Proximal gradient methods

Yuxin Chen

Wharton Statistics & Data Science, Fall 2023

Outline

- Proximal gradient descent for composite functions
- Proximal mapping / operator
- Convergence analysis

Proximal gradient descent for composite functions

Composite models

$$\begin{aligned} \text{minimize}_{\boldsymbol{x}} \quad & F(\boldsymbol{x}) := f(\boldsymbol{x}) + h(\boldsymbol{x}) \\ \text{subject to} \quad & \boldsymbol{x} \in \mathbb{R}^n \end{aligned}$$

- f: convex and smooth
- h: convex (may not be differentiable)

let $F^{\mathsf{opt}} := \min_{\boldsymbol{x}} F(\boldsymbol{x})$ be the optimal cost

Examples

- ℓ_1 regularized minimization

$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}) + \underbrace{\|\boldsymbol{x}\|_1}_{h(\boldsymbol{x}):\, \ell_1 \text{ norm}}$$

  - use ℓ_1 regularization to promote sparsity

- nuclear norm regularized minimization

$$\text{minimize}_{\boldsymbol{X}} \quad f(\boldsymbol{X}) + \underbrace{\|\boldsymbol{X}\|_*}_{h(\boldsymbol{X}):\, \text{nuclear norm}}$$

  - use nuclear norm regularization to promote low-rank structure

A proximal view of gradient descent

To motivate proximal gradient methods, we first revisit gradient descent

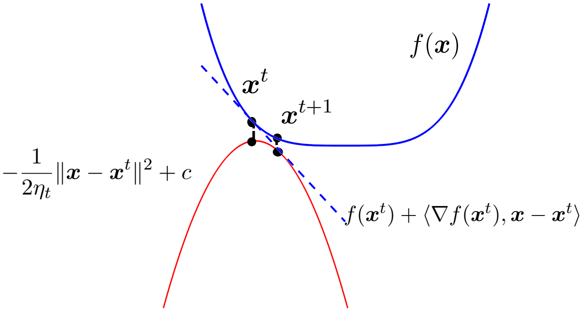
$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t \nabla f(\boldsymbol{x}^t)$$

$$\Updownarrow$$

$$\boldsymbol{x}^{t+1} = \arg\min_{\boldsymbol{x}} \Big\{ \underbrace{f(\boldsymbol{x}^t) + \langle \nabla f(\boldsymbol{x}^t), \boldsymbol{x} - \boldsymbol{x}^t \rangle}_{\text{first-order approximation}} + \underbrace{\frac{1}{2\eta_t} \|\boldsymbol{x} - \boldsymbol{x}^t\|_2^2}_{\text{proximal term}} \Big\}$$

A proximal view of gradient descent

$$\boldsymbol{x}^{t+1} = \arg\min_{\boldsymbol{x}} \left\{ f(\boldsymbol{x}^t) + \langle \nabla f(\boldsymbol{x}^t), \boldsymbol{x} - \boldsymbol{x}^t \rangle + \frac{1}{2\eta_t} \|\boldsymbol{x} - \boldsymbol{x}^t\|_2^2 \right\}$$



By the optimality condition, $\boldsymbol{x}^{t+1}$ is the point where $f(\boldsymbol{x}^t) + \langle \nabla f(\boldsymbol{x}^t), \boldsymbol{x} - \boldsymbol{x}^t \rangle$ and $-\frac{1}{2\eta_t}\|\boldsymbol{x} - \boldsymbol{x}^t\|_2^2$ have the same slope

How about projected gradient descent?

$$\boldsymbol{x}^{t+1} = \mathcal{P}_{\mathcal{C}}\big(\boldsymbol{x}^t - \eta_t \nabla f(\boldsymbol{x}^t)\big)$$

$$\Updownarrow$$

$$\begin{aligned}
\boldsymbol{x}^{t+1} &= \arg\min_{\boldsymbol{x}} \left\{ f(\boldsymbol{x}^t) + \langle \nabla f(\boldsymbol{x}^t), \boldsymbol{x} - \boldsymbol{x}^t \rangle + \frac{1}{2\eta_t} \|\boldsymbol{x} - \boldsymbol{x}^t\|_2^2 + \mathbb{1}_{\mathcal{C}}(\boldsymbol{x}) \right\} \\
&= \arg\min_{\boldsymbol{x}} \left\{ \frac{1}{2} \big\| \boldsymbol{x} - (\boldsymbol{x}^t - \eta_t \nabla f(\boldsymbol{x}^t)) \big\|_2^2 + \eta_t \, \mathbb{1}_{\mathcal{C}}(\boldsymbol{x}) \right\} \qquad (6.1)
\end{aligned}$$

where $\mathbb{1}_{\mathcal{C}}(\boldsymbol{x}) = \begin{cases} 0, & \text{if } \boldsymbol{x} \in \mathcal{C} \\ \infty, & \text{else} \end{cases}$

