Convex Optimization

(EE227A: UC Berkeley)

Lecture 15
(Gradient methods – III)

12 March, 2013

Suvrit Sra

Optimal gradient methods

Optimal gradient methods

♠ We saw following efficiency estimates for the gradient method

$$f \in C_L^1 : \qquad f(x^k) - f^* \le \frac{2L\|x^0 - x^*\|_2^2}{k+4}$$

$$f \in S_{L,\mu}^1 : \qquad f(x^k) - f^* \le \tfrac{L}{2}\left(\frac{L-\mu}{L+\mu}\right)^{2k} \|x^0 - x^*\|_2^2.$$

Optimal gradient methods

♠ We saw following efficiency estimates for the gradient method

$$f \in C_L^1: \qquad f(x^k) - f^* \leq \frac{2L\|x^0 - x^*\|_2^2}{k+4}$$

$$f \in S_{L,\mu}^1: \qquad f(x^k) - f^* \leq \tfrac{L}{2}\left(\frac{L-\mu}{L+\mu}\right)^{2k}\|x^0 - x^*\|_2^2.$$

♠ We also saw **lower complexity bounds**

$$f \in C_L^1: \qquad f(x^k) - f(x^*) \geq \frac{3L\|x^0 - x^*\|_2^2}{32(k+1)^2}$$

$$f S_{L,\mu}^\infty: \qquad f(x^k) - f(x^*) \geq \frac{\mu}{2}\left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^{2k}\|x^0 - x^*\|_2^2.$$

Optimal gradient methods

♠ We saw following efficiency estimates for the gradient method

$$f \in C_L^1 : \qquad f(x^k) - f^* \leq \frac{2L\|x^0 - x^*\|_2^2}{k+4}$$

$$f \in S_{L,\mu}^1 : \qquad f(x^k) - f^* \leq \tfrac{L}{2}\left(\frac{L-\mu}{L+\mu}\right)^{2k}\|x^0 - x^*\|_2^2.$$

♠ We also saw **lower complexity bounds**

$$f \in C_L^1 : \qquad f(x^k) - f(x^*) \geq \frac{3L\|x^0 - x^*\|_2^2}{32(k+1)^2}$$

$$f S_{L,\mu}^\infty : \qquad f(x^k) - f(x^*) \geq \frac{\mu}{2}\left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^{2k}\|x^0 - x^*\|_2^2.$$

Can we close the gap?

Gradient with "momentum"

Polyak's method (aka heavy-ball) for $f \in S^1_{L,\mu}$

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k) + \beta_k (x^k - x^{k-1})$$

Gradient with "momentum"

Polyak