

# **Proximal gradient methods**

**Yuanming Shi**

ShanghaiTech University

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

# Examples

---

- $\ell_1$  regularized minimization

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

- use  $\ell_1$  regularization to promote sparsity
- nuclear norm regularized minimization

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

- 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

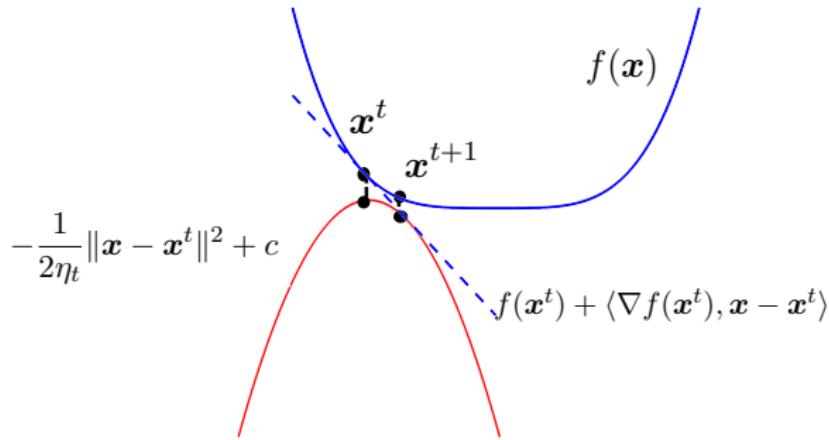
$$\mathbf{x}^{t+1} = \mathbf{x}^t - \eta_t \nabla f(\mathbf{x}^t)$$

$\Updownarrow$

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

# A proximal view of gradient descent

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



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

# How about projected gradient descent?

---

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

$\Updownarrow$

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

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

# Proximal operator

---

Define the proximal operator

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

for any convex function  $h$

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

$$\mathbf{x}^{t+1} = \text{prox}_{\eta_t \mathbb{1}_C} (\mathbf{x}^t - \eta_t \nabla f(\mathbf{x}^t)) \quad (6.2)$$

# Proximal gradient methods

---

One can generalize (6.2) to accommodate more general  $h$

---

## Algorithm 6.1 Proximal gradient algorithm

---

```
1: for  $t = 0, 1, \dots$  do
2:    $\mathbf{x}^{t+1} = \text{prox}_{\eta_t h}(\mathbf{x}^t - \eta_t \nabla f(\mathbf{x}^t))$ 
```

---

- alternates between gradient updates on  $f$  and proximal minimization on  $h$
- useful if  $\text{prox}_h$  is inexpensive