Proximal operator

Define the proximal operator

$$\mathsf{prox}_h(\boldsymbol{x}) \ := \ \arg\min_{\boldsymbol{z}} \left\{ \frac{1}{2} \left\| \boldsymbol{z} - \boldsymbol{x} \right\|_2^2 + h(\boldsymbol{z}) \right\}$$

for any convex function h

This allows one to express projected GD update (6.1) as

$$\boldsymbol{x}^{t+1} = \mathsf{prox}_{\eta_t \mathbb{1}_{\mathcal{C}}}\left(\boldsymbol{x}^t - \eta_t \nabla f(\boldsymbol{x}^t)\right) \tag{6.2}$$

Proximal gradient methods

One can generalize (6.2) to accommodate more general h

Algorithm 6.1 Proximal gradient algorithm

1: **for** $t = 0, 1, \cdots$ **do**
2: $\quad \boldsymbol{x}^{t+1} = \mathsf{prox}_{\eta_t h}\big(\boldsymbol{x}^t - \eta_t \nabla f(\boldsymbol{x}^t)\big)$

- alternates between gradient updates on f and proximal minimization on h
- useful if prox_h is inexpensive

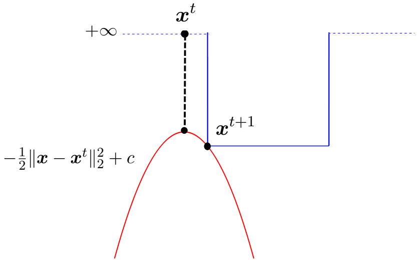
Proximal mapping / operator

Why consider proximal operators?

$$\mathsf{prox}_h(\boldsymbol{x}) \ := \ \arg\min_{\boldsymbol{z}} \left\{ \frac{1}{2} \|\boldsymbol{z} - \boldsymbol{x}\|_2^2 + h(\boldsymbol{z}) \right\}$$

- well-defined under very general conditions (including nonsmooth convex functions)
- can be evaluated efficiently for many widely used functions (in particular, regularizers)
- this abstraction is conceptually and mathematically simple, and covers many well-known optimization algorithms

Example: indicator functions



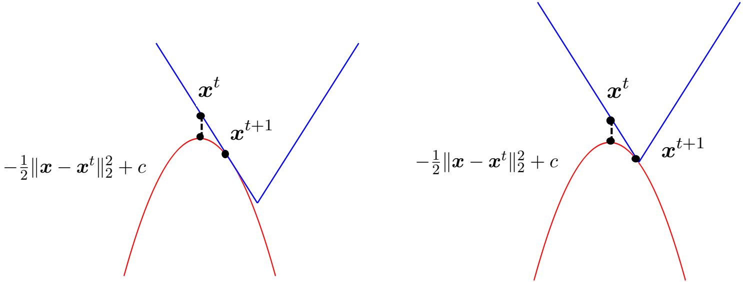
If $h = \mathbb{1}_{\mathcal{C}}$ is the "indicator" function

$$h(\boldsymbol{x}) = \begin{cases} 0, & \text{if } \boldsymbol{x} \in \mathcal{C} \\ \infty, & \text{else} \end{cases}$$

then

$$\text{prox}_h(\boldsymbol{x}) = \arg\min_{\boldsymbol{z}\in\mathcal{C}} \|\boldsymbol{z}-\boldsymbol{x}\|_2 \quad \text{(Euclidean projection)}$$

Example: ℓ_1 norm



If $h(\boldsymbol{x}) = \lambda\|\boldsymbol{x}\|_1$, then

$$(\mathsf{prox}_{\lambda h}(\boldsymbol{x}))_i = \psi_{\mathrm{st}}(x_i; \lambda) \qquad \text{(soft-thresholding)}$$

where $\psi_{\mathrm{st}}(x) = \begin{cases} x - \lambda, & \text{if } x > \lambda \\ x + \lambda, & \text{if } x < -\lambda \\ 0, & \text{else} \end{cases}$

Basic rules

- If $f(\boldsymbol{x}) = ag(\boldsymbol{x}) + b$ with $a > 0$, then

$$\mathsf{prox}_f(\boldsymbol{x}) = \mathsf{prox}_{ag}(\boldsymbol{x})$$

- **affine addition:** if $f(\boldsymbol{x}) = g(\boldsymbol{x}) + \boldsymbol{a}^\top \boldsymbol{x} + b$, then

$$\mathsf{prox}_f(\boldsymbol{x}) = \mathsf{prox}_g(\boldsymbol{x} - \boldsymbol{a})$$

Basic rules

- **quadratic addition:** if $f(\boldsymbol{x}) = g(\boldsymbol{x}) + \frac{\rho}{2}\|\boldsymbol{x} - \boldsymbol{a}\|_2^2$, then

$$\mathsf{prox}_f(\boldsymbol{x}) = \mathsf{prox}_{\frac{1}{1+\rho}g}\left(\frac{1}{1+\rho}\boldsymbol{x} + \frac{\rho}{1+\rho}\boldsymbol{a}\right)$$

- **scaling and translation:** if $f(\boldsymbol{x}) = g(a\boldsymbol{x} + \boldsymbol{b})$ with $a \neq 0$, then

$$\mathsf{prox}_f(\boldsymbol{x}) = \frac{1}{a}\Big(\mathsf{prox}_{a^2 g}(a\boldsymbol{x} + \boldsymbol{b}) - \boldsymbol{b}\Big) \quad \text{(homework)}$$

