we get

$$F(x^{k+1}) \le f(x^k) + g(x^k) - \frac{L}{2}\|x^{k+1} - x^k\|_2^2 \le F(x^k)$$

showing that $\{F(x^k)\}$ is nonincreasing.

Now, pick any $\bar{x} \in \operatorname{Arg\,min} F$ and let $x = \bar{x}$, then

$$F(x^{k+1}) \le f(x^k) + \nabla f(x^k)^T(\bar{x} - x^k) + \frac{L}{2}\|\bar{x} - x^k\|_2^2 + g(\bar{x}) - \frac{L}{2}\|x^{k+1} - \bar{x}\|_2^2$$
$$\le f(\bar{x}) + g(\bar{x}) + \frac{L}{2}\|\bar{x} - x^k\|_2^2 - \frac{L}{2}\|x^{k+1} - \bar{x}\|_2^2.$$

Hence,

$$(k+1)[F(x^{k+1}) - F(\bar{x})] \le \textstyle\sum_{i=0}^{k}(F(x^{i+1}) - F(\bar{x}))$$
$$\le \frac{L}{2}\sum_{i=0}^{k}[\|\bar{x} - x^i\|_2^2 - \|x^{i+1} - \bar{x}\|_2^2] \le \frac{L}{2}\|\bar{x} - x^0\|_2^2.$$

Remarks on PG

Remark:

- The efficiency of PG relies heavily on the efficiency for computing proximal mapping. Each iteration involves one evaluation of ∇f and one proximal mapping computation.
- When $g = 0$, the PG is just the steepest descent with constant stepsize in Lecture 2. This suggests that PG may require lots of iterations in practice.
- PG requires explicit knowledge of L, which can be restrictive in applications. There are variants using line-search strategies to replace the constant L.
- In this lecture, we look at another kind of acceleration strategy based on Nesterov's extrapolation techniques; see [Nesterov '83] and [Nesterov '06]. Our discussions follow closely [Tseng unpub] and [Tseng '10].

Accelerated PG

Consider (1) and recall that L is a Lipschitz continuity modulus of ∇f.

Accelerated PG: Let $\theta_0 = \theta_{-1} = 1$, $x^0 = x^{-1} \in \operatorname{dom} g$. For $k \geq 0$, compute

$$\begin{cases} y^k &= x^k + \theta_k(\theta_{k-1}^{-1} - 1)(x^k - x^{k-1}), \\ x^{k+1} &= \operatorname{Prox}_{\frac{1}{L}g}\left(y^k - \frac{1}{L}\nabla f(y^k)\right), \end{cases}$$

and choose $\theta_{k+1} \in (0, 1]$ so that

$$\frac{1-\theta_{k+1}}{\theta_{k+1}^2} \leq \frac{1}{\theta_k^2}. \tag{3}$$

Accelerated PG

Consider (1) and recall that L is a Lipschitz continuity modulus of ∇f.

Accelerated PG: Let $\theta_0 = \theta_{-1} = 1$, $x^0 = x^{-1} \in \operatorname{dom} g$. For $k \geq 0$, compute

$$\begin{cases} y^k &= x^k + \theta_k(\theta_{k-1}^{-1} - 1)(x^k - x^{k-1}), \\ x^{k+1} &= \operatorname{Prox}_{\frac{1}{L}g}\left(y^k - \frac{1}{L}\nabla f(y^k)\right), \end{cases}$$

and choose $\theta_{k+1} \in (0,1]$ so that

$$\frac{1-\theta_{k+1}}{\theta_{k+1}^2} \leq \frac{1}{\theta_k^2}. \tag{3}$$

Remark:

- Notice that $\theta_k \equiv 1$ verifies (3): This choice gives back PG!

Accelerated PG

Consider (1) and recall that L is a Lipschitz continuity modulus of ∇f.