## **Proximal mapping / operator**

# Why consider proximal operators?

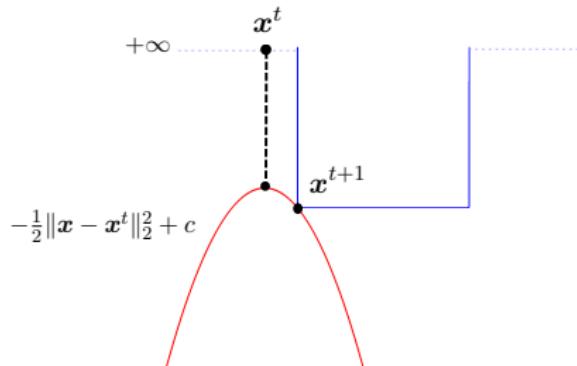
---

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

- 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

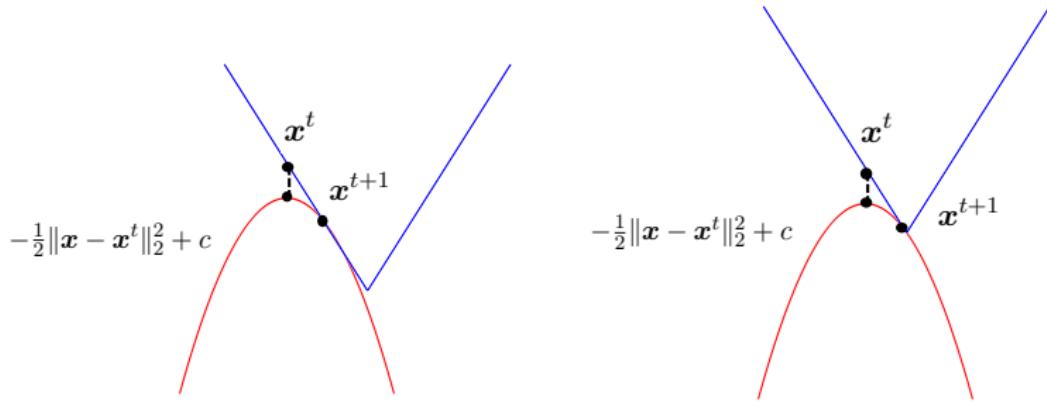
$$h(x) = \begin{cases} 0, & \text{if } x \in \mathcal{C} \\ \infty, & \text{else} \end{cases}$$

then

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

## Example: $\ell_1$ norm

---



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

$$(\text{prox}_{\lambda h}(\mathbf{x}))_i = \psi_{\text{st}}(x_i; \lambda) \quad (\text{soft-thresholding})$$

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

# Basic rules

---

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

$$\text{prox}_f(\mathbf{x}) = \text{prox}_{ag}(\mathbf{x})$$

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

$$\text{prox}_f(\mathbf{x}) = \text{prox}_g(\mathbf{x} - \mathbf{a})$$

# Basic rules

---

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

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

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

$$\text{prox}_f(\mathbf{x}) = \frac{1}{a}\left(\text{prox}_{a^2g}(a\mathbf{x} + \mathbf{b}) - \mathbf{b}\right) \quad (\text{homework})$$

# Proof for quadratic addition

---

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

# Basic rules

---

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

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

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

does not require  $\mathbf{Q}^\top\mathbf{Q}=\alpha^{-1}\mathbf{I}$

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

- for general  $\mathbf{Q}$ , it is not easy to derive  $\text{prox}_f$  from  $\text{prox}_g$

# Basic rules

---

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

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

# Proof for norm composition

---

Observe that

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

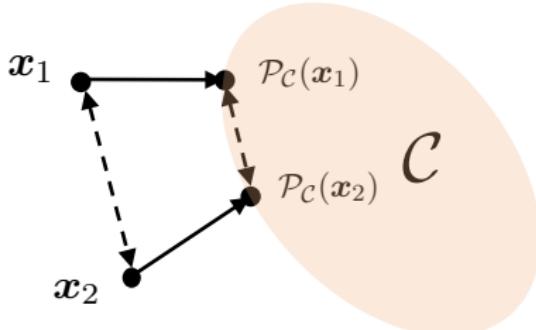
From the above calculation, we know the optimal point is

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

thus concluding proof

# Nonexpansiveness of proximal operators

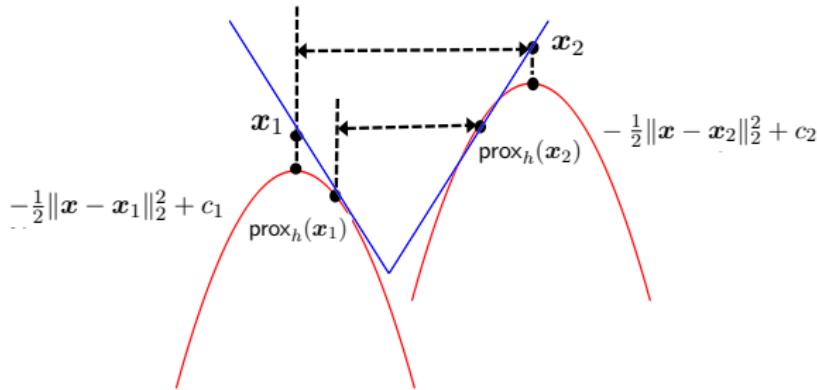
---



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

$$\|\mathcal{P}_{\mathcal{C}}(\mathbf{x}_1) - \mathcal{P}_{\mathcal{C}}(\mathbf{x}_2)\|_2 \leq \|\mathbf{x}_1 - \mathbf{x}_2\|_2$$