Proof for quadratic addition

$$
\begin{aligned}
\mathsf{prox}_f(\boldsymbol{x}) &= \arg\min_{\boldsymbol{x}} \left\{ \frac{1}{2}\|\boldsymbol{z}-\boldsymbol{x}\|_2^2 + g(\boldsymbol{z}) + \frac{\rho}{2}\|\boldsymbol{z}-\boldsymbol{a}\|_2^2 \right\} \\
&= \arg\min_{\boldsymbol{x}} \left\{ \frac{1+\rho}{2}\|\boldsymbol{z}\|_2^2 - \langle \boldsymbol{z}, \boldsymbol{x}+\rho\boldsymbol{a}\rangle + g(\boldsymbol{z}) \right\} \\
&= \arg\min_{\boldsymbol{x}} \left\{ \frac{1}{2}\|\boldsymbol{z}\|_2^2 - \frac{1}{1+\rho}\langle \boldsymbol{z}, \boldsymbol{x}+\rho\boldsymbol{a}\rangle + \frac{1}{1+\rho}g(\boldsymbol{z}) \right\} \\
&= \arg\min_{\boldsymbol{x}} \left\{ \frac{1}{2}\left\|\boldsymbol{z} - \left(\frac{1}{1+\rho}\boldsymbol{x} + \frac{\rho}{1+\rho}\boldsymbol{a}\right)\right\|_2^2 + \frac{1}{1+\rho}g(\boldsymbol{z}) \right\} \\
&= \mathsf{prox}_{\frac{1}{1+\rho}g}\left(\frac{1}{1+\rho}\boldsymbol{x} + \frac{\rho}{1+\rho}\boldsymbol{a}\right)
\end{aligned}
$$

Basic rules

- **orthogonal mapping:** if $f(\boldsymbol{x}) = g(\boldsymbol{Q}\boldsymbol{x})$ with $\boldsymbol{Q}$ orthogonal ($\boldsymbol{Q}\boldsymbol{Q}^\top = \boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}$), then

$$\mathrm{prox}_f(\boldsymbol{x}) = \boldsymbol{Q}^\top \mathrm{prox}_g(\boldsymbol{Q}\boldsymbol{x}) \quad \text{(homework)}$$

- **orthogonal affine mapping:** if $f(\boldsymbol{x}) = g(\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{b})$ with $\underbrace{\boldsymbol{Q}\boldsymbol{Q}^\top = \alpha^{-1}\boldsymbol{I}}_{\text{does not require } \boldsymbol{Q}^\top\boldsymbol{Q}=\alpha^{-1}\boldsymbol{I}}$, then

$$\mathrm{prox}_f(\boldsymbol{x}) = \left(\boldsymbol{I} - \alpha\boldsymbol{Q}^\top\boldsymbol{Q}\right)\boldsymbol{x} + \alpha\boldsymbol{Q}^\top\left(\mathrm{prox}_{\alpha^{-1}g}(\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{b}) - \boldsymbol{b}\right)$$

  - for general $\boldsymbol{Q}$, it is not easy to derive prox_f from prox_g

Basic rules

- **norm composition:** if $f(\boldsymbol{x}) = g(\|\boldsymbol{x}\|_2)$ with domain$(g) = [0, \infty)$, then

$$\mathsf{prox}_f(\boldsymbol{x}) = \mathsf{prox}_g(\|\boldsymbol{x}\|_2)\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2} \quad \forall \boldsymbol{x} \neq \boldsymbol{0}$$

Proof for norm composition

Observe that

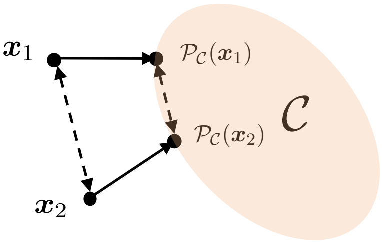
$$
\begin{aligned}
&\min_{\boldsymbol{z}} \left\{ f(\boldsymbol{z}) + \frac{1}{2}\|\boldsymbol{z}-\boldsymbol{x}\|_2^2 \right\} \\
&= \min_{\boldsymbol{z}} \left\{ g(\|\boldsymbol{z}\|_2) + \frac{1}{2}\|\boldsymbol{z}\|_2^2 - \boldsymbol{z}^\top \boldsymbol{x} + \frac{1}{2}\|\boldsymbol{x}\|_2^2 \right\} \\
&= \min_{\alpha \geq 0} \min_{\boldsymbol{z}:\|\boldsymbol{z}\|_2=\alpha} \left\{ g(\alpha) + \frac{1}{2}\alpha^2 - \boldsymbol{z}^\top \boldsymbol{x} + \frac{1}{2}\|\boldsymbol{x}\|_2^2 \right\} \\
&= \min_{\alpha \geq 0} \left\{ g(\alpha) + \frac{1}{2}\alpha^2 - \alpha\|\boldsymbol{x}\|_2 + \frac{1}{2}\|\boldsymbol{x}\|_2^2 \right\} \qquad \text{(Cauchy-Schwarz)} \\
&= \min_{\alpha \geq 0} \left\{ g(\alpha) + \frac{1}{2}\left(\alpha - \|\boldsymbol{x}\|_2\right)^2 \right\}
\end{aligned}
$$