Accelerated PG: Let $\theta_0 = \theta_{-1} = 1$, $x^0 = x^{-1} \in \operatorname{dom} g$. For $k \geq 0$, compute

$$\begin{cases} y^k &= x^k + \theta_k(\theta_{k-1}^{-1} - 1)(x^k - x^{k-1}), \\ x^{k+1} &= \mathrm{Prox}_{\frac{1}{L}g}\left(y^k - \frac{1}{L}\nabla f(y^k)\right), \end{cases}$$

and choose $\theta_{k+1} \in (0, 1]$ so that

$$\frac{1-\theta_{k+1}}{\theta_{k+1}^2} \leq \frac{1}{\theta_k^2}. \tag{3}$$

Remark:

- Notice that $\theta_k \equiv 1$ verifies (3): This choice gives back PG! Accelerated PG typically chooses $\theta_k = O(1/k)$. We will stick to this latter choice.

Accelerated PG

Consider (1) and recall that L is a Lipschitz continuity modulus of ∇f.

Accelerated PG: Let $\theta_0 = \theta_{-1} = 1$, $x^0 = x^{-1} \in \operatorname{dom} g$. For $k \geq 0$, compute

$$\begin{cases} y^k &= x^k + \theta_k(\theta_{k-1}^{-1} - 1)(x^k - x^{k-1}), \\ x^{k+1} &= \operatorname{Prox}_{\frac{1}{L}g}\left(y^k - \frac{1}{L}\nabla f(y^k)\right), \end{cases}$$

and choose $\theta_{k+1} \in (0,1]$ so that

$$\frac{1-\theta_{k+1}}{\theta_{k+1}^2} \leq \frac{1}{\theta_k^2}. \tag{3}$$

Remark:

- Notice that $\theta_k \equiv 1$ verifies (3): This choice gives back PG! Accelerated PG typically chooses $\theta_k = O(1/k)$. We will stick to this latter choice.
- A common choice of θ_k is $\frac{2}{k+2}$. Another popular choice is

$$\theta_{k+1} = \frac{1}{2}\left(\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2\right); \tag{4}$$

  one can show using (4) and induction that $\theta_k \leq \frac{2}{k+2}$. Exercise!

Accelerated PG

Consider (1) and recall that L is a Lipschitz continuity modulus of ∇f.

Accelerated PG: Let $\theta_0 = \theta_{-1} = 1$, $x^0 = x^{-1} \in \operatorname{dom} g$. For $k \geq 0$, compute

$$\begin{cases} y^k &= x^k + \theta_k(\theta_{k-1}^{-1} - 1)(x^k - x^{k-1}), \\ x^{k+1} &= \operatorname{Prox}_{\frac{1}{L}g}\left(y^k - \frac{1}{L}\nabla f(y^k)\right), \end{cases}$$

and choose $\theta_{k+1} \in (0, 1]$ so that

$$\frac{1-\theta_{k+1}}{\theta_{k+1}^2} \leq \frac{1}{\theta_k^2}. \tag{3}$$

Remark:

- Notice that $\theta_k \equiv 1$ verifies (3): This choice gives back PG! Accelerated PG typically chooses $\theta_k = O(1/k)$. We will stick to this latter choice.
- A common choice of θ_k is $\frac{2}{k+2}$. Another popular choice is

$$\theta_{k+1} = \frac{1}{2}\left(\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2\right); \tag{4}$$

  one can show using (4) and induction that $\theta_k \leq \frac{2}{k+2}$. Exercise!

The update of y^k is an *extrapolation* step as $\theta_k(\theta_{k-1}^{-1} - 1) \geq 0$.