# Nonexpansiveness of proximal operators



in some sense,  
proximal operator  
behaves like projection

## Fact 6.1

- **(firm nonexpansiveness)**

$$\langle \text{prox}_h(x_1) - \text{prox}_h(x_2), x_1 - x_2 \rangle \geq \|\text{prox}_h(x_1) - \text{prox}_h(x_2)\|_2^2$$

- **(nonexpansiveness)**

$$\|\text{prox}_h(x_1) - \text{prox}_h(x_2)\|_2 \leq \|x_1 - x_2\|_2$$

## Proof of Fact 6.1

Let  $z_1 = \text{prox}_h(x_1)$  and  $z_2 = \text{prox}_h(x_2)$ . Subgradient characterizations of  $z_1$  and  $z_2$  read

$$x_1 - z_1 \in \partial h(z_1) \quad \text{and} \quad x_2 - z_2 \in \partial h(z_2)$$

The nonexpansiveness claim  $\|z_1 - z_2\|_2 \leq \|x_1 - x_2\|_2$  would follow if

$$\underbrace{(x_1 - x_2)^\top (z_1 - z_2)}_{\text{firm nonexpansiveness}} \geq \|z_1 - z_2\|_2^2 \quad (\text{together with Cauchy-Schwarz})$$

$$\begin{aligned} &\iff (x_1 - z_1 - x_2 + z_2)^\top (z_1 - z_2) \geq 0 \\ &\iff \begin{cases} h(z_2) \geq h(z_1) + \langle \underbrace{x_1 - z_1}_{\in \partial h(z_1)}, z_2 - z_1 \rangle \\ h(z_1) \geq h(z_2) + \langle \underbrace{x_2 - z_2}_{\in \partial h(z_2)}, z_1 - z_2 \rangle \end{cases} \end{aligned}$$

add these inequalities

# Resolvent of subdifferential operator

One can interpret prox via the resolvent of subdifferential operator

## Fact 6.2

Suppose that  $f$  is convex. Then one can write

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

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

## Justification of Fact 6.2

---

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

$$\iff \mathbf{0} \in \partial f(\mathbf{z}) + \mathbf{z} - \mathbf{x} \quad (\text{optimality condition})$$