From the above calculation, we know the optimal point is

$$
\alpha^* = \mathsf{prox}_g(\|\boldsymbol{x}\|_2) \qquad \text{and} \qquad \boldsymbol{z}^* = \alpha^* \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2} = \mathsf{prox}_g(\|\boldsymbol{x}\|_2)\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2},
$$

thus concluding proof

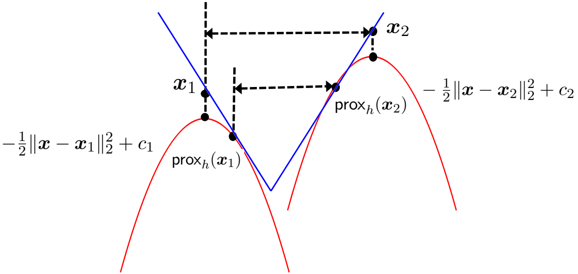
Nonexpansiveness of proximal operators



Recall that when $h(\boldsymbol{x}) = \mathbb{1}_{\mathcal{C}}(\boldsymbol{x})$, $\mathrm{prox}_h(\boldsymbol{x})$ is the Euclidean projection $\mathcal{P}_{\mathcal{C}}$ onto $\mathcal{C}$, which is nonexpansive for convex $\mathcal{C}$:

$$\|\mathcal{P}_{\mathcal{C}}(\boldsymbol{x}_1) - \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}_2)\|_2 \le \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2$$

Nonexpansiveness of proximal operators



in some sense, proximal operator behaves like projection

Fact 6.1

- **(firm nonexpansiveness)**

$$\langle \mathrm{prox}_h(\boldsymbol{x}_1)-\mathrm{prox}_h(\boldsymbol{x}_2), \boldsymbol{x}_1-\boldsymbol{x}_2\rangle \geq \|\mathrm{prox}_h(\boldsymbol{x}_1)-\mathrm{prox}_h(\boldsymbol{x}_2)\|_2^2$$

- **(nonexpansiveness)**

$$\|\mathrm{prox}_h(\boldsymbol{x}_1)-\mathrm{prox}_h(\boldsymbol{x}_2)\|_2 \leq \|\boldsymbol{x}_1-\boldsymbol{x}_2\|_2$$

Proof of Fact 6.1

Let $\boldsymbol{z}_1 = \mathrm{prox}_h(\boldsymbol{x}_1)$ and $\boldsymbol{z}_2 = \mathrm{prox}_h(\boldsymbol{x}_2)$. Subgradient characterizations of $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ read

$$\boldsymbol{x}_1 - \boldsymbol{z}_1 \in \partial h(\boldsymbol{z}_1) \quad \text{and} \quad \boldsymbol{x}_2 - \boldsymbol{z}_2 \in \partial h(\boldsymbol{z}_2)$$

The nonexpansiveness claim $\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|_2 \leq \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2$ would follow if

$$\underbrace{(\boldsymbol{x}_1 - \boldsymbol{x}_2)^\top (\boldsymbol{z}_1 - \boldsymbol{z}_2) \geq \|\boldsymbol{z}_1 - \boldsymbol{z}_2\|_2^2}_{\text{firm nonexpansiveness}} \quad \text{(together with Cauchy-Schwarz)}$$

$$\Longleftarrow \quad (\boldsymbol{x}_1 - \boldsymbol{z}_1 - \boldsymbol{x}_2 + \boldsymbol{z}_2)^\top (\boldsymbol{z}_1 - \boldsymbol{z}_2) \geq 0$$

$$\overset{\text{add these inequalities}}{\Longleftarrow} \quad \begin{cases} h(\boldsymbol{z}_2) \geq h(\boldsymbol{z}_1) + \langle \underbrace{\boldsymbol{x}_1 - \boldsymbol{z}_1}_{\in \partial h(\boldsymbol{z}_1)}, \ \boldsymbol{z}_2 - \boldsymbol{z}_1 \rangle \\ h(\boldsymbol{z}_1) \geq h(\boldsymbol{z}_2) + \langle \underbrace{\boldsymbol{x}_2 - \boldsymbol{z}_2}_{\in \partial h(\boldsymbol{z}_2)}, \ \boldsymbol{z}_1 - \boldsymbol{z}_2 \rangle \end{cases}$$

Resolvent of subdifferential operator

One can interpret prox via the resolvant of subdifferential operator

Fact 6.2

Suppose that f is convex. Then one can write

$$\boldsymbol{z} = \text{prox}_f(\boldsymbol{x}) \quad \Longleftrightarrow \quad \boldsymbol{z} = \underbrace{(\mathcal{I} + \partial f)^{-1}}_{\text{resolvent of operator } \partial f}(\boldsymbol{x})$$

where $\mathcal{I}$ is the identity mapping

Justification of Fact 6.2

$$\boldsymbol{z} = \arg\min_{\boldsymbol{u}} \left\{ f(\boldsymbol{u}) + \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{x}\|_2^2 \right\}$$

$$\Longleftrightarrow \quad \boldsymbol{0} \in \partial f(\boldsymbol{z}) + \boldsymbol{z} - \boldsymbol{x} \qquad \text{(optimality condition)}$$

$$\Longleftrightarrow \quad \boldsymbol{x} \in \left(\mathcal{I} + \partial f\right)(\boldsymbol{z})$$

$$\Longleftrightarrow \quad \boldsymbol{z} = \left(\mathcal{I} + \partial f\right)^{-1}(\boldsymbol{x})$$