Accelerated PG

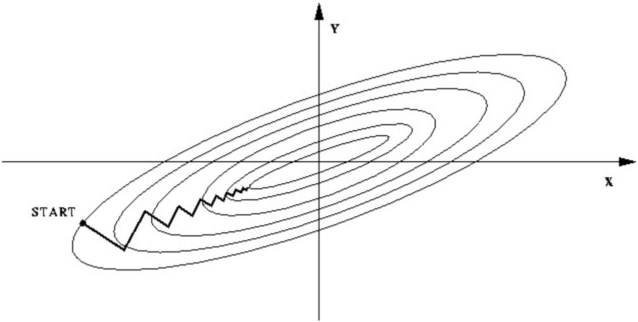


Figure: Trajectory of gradient descent on quadratic problems

Accelerated PG

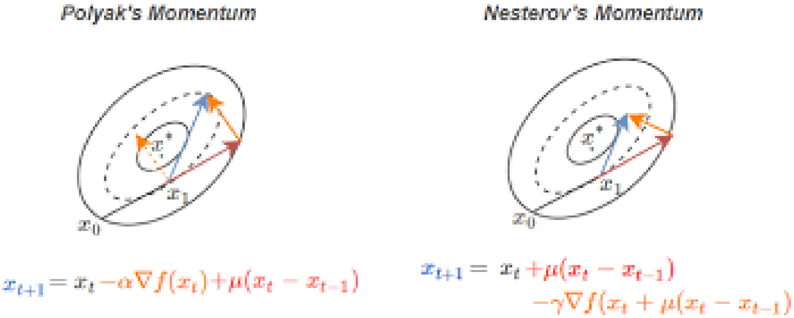


Figure: Correcting gradient direction by momentum

Accelerated PG

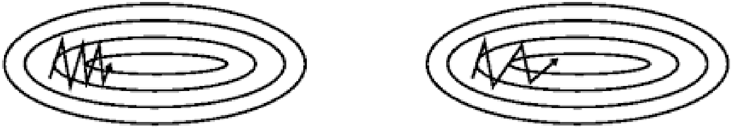


Figure: GD with and without momentum

Convergence of accelerated PG

Theorem 7.3 (accelerated PG complexity)
Consider (1) and let $\{x^k\}$ be generated by the accelerated proximal gradient algorithm on Slide 14. Then for all $k \geq 1$, it holds that

$$F(x^k) - F(\bar{x}) \leq \frac{L\theta_{k-1}^2}{2}\|x^0 - \bar{x}\|_2^2,$$

where $F := f + g$ and $\bar{x}$ is any element in $\operatorname{Arg\,min} F$.

Convergence of accelerated PG

Theorem 7.3 (accelerated PG complexity)
Consider (1) and let $\{x^k\}$ be generated by the accelerated proximal gradient algorithm on Slide 14. Then for all $k \geq 1$, it holds that

$$F(x^k) - F(\bar{x}) \leq \frac{L\theta_{k-1}^2}{2}\|x^0 - \bar{x}\|_2^2,$$

where $F := f + g$ and $\bar{x}$ is any element in $\operatorname{Arg\,min} F$.

Proof: Let $y \in \operatorname{dom} g$. Note from Taylor's inequality that for all $k \geq 0$

$$F(x^{k+1}) \leq f(y^k) + \nabla f(y^k)^T(x^{k+1} - y^k) + \frac{L}{2}\|x^{k+1} - y^k\|_2^2 + g(x^{k+1})$$

$$\leq f(y^k) + \nabla f(y^k)^T(y - y^k) + \frac{L}{2}\|y - y^k\|_2^2 + g(y) - \frac{L}{2}\|x^{k+1} - y\|_2^2,$$

where the 2nd inequality follows from (2) and Theorem 4.12.

Convergence of accelerated PG cont.

Proof of Theorem 7.3 cont.: Using convexity of f, we further have

$$F(x^{k+1}) \leq F(y) + \frac{L}{2}\|y - y^k\|_2^2 - \frac{L}{2}\|x^{k+1} - y\|_2^2.$$

Convergence of accelerated PG cont.

Proof of Theorem 7.3 cont.: Using convexity of f, we further have

$$F(x^{k+1}) \leq F(y) + \frac{L}{2}\|y - y^k\|_2^2 - \frac{L}{2}\|x^{k+1} - y\|_2^2.$$

Now, let $\bar{x} \in \operatorname{Arg\,min} F$ and set $y = (1-\theta_k)x^k + \theta_k \bar{x}$: This is a convex combination because $\theta_k \in [0, 1]$. Hence,

$$\begin{aligned}
F(x^{k+1}) \leq\ & F((1-\theta_k)x^k + \theta_k\bar{x}) + \frac{L}{2}\|(1-\theta_k)x^k + \theta_k\bar{x} - y^k\|_2^2 \\
& - \frac{L}{2}\|(1-\theta_k)x^k + \theta_k\bar{x} - x^{k+1}\|_2^2 \\
=\ & F((1-\theta_k)x^k + \theta_k\bar{x}) + \frac{L\theta_k^2}{2}\|\bar{x} + (\theta_k^{-1} - 1)x^k - \theta_k^{-1}y^k\|_2^2 \\
& - \frac{L\theta_k^2}{2}\|\bar{x} + (\theta_k^{-1} - 1)x^k - \theta_k^{-1}x^{k+1}\|_2^2
\end{aligned}$$

Convergence of accelerated PG cont.