$$\iff \mathbf{x} \in (\mathcal{I} + \partial f)(\mathbf{z})$$

$$\iff \mathbf{z} = (\mathcal{I} + \partial f)^{-1}(\mathbf{x})$$

# Moreau decomposition

---

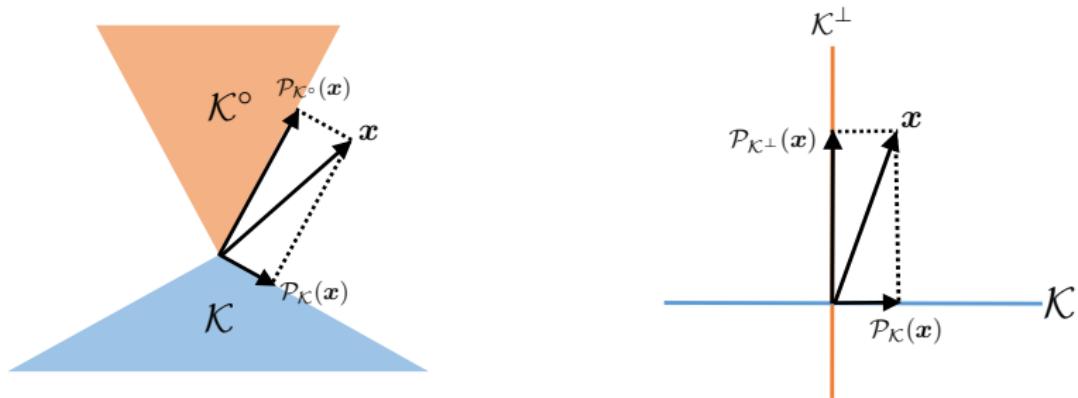
## Fact 6.3

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

$$\mathbf{x} = \text{prox}_f(\mathbf{x}) + \text{prox}_{f^*}(\mathbf{x})$$

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

# Moreau decomposition for convex cones



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

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

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

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

## Proof of Fact 6.3

---

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

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

This together with **conjugate subgradient theorem (homework)** yields

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

In view of the optimality condition, this means

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

$$\implies \mathbf{x} = \mathbf{u} + (\mathbf{x} - \mathbf{u}) = \text{prox}_f(\mathbf{x}) + \text{prox}_{f^*}(\mathbf{x})$$

## Example: prox of support function

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

$$\text{prox}_{S_{\mathcal{C}}}(\mathbf{x}) = \mathbf{x} - \mathcal{P}_{\mathcal{C}}(\mathbf{x}) \quad (6.3)$$

**Proof:** First of all, it is easy to verify that [\(exercise\)](#)

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

Then the Moreau decomposition gives

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



## Example: $\ell_\infty$ norm

---

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

where  $\mathcal{B}_{\|\cdot\|_1} := \{\mathbf{z} \mid \|\mathbf{z}\|_1 \leq 1\}$  is unit  $\ell_1$  ball

**Remark:** projection onto  $\ell_1$  ball can be computed efficiently

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

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

□

## Example: max function

---

Let  $g(\mathbf{x}) = \max\{x_1, \dots, x_n\}$ , then

$$\text{prox}_g(\mathbf{x}) = \mathbf{x} - \mathcal{P}_\Delta(\mathbf{x})$$

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

**Remark:** projection onto  $\Delta$  can be computed efficiently

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

$$\text{prox}_g(\mathbf{x}) = \mathbf{x} - \mathcal{P}_\Delta(\mathbf{x})$$



# Extended Moreau decomposition

---

A useful extension (homework):

## Fact 6.4

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

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

# **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

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

- 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  $\mathbf{y}^+ = \text{prox}_{\frac{1}{L}h}(\mathbf{y} - \frac{1}{L}\nabla f(\mathbf{y}))$ , then

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

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

Take  $\mathbf{x} = \mathbf{y} = \mathbf{x}^t$  (and hence  $\mathbf{y}^+ = \mathbf{x}^{t+1}$ ) 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

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

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

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

which immediately concludes the proof

# Proof of Lemma 6.6

---

Define

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

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

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

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

*Remark: we are propagating the smoothness of  $f$  to the strong convexity of another function  $\phi$*

- From the smoothness condition of  $f$ ,

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

## Proof of Lemma 6.6 (cont.)

---

Taken collectively, these yield

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

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

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

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

$$F(\mathbf{x}^t) - F^{\text{opt}} \leq \frac{L\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{2t}$$

- 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  $\mathbf{x} = \mathbf{x}^*$ ,  $\mathbf{y} = \mathbf{x}^t$  to obtain

$$\begin{aligned} F(\mathbf{x}^{t+1}) - F(\mathbf{x}^*) &\leq \frac{L}{2} \|\mathbf{x}^t - \mathbf{x}^*\|_2^2 - \frac{L}{2} \|\mathbf{x}^{t+1} - \mathbf{x}^*\|_2^2 - \underbrace{g(\mathbf{x}^*, \mathbf{x}^t)}_{\geq 0 \text{ by convexity}} \\ &\leq \frac{L}{2} \|\mathbf{x}^t - \mathbf{x}^*\|_2^2 - \frac{L}{2} \|\mathbf{x}^{t+1} - \mathbf{x}^*\|_2^2 \end{aligned}$$