Moreau decomposition

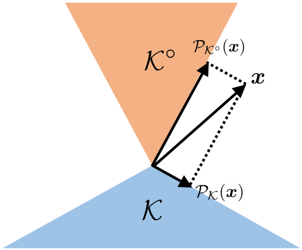
Fact 6.3

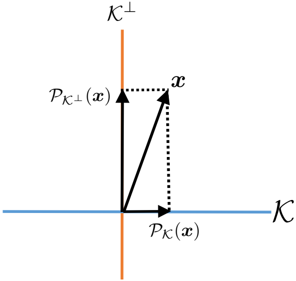
Suppose f is closed and convex, and $f^*(\boldsymbol{x}) := \sup_{\boldsymbol{z}}\{\langle \boldsymbol{x}, \boldsymbol{z}\rangle - f(\boldsymbol{z})\}$ is the convex conjugate of f. Then

$$\boldsymbol{x} = \mathsf{prox}_f(\boldsymbol{x}) + \mathsf{prox}_{f^*}(\boldsymbol{x})$$

- key relationship between proximal mapping and duality
- generalization of orthogonal decomposition

Moreau decomposition for convex cones





When $\mathcal{K}$ is a closed convex cone, $(\mathbb{1}_{\mathcal{K}})^*(\boldsymbol{x}) = \mathbb{1}_{\mathcal{K}^\circ}(\boldsymbol{x})$ (exercise) with $\mathcal{K}^\circ := \{\boldsymbol{x} \mid \langle \boldsymbol{x}, \boldsymbol{z} \rangle \leq 0, \forall \boldsymbol{z} \in \mathcal{K}\}$ polar cone of $\mathcal{K}$. This gives

$$\boldsymbol{x} = \mathcal{P}_{\mathcal{K}}(\boldsymbol{x}) + \mathcal{P}_{\mathcal{K}^\circ}(\boldsymbol{x})$$

- a special case: if $\mathcal{K}$ is a subspace, then $\mathcal{K}^\circ = \mathcal{K}^\perp$, and hence

$$\boldsymbol{x} = \mathcal{P}_{\mathcal{K}}(\boldsymbol{x}) + \mathcal{P}_{\mathcal{K}^\perp}(\boldsymbol{x})$$

Proof of Fact 6.3

Let $\boldsymbol{u} = \text{prox}_f(\boldsymbol{x})$, then from the optimality condition we know that

$$\boldsymbol{x} - \boldsymbol{u} \in \partial f(\boldsymbol{u}).$$

This together with conjugate subgradient theorem (homework) yields

$$\boldsymbol{u} \in \partial f^*(\boldsymbol{x} - \boldsymbol{u})$$

In view of the optimality condition, this means

$$\boldsymbol{x} - \boldsymbol{u} = \text{prox}_{f^*}(\boldsymbol{x})$$

$$\Longrightarrow \qquad \boldsymbol{x} = \boldsymbol{u} + (\boldsymbol{x} - \boldsymbol{u}) = \text{prox}_f(\boldsymbol{x}) + \text{prox}_{f^*}(\boldsymbol{x})$$

Example: prox of support function

For any closed and convex set $\mathcal{C}$, the *support function* $S_{\mathcal{C}}$ is defined as $S_{\mathcal{C}}(\boldsymbol{x}) = \sup_{\boldsymbol{z} \in \mathcal{C}} \langle \boldsymbol{x}, \boldsymbol{z} \rangle$. Then

$$\mathsf{prox}_{S_{\mathcal{C}}}(\boldsymbol{x}) = \boldsymbol{x} - \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}) \tag{6.3}$$

Proof: First of all, it is easy to verify that (exercise)

$$S_{\mathcal{C}}^*(\boldsymbol{x}) = \mathbb{1}_{\mathcal{C}}(\boldsymbol{x})$$

Then the Moreau decomposition gives

$$\begin{aligned} \mathsf{prox}_{S_{\mathcal{C}}}(\boldsymbol{x}) &= \boldsymbol{x} - \mathsf{prox}_{S_{\mathcal{C}}^*}(\boldsymbol{x}) \\ &= \boldsymbol{x} - \mathsf{prox}_{\mathbb{1}_{\mathcal{C}}}(\boldsymbol{x}) \\ &= \boldsymbol{x} - \mathcal{P}_{\mathcal{C}}(\boldsymbol{x}) \end{aligned}$$