Proof of Theorem 7.3 cont.: Here comes the magic

Convergence of accelerated PG cont.

Proof of Theorem 7.3 cont.: Here comes the magic — Observe that y^k is defined in such a way so that

$$\begin{aligned}
z^k &:= -(\theta_k^{-1} - 1)x^k + \theta_k^{-1} y^k \\
&= -(\theta_k^{-1} - 1)x^k + \theta_k^{-1} x^k + (\theta_{k-1}^{-1} - 1)(x^k - x^{k-1}) \\
&= -(\theta_{k-1}^{-1} - 1)x^{k-1} + \theta_{k-1}^{-1} x^k.
\end{aligned}$$

Convergence of accelerated PG cont.

Proof of Theorem 7.3 cont.: Here comes the magic — Observe that y^k is defined in such a way so that

$$\begin{aligned} z^k &:= -(\theta_k^{-1} - 1)x^k + \theta_k^{-1} y^k \\ &= -(\theta_k^{-1} - 1)x^k + \theta_k^{-1} x^k + (\theta_{k-1}^{-1} - 1)(x^k - x^{k-1}) \\ &= -(\theta_{k-1}^{-1} - 1)x^{k-1} + \theta_{k-1}^{-1} x^k. \end{aligned}$$

Thus, we have further that

$$\begin{aligned} F(x^{k+1}) &\leq F((1-\theta_k)x^k + \theta_k \bar{x}) + \frac{L\theta_k^2}{2}\|\bar{x} - z^k\|_2^2 - \frac{L\theta_k^2}{2}\|\bar{x} - z^{k+1}\|_2^2 \\ &\leq (1-\theta_k)F(x^k) + \theta_k F(\bar{x}) + \frac{L\theta_k^2}{2}\|\bar{x} - z^k\|_2^2 - \frac{L\theta_k^2}{2}\|\bar{x} - z^{k+1}\|_2^2. \end{aligned}$$

Convergence of accelerated PG cont.

Proof of Theorem 7.3 cont.: Rearranging terms, we have for all $k \geq 0$

$$F(x^{k+1}) - F(\bar{x}) \leq (1-\theta_k)[F(x^k) - F(\bar{x})] + \frac{L\theta_k^2}{2}\|\bar{x} - z^k\|_2^2 - \frac{L\theta_k^2}{2}\|\bar{x} - z^{k+1}\|_2^2.$$

Hence

$$\begin{aligned}
&\frac{(1-\theta_{k+1})}{\theta_{k+1}^2}[F(x^{k+1}) - F(\bar{x})] + \frac{L}{2}\|\bar{x} - z^{k+1}\|_2^2 \\
&\leq \frac{1}{\theta_k^2}[F(x^{k+1}) - F(\bar{x})] + \frac{L}{2}\|\bar{x} - z^{k+1}\|_2^2 \\
&\leq \frac{(1-\theta_k)}{\theta_k^2}[F(x^k) - F(\bar{x})] + \frac{L}{2}\|\bar{x} - z^k\|_2^2 \\
&\leq \cdots \leq \frac{1-\theta_0}{\theta_0^2}[F(x^0) - F(\bar{x})] + \frac{L}{2}\|\bar{x} - x^0\|_2^2 = \frac{L}{2}\|\bar{x} - x^0\|_2^2,
\end{aligned}$$

since $z^0 = x^0$ and $\theta_0 = 1$.