Apply it recursively and add up all inequalities to get

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

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

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

# Convergence for strongly convex problems

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

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

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

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

## Proof of Theorem 6.9

---

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

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

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

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

Applying it recursively concludes the proof

# Numerical example: LASSO

---

*taken from UCLA EE236C*

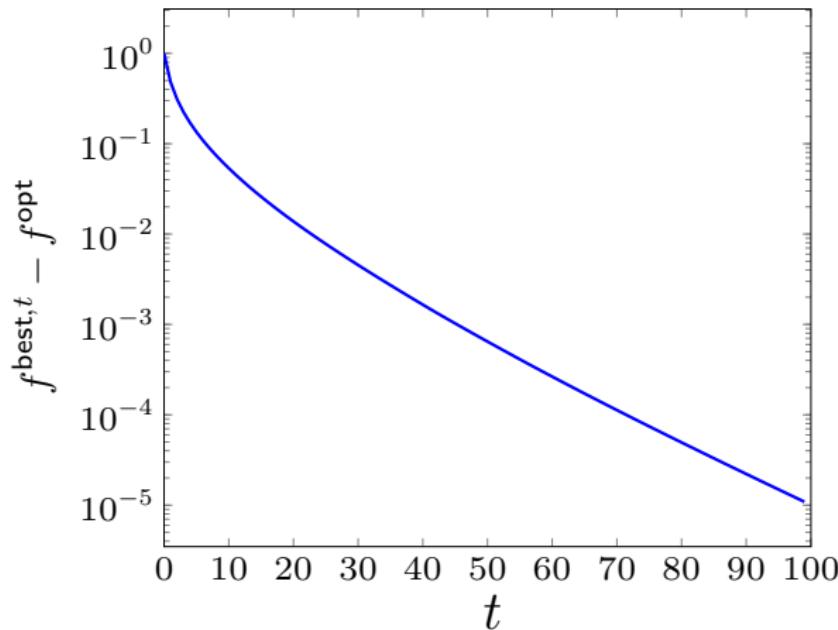
$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}) = \frac{1}{2} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 + \|\boldsymbol{x}\|_1$$

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

$$f(\mathbf{x}^t - \eta \nabla f(\mathbf{x}^t)) \leq f(\mathbf{x}^t) - \frac{\eta}{2} \|\nabla f(\mathbf{x}^t)\|_2^2$$

## Backtracking line search

---

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

$$f(\mathbf{x}^t - \eta \nabla f(\mathbf{x}^t)) \leq f(\mathbf{x}^t) - \frac{\eta}{2} \|\nabla f(\mathbf{x}^t)\|_2^2$$

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

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

# Backtracking line search

---

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

---

## Algorithm 6.2 Backtracking line search for proximal gradient methods

---

- 1: Initialize  $\eta = 1$ ,  $0 < \alpha \leq 1/2$ ,  $0 < \beta < 1$
  - 2: **while**  $f(\mathcal{T}_{L_t}(\mathbf{x}^t)) > f(\mathbf{x}^t) - \langle \nabla f(\mathbf{x}^t), \mathbf{x}^t - \mathcal{T}_{L_t}(\mathbf{x}^t) \rangle + \frac{L_t}{2} \|\mathcal{T}_{L_t}(\mathbf{x}^t) - \mathbf{x}^t\|_2^2$   
**do**
  - 3:      $L_t \leftarrow \frac{1}{\beta} L_t$     (or  $\frac{1}{L_t} \leftarrow \beta \frac{1}{L_t}$ )
- 

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

## 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	$\eta_t = \frac{1}{L}$	$O\left((1 - \frac{1}{\kappa})^t\right)$	$O\left(\kappa \log \frac{1}{\varepsilon}\right)$

# Reference

---

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