□

Example: ℓ_∞ norm

$$\text{prox}_{\|\cdot\|_\infty}(\boldsymbol{x}) = \boldsymbol{x} - \mathcal{P}_{\mathcal{B}_{\|\cdot\|_1}}(\boldsymbol{x})$$

where $\mathcal{B}_{\|\cdot\|_1} := \{\boldsymbol{z} \mid \|\boldsymbol{z}\|_1 \leq 1\}$ is unit ℓ_1 ball

Remark: projection onto ℓ_1 ball can be computed efficiently

Proof: Since $\|\boldsymbol{x}\|_\infty = \sup_{\boldsymbol{z}:\|\boldsymbol{z}\|_1 \leq 1} \langle \boldsymbol{x}, \boldsymbol{z} \rangle = S_{\mathcal{B}_{\|\cdot\|_1}}(\boldsymbol{x})$, we can invoke (6.3) to arrive at

$$\text{prox}_{\|\cdot\|_\infty}(\boldsymbol{x}) = \text{prox}_{S_{\mathcal{B}_{\|\cdot\|_1}}}(\boldsymbol{x}) = \boldsymbol{x} - \mathcal{P}_{\mathcal{B}_{\|\cdot\|_1}}(\boldsymbol{x})$$

□

Example: max function

Let $g(\boldsymbol{x}) = \max\{x_1, \cdots, x_n\}$, then

$$\mathsf{prox}_g(\boldsymbol{x}) = \boldsymbol{x} - \mathcal{P}_\Delta(\boldsymbol{x})$$

where $\Delta := \{\boldsymbol{z} \in \mathbb{R}^n_+ \mid \boldsymbol{1}^\top \boldsymbol{z} = 1\}$ is probability simplex

Remark: projection onto Δ can be computed efficiently

Proof: Since $g(\boldsymbol{x}) = \max\{x_1, \cdots, x_n\} = S_\Delta(\boldsymbol{x})$ (support function of Δ), we can invoke (6.3) to reach

$$\mathsf{prox}_g(\boldsymbol{x}) = \boldsymbol{x} - \mathcal{P}_\Delta(\boldsymbol{x})$$

$\square$

Extended Moreau decomposition

A useful extension (homework):

Fact 6.4

Suppose f is closed and convex, and $\lambda > 0$. Then

$$\boldsymbol{x} = \mathsf{prox}_{\lambda f}(\boldsymbol{x}) + \lambda \mathsf{prox}_{\frac{1}{\lambda} f^*}(\boldsymbol{x}/\lambda)$$

Convergence analysis

Cost monotonicity

The objective value is *non-increasing* in t:

Lemma 6.5

Suppose f *is convex and* L*-smooth. If* $\eta_t \equiv 1/L$*, then*

$$F(\boldsymbol{x}^{t+1}) \leq F(\boldsymbol{x}^t)$$

- different from subgradient methods (for which the objective values might be non-monotonic in t)
- constant stepsizes are recommended when f is convex and smooth

Proof of cost monotonicity

Main pillar: a fundamental inequality

Lemma 6.6

Let $\boldsymbol{y}^{+} = \mathrm{prox}_{\frac{1}{L}h}(\boldsymbol{y} - \frac{1}{L}\nabla f(\boldsymbol{y}))$, *then*

$$F(\boldsymbol{y}^{+}) - F(\boldsymbol{x}) \leq \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 - \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}^{+}\|_2^2 - \underbrace{g(\boldsymbol{x}, \boldsymbol{y})}_{\geq 0 \text{ by convexity}}$$

where $g(\boldsymbol{x}, \boldsymbol{y}) := f(\boldsymbol{x}) - f(\boldsymbol{y}) - \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle$

Take $\boldsymbol{x} = \boldsymbol{y} = \boldsymbol{x}^{t}$ (and hence $\boldsymbol{y}^{+} = \boldsymbol{x}^{t+1}$) to complete the proof

Monotonicity in estimation errors

Proximal gradient iterates are not only monotonic w.r.t. cost, but also monotonic in estimation error

Lemma 6.7

Suppose f is convex and L-smooth. If $\eta_t \equiv 1/L$, then

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2 \leq \|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2$$

Proof: from Lemma 6.6, taking $\boldsymbol{x} = \boldsymbol{x}^*$, $\boldsymbol{y} = \boldsymbol{x}^t$ (and hence $\boldsymbol{y}^+ = \boldsymbol{x}^{t+1}$) yields

$$\underbrace{F(\boldsymbol{x}^{t+1}) - F(\boldsymbol{x}^*)}_{\geq 0} + \underbrace{g(\boldsymbol{x}, \boldsymbol{y})}_{\geq 0} \leq \frac{L}{2}\|\boldsymbol{x}^* - \boldsymbol{x}^t\|_2^2 - \frac{L}{2}\|\boldsymbol{x}^* - \boldsymbol{x}^{t+1}\|_2^2$$

which immediately concludes the proof

Proof of Lemma 6.6

Define

$$\phi(\boldsymbol{z}) = f(\boldsymbol{y}) + \langle \nabla f(\boldsymbol{y}), \boldsymbol{z} - \boldsymbol{y} \rangle + \frac{L}{2}\|\boldsymbol{z} - \boldsymbol{y}\|_2^2 + h(\boldsymbol{z})$$

It is easily seen that $\boldsymbol{y}^+ = \arg\min_{\boldsymbol{z}} \phi(\boldsymbol{z})$. Two important properties:

- Since $\phi(\boldsymbol{z})$ is L-strongly convex, one has

$$\phi(\boldsymbol{x}) \geq \phi(\boldsymbol{y}^+) + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}^+\|_2^2$$

  Remark: we are propergating the smoothness of f to the strong convexity of another function ϕ

- From the smoothness condition of f,

$$\phi(\boldsymbol{y}^+) = \underbrace{f(\boldsymbol{y}) + \langle \nabla f(\boldsymbol{y}), \boldsymbol{y}^+ - \boldsymbol{y} \rangle + \frac{L}{2}\|\boldsymbol{y}^+ - \boldsymbol{y}\|_2^2}_{\text{upper bound on } f(\boldsymbol{y}^+)} + h(\boldsymbol{y}^+)$$

$$\geq f(\boldsymbol{y}^+) + h(\boldsymbol{y}^+) = F(\boldsymbol{y}^+)$$

Proof of Lemma 6.6 (cont.)

Taken collectively, these yield

$$\phi(\boldsymbol{x}) \geq F(\boldsymbol{y}^+) + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}^+\|_2^2,$$

which together with the definition of $\phi(\boldsymbol{x})$ gives

$$\underbrace{f(\boldsymbol{y}) + \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle + h(\boldsymbol{x})}_{=f(\boldsymbol{x})+h(\boldsymbol{x})-g(\boldsymbol{x},\boldsymbol{y})=F(\boldsymbol{x})-g(\boldsymbol{x},\boldsymbol{y})} + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \geq F(\boldsymbol{y}^+) + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}^+\|_2^2$$

which finishes the proof

Convergence for convex problems

Theorem 6.8 (Convergence of proximal gradient methods for convex problems)

Suppose f is convex and L-smooth. If $\eta_t \equiv 1/L$, then

$$F(\boldsymbol{x}^t) - F^{\mathsf{opt}} \le \frac{L\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2}{2t}$$

- achieves better iteration complexity (i.e. $O(1/\varepsilon)$) than subgradient method (i.e. $O(1/\varepsilon^2)$)
- fast if prox can be efficiently implemented

Proof of Theorem 6.8

With Lemma 6.6 in mind, set $\boldsymbol{x} = \boldsymbol{x}^*$, $\boldsymbol{y} = \boldsymbol{x}^t$ to obtain

$$F(\boldsymbol{x}^{t+1}) - F(\boldsymbol{x}^*) \le \frac{L}{2}\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2 - \frac{L}{2}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2 - \underbrace{g(\boldsymbol{x}^*, \boldsymbol{x}^t)}_{\ge 0 \text{ by convexity}}$$

$$\le \frac{L}{2}\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2 - \frac{L}{2}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2$$

Apply it recursively and add up all inequalities to get

$$\sum_{k=0}^{t-1} \Big(F(\boldsymbol{x}^{k+1}) - F(\boldsymbol{x}^*) \Big) \le \frac{L}{2}\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2 - \frac{L}{2}\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2$$

This combined with monotonicity of $F(\boldsymbol{x}^t)$ (cf. Lemma 6.6) yields

$$F(\boldsymbol{x}^t) - F(\boldsymbol{x}^*) \le \frac{\frac{L}{2}\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2}{t}$$

Convergence for strongly convex problems

Theorem 6.9 (Convergence of proximal gradient methods for strongly convex problems)

Suppose f is μ-strongly convex and L-smooth. If $\eta_t \equiv 1/L$, then

$$\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2 \leq \left(1 - \frac{\mu}{L}\right)^t \|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2$$

- linear convergence: attains ε accuracy within $O(\log \frac{1}{\varepsilon})$ iterations

Proof of Theorem 6.9

Taking $\boldsymbol{x} = \boldsymbol{x}^*$, $\boldsymbol{y} = \boldsymbol{x}^t$ (and hence $\boldsymbol{y}^+ = \boldsymbol{x}^{t+1}$) in Lemma 6.6 gives

$$F(\boldsymbol{x}^{t+1}) - F(\boldsymbol{x}^*) \leq \frac{L}{2}\|\boldsymbol{x}^* - \boldsymbol{x}^t\|_2^2 - \frac{L}{2}\|\boldsymbol{x}^* - \boldsymbol{x}^{t+1}\|_2^2 - \underbrace{g(\boldsymbol{x}^*, \boldsymbol{x}^t)}_{\geq \frac{\mu}{2}\|\boldsymbol{x}^* - \boldsymbol{x}^t\|_2^2}$$

$$\leq \frac{L-\mu}{2}\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2 - \frac{L}{2}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2$$

This taken collectively with $F(\boldsymbol{x}^{t+1}) - F(\boldsymbol{x}^*) \geq 0$ yields

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2 \leq \left(1 - \frac{\mu}{L}\right)\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2$$

Applying it recursively concludes the proof

Numerical example: LASSO

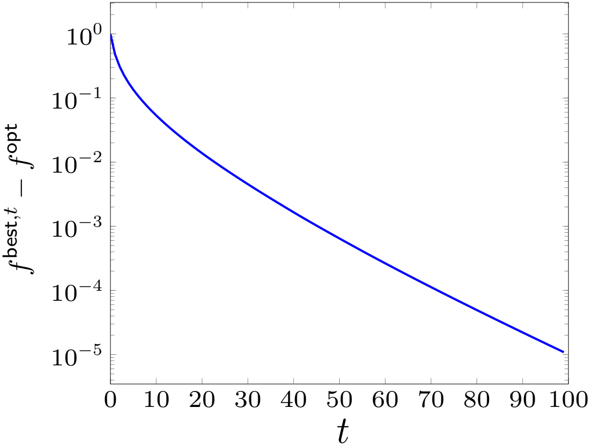
taken from UCLA EE236C

$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 + \|\boldsymbol{x}\|_1$$

with i.i.d. Gaussian $\boldsymbol{A} \in \mathbb{R}^{2000\times 1000}$, $\eta_t = 1/L$, $L = \lambda_{\max}(\boldsymbol{A}^\top \boldsymbol{A})$

Numerical example: LASSO

taken from UCLA EE236C



Backtracking line search

Recall that for the unconstrained case, backtracking line search is based on a sufficient decrease criterion

$$f\big(\boldsymbol{x}^t - \eta \nabla f(\boldsymbol{x}^t)\big) \le f(\boldsymbol{x}^t) - \frac{\eta}{2}\|\nabla f(\boldsymbol{x}^t)\|_2^2$$

Backtracking line search

Recall that for the unconstrained case, backtracking line search is based on a sufficient decrease criterion

$$f\big(\boldsymbol{x}^t - \eta \nabla f(\boldsymbol{x}^t)\big) \le f(\boldsymbol{x}^t) - \frac{\eta}{2}\|\nabla f(\boldsymbol{x}^t)\|_2^2$$

As a result, this is equivalent to updating $\eta_t = 1/L_t$ until

$$\begin{aligned} f\big(\boldsymbol{x}^t - \eta_t \nabla f(\boldsymbol{x}^t)\big) &\le f(\boldsymbol{x}^t) - \frac{1}{L_t}\langle \nabla f(\boldsymbol{x}^t), \nabla f(\boldsymbol{x}^t)\rangle + \frac{1}{2L_t}\|\nabla f(\boldsymbol{x}^t)\|_2^2 \\ &= f(\boldsymbol{x}^t) - \langle \nabla f(\boldsymbol{x}^t), \boldsymbol{x}^t - \boldsymbol{x}^{t+1}\rangle + \frac{L_t}{2}\|\boldsymbol{x}^t - \boldsymbol{x}^{t+1}\|_2^2 \end{aligned}$$

Backtracking line search

Let $\mathcal{T}_L(\boldsymbol{x}) := \text{prox}_{\frac{1}{L}h}(\boldsymbol{x} - \frac{1}{L}\nabla f(\boldsymbol{x}))$:

Algorithm 6.2 Backtracking line search for proximal gradient methods

1: Initialize $\eta = 1$, $0 < \alpha \leq 1/2$, $0 < \beta < 1$

2: **while** $f\big(\mathcal{T}_{L_t}(\boldsymbol{x}^t)\big) > f(\boldsymbol{x}^t) - \langle \nabla f(\boldsymbol{x}^t), \boldsymbol{x}^t - \mathcal{T}_{L_t}(\boldsymbol{x}^t)\rangle + \frac{L_t}{2}\big\|\mathcal{T}_{L_t}(\boldsymbol{x}^t) - \boldsymbol{x}^t\big\|_2^2$ **do**

3: $\quad L_t \leftarrow \frac{1}{\beta}L_t$ (or $\frac{1}{L_t} \leftarrow \beta\frac{1}{L_t}$)

- here, $\frac{1}{L_t}$ corresponds to η_t, and $\mathcal{T}_{L_t}(\boldsymbol{x}^t)$ generalizes $\boldsymbol{x}^{t+1}$

Summary: proximal gradient methods

	stepsize rule	convergence rate	iteration complexity
convex & smooth (w.r.t. f) problems	$\eta_t = \frac{1}{L}$	$O\left(\frac{1}{t}\right)$	$O\left(\frac{1}{\varepsilon}\right)$
strongly convex & smooth (w.r.t. f) problems	$\eta_t = \frac{1}{L}$	$O\Big((1 - \frac{1}{\kappa})^t\Big)$	$O(\kappa \log \frac{1}{\varepsilon})$

Reference

- "*Proximal algorithms*," N. Parikh and S. Boyd, *Foundations and Trends in Optimization*, 2013.
- "*First-order methods in optimization*," A. Beck, Vol. 25, SIAM, 2017.
- "*Convex optimization and algorithms*," D. Bertsekas, 2015.
- "*Convex optimization: algorithms and complexity*," S. Bubeck, Foundations and trends in machine learning, 2015.
- "*Mathematical optimization, MATH301 lecture notes*," E. Candes, Stanford.
- "*Optimization methods for large-scale systems, EE236C lecture notes*," L. Vandenberghe, UCLA.