Remark on accelerated PG

Remark: With θ_k chosen to be $O(\frac{1}{k})$:

- APG is in general not a descent algorithm!
- This class of method is commonly known as optimal methods: According to Section 2.1.2 in [Nesterov '06], there exists a convex function f with Lipschitz gradient and $\operatorname{Arg\,min} f \neq \emptyset$ such that for any first-order method that generates iterates as

$$x^k \in x^0 + \operatorname{span}\{\nabla f(x^0), \nabla f(x^1), \ldots, \nabla f(x^{k-1})\}, \quad k \geq 1,$$

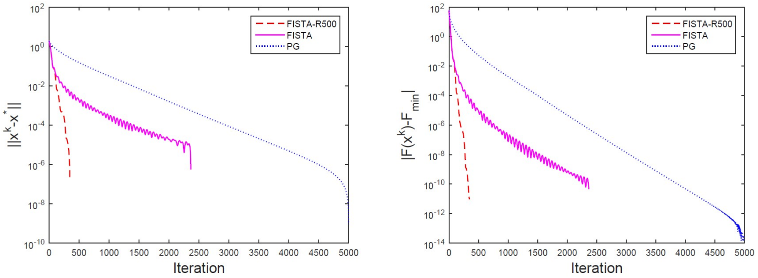
  it holds that for any $\bar{x} \in \operatorname{Arg\,min} f$,

$$f(x^k) - f(\bar{x}) \geq \frac{3L\|x^0 - \bar{x}\|_2^2}{32(k+1)^2}$$

  whenever $1 \leq k \leq \frac{1}{2}(n-1)$.
- APG motivated researches on extrapolation, and many variants have since been proposed. See [Becker-Candès-Grant '11] and the associated software TFOCS.

Restart strategy

- The optimality result suggests that Nesterov's extrapolation techniques is only optimal when k is not too large.
- This suggests restarting the θ_k from time to time, i.e., set $\theta_{k-1} = \theta_k = 1$ after certain number of iterations and/or some criterion is satisfied; see [O'Donoghue-Candès '15].
- The restart strategy has been implemented in TFOCS.



The pictures are taken from [Wen-Chen-Pong '17].

Case study

As an example, we illustrate how to design termination criterion for the following convex optimization problem; see for example [Wen-Chen-Pong '17].

Consider

$$\underset{x \in \mathbb{R}^n}{\text{Minimize}} \quad F(x) := \frac{1}{2}\|Ax - b\|_2^2 + \mu\|x\|_1, \tag{5}$$

where $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ and $\mu > 0$.

Case study

As an example, we illustrate how to design termination criterion for the following convex optimization problem; see for example [Wen-Chen-Pong '17].

Consider

$$\underset{x \in \mathbb{R}^n}{\text{Minimize}} \quad F(x) := \frac{1}{2}\|Ax - b\|_2^2 + \mu\|x\|_1, \tag{5}$$

where $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ and $\mu > 0$.

- Note that (5) has a solution in view of Theorem 1.4 because $\{x : \; F(x) \le F(x^0)\}$ is closed and bounded for any $x^0 \in \mathbb{R}^n$.

Case study

As an example, we illustrate how to design termination criterion for the following convex optimization problem; see for example [Wen-Chen-Pong '17].

Consider

$$\operatorname*{Minimize}_{x\in\mathbb{R}^n} \quad F(x) := \frac{1}{2}\|Ax-b\|_2^2 + \mu\|x\|_1, \tag{5}$$

where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ and $\mu > 0$.

- Note that (5) has a solution in view of Theorem 1.4 because $\{x : \; F(x) \le F(x^0)\}$ is closed and bounded for any $x^0 \in \mathbb{R}^n$.
- If we set $g(x) = \mu\|x\|_1$ and $h(y) = \frac{1}{2}\|y-b\|_2^2$, the above problem is $\inf_{x\in\mathbb{R}^n}\{h(Ax)+g(x)\}$. Notice that $\operatorname{dom} g = \mathbb{R}^n$ and $\operatorname{dom} h = \mathbb{R}^m$. Thus, Theorem 5.4 states that (5) has the same optimal value as Notice that the $-$ve sign is moved from h^* to g^*

$$\sup_{u\in\mathbb{R}^m}\{-g^*(-A^Tu) - h^*(u)\}.$$

Case study

As an example, we illustrate how to design termination criterion for the following convex optimization problem; see for example [Wen-Chen-Pong '17].

Consider

$$\underset{x\in\mathbb{R}^n}{\text{Minimize}} \quad F(x) := \frac{1}{2}\|Ax-b\|_2^2 + \mu\|x\|_1, \tag{5}$$

where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ and $\mu > 0$.

- Note that (5) has a solution in view of Theorem 1.4 because $\{x : F(x) \le F(x^0)\}$ is closed and bounded for any $x^0 \in \mathbb{R}^n$.
- If we set $g(x) = \mu\|x\|_1$ and $h(y) = \frac{1}{2}\|y-b\|_2^2$, the above problem is $\inf_{x\in\mathbb{R}^n}\{h(Ax)+g(x)\}$. Notice that $\operatorname{dom} g = \mathbb{R}^n$ and $\operatorname{dom} h = \mathbb{R}^m$. Thus, Theorem 5.4 states that (5) has the same optimal value as Notice that the $-$ve sign is moved from h^* to g^*

$$\sup_{u\in\mathbb{R}^m}\{-g^*(-A^Tu) - h^*(u)\}.$$

Moreover, this dual problem also has an optimal solution.

Case study cont.

Notice that

$$g^*(v) = \sup_{y \in \mathbb{R}^n} \left\{ y^T v - \mu \|y\|_1 \right\} = \delta_{\|\cdot\|_\infty \le \mu}(v),$$

$$h^*(w) = \sup_{x \in \mathbb{R}^m} \left\{ w^T x - \frac{1}{2}\|x - b\|_2^2 \right\} = \frac{1}{2}\|w\|_2^2 + b^T w.$$

Hence, the dual problem is given by

$$\begin{array}{ll} \underset{u \in \mathbb{R}^m}{\text{Maximize}} & D(u) := -\frac{1}{2}\|u\|_2^2 - b^T u \\ \text{Subject to} & \|A^T u\|_\infty \le \mu. \end{array} \tag{6}$$

Case study cont.

Notice that

$$g^*(v) = \sup_{y \in \mathbb{R}^n} \left\{ y^T v - \mu \|y\|_1 \right\} = \delta_{\|\cdot\|_\infty \le \mu}(v),$$

$$h^*(w) = \sup_{x \in \mathbb{R}^m} \left\{ w^T x - \frac{1}{2}\|x - b\|_2^2 \right\} = \frac{1}{2}\|w\|_2^2 + b^T w.$$

Hence, the dual problem is given by

$$\begin{array}{ll} \underset{u \in \mathbb{R}^m}{\text{Maximize}} & D(u) := -\frac{1}{2}\|u\|_2^2 - b^T u \\ \text{Subject to} & \|A^T u\|_\infty \le \mu. \end{array} \tag{6}$$

- Note that we always have $F(x) \ge \inf F \ge D(u)$ whenever u satisfies $\|A^T u\|_\infty \le \mu$.
- Suppose we apply PG / APG to minimize F in (5). Then $F(x^k) \ge \inf F$ for all k and $F(x^k) \to \inf F$. Can we construct (approximately) feasible $\{u^k\}$ suitably so that $D(u^k) \to \inf F$?

Case study cont.

Let $\bar{x}$ solve (5). Then Theorem 5.5 and Proposition 4.4(ii) give

$$0 \in \partial(h \circ A + g)(\overline{x}) = A^T \nabla h(A\overline{x}) + \partial g(\overline{x}).$$

Using Young's inequality and Theorem 5.1(iii), we have

$$\begin{aligned} & -A^T \nabla h(A\overline{x}) \in \partial g(\overline{x}) \\ \iff\ & \overline{x} \in \partial g^*(-A^T\overline{y}) \ \text{ and } \ \overline{y} = \nabla h(A\overline{x}) \\ \iff\ & \overline{x} \in \partial g^*(-A^T\overline{y}) \ \text{ and } \ A\overline{x} \in \partial h^*(\overline{y}). \end{aligned}$$

Case study cont.

Let $\bar{x}$ solve (5). Then Theorem 5.5 and Proposition 4.4(ii) give

$$0 \in \partial(h \circ A + g)(\overline{x}) = A^T \nabla h(A\overline{x}) + \partial g(\overline{x}).$$

Using Young's inequality and Theorem 5.1(iii), we have

$$\begin{aligned} & -A^T \nabla h(A\overline{x}) \in \partial g(\overline{x}) \\ \iff \; & \overline{x} \in \partial g^*(-A^T \overline{y}) \;\text{ and }\; \overline{y} = \nabla h(A\overline{x}) \\ \iff \; & \overline{x} \in \partial g^*(-A^T \overline{y}) \;\text{ and }\; A\overline{x} \in \partial h^*(\overline{y}). \end{aligned}$$

The last relation above gives

$$\begin{aligned} 0 &\in \partial h^*(\overline{y}) - A\overline{x} \subseteq \partial h^*(\overline{y}) - A\partial g^*(-A^T \overline{y}) \\ &\subseteq \partial[h^* + g^* \circ (-A^T)](\overline{y}) \end{aligned}$$

Thus, $\overline{y} := \nabla h(A\overline{x})$ solves the dual problem.

Case study cont.

Thus:

- Suppose $\{x^k\}$ converges to / clusters at a minimizer $\overline{x}$ of (5).
- Now, $\nabla h(A\overline{x})$ solves (6). Hence, $\{\nabla h(Ax^k)\}$ will be a sequence that converges to / clusters at a solution of (6).

Case study cont.

Thus:

- Suppose $\{x^k\}$ converges to / clusters at a minimizer $\overline{x}$ of (5).
- Now, $\nabla h(A\overline{x})$ solves (6). Hence, $\{\nabla h(Ax^k)\}$ will be a sequence that converges to / clusters at a solution of (6).

Next, we define:

$$u^k = \min\left\{1, \frac{\mu}{\|A^T\nabla h(Ax^k)\|_\infty}\right\}\nabla h(Ax^k).$$

Then:

(a) $u^k \to \nabla h(A\overline{x})$; (b) $\|A^T u^k\|_\infty \le \mu$ for all k.

Thus, $\{u^k\}$ is a maximizing sequence of (6). Hence,

$$\inf F \leftarrow F(x^k) \ge \inf F \ge D(u^k) \to \inf F.$$

References

S. R. Becker, E. J. Candès and M. C. Grant.
Templates for convex cone problems with applications to sparse signal recovery.
Math. Program. Comput. 3, pp. 165–218, 2011.

B. O'Donoghue and E. J. Candès.
Adaptive restart for accelerated gradient schemes.
Found. Comput. Math. 15, pp. 715–732, 2015.

S. Foucart and H. Rauhut.
A Mathematical Introduction to Compressive Sensing.
Springer, New York (2013).

P. L. Lions and B. Mercier.
Splitting algorithms for the sum of two nonlinear operators.
SIAM J. Numer. Anal. 16, pp. 964–979, 1979.

References

Y. Nesterov.
A method for solving a convex programming problem with convergence rate $O(1/k^2)$.
Soviet Math. Dokl. 27, pp. 372–376, 1983.

Y. Nesterov.
Introductory Lectures on Convex Optimization: A Basic Course.
Kluwer Academic Publishers, Boston (2004).

R. Tibshirani.
Regression shrinkage and selection via the Lasso.
J. R. Statist. Soc. B 58, pp. 267–288, 1996.

P. Tseng.
Applications of a splitting algorithm to decomposition in convex programming and variational inequalities.
SIAM J. Control Optim. 29, pp. 119–138, 1991.

References

P. Tseng.
On accelerated proximal gradient methods for convex-concave optimization.
Unpublished manuscript. Available at https://www.mit.edu/~dimitrib/PTseng/papers/apgm.pdf.

P. Tseng.
Approximation accuracy, gradient methods and error bound for structured convex optimization.
Math. Program. 125, pp. 263–295, 2010.

B. Wen, X. Chen and T. K. Pong.
Linear convergence of proximal gradient algorithm with extrapolation for a class of nonconvex nonsmooth minimization problems.
SIAM J. Optim. 27, pp. 124–145, 2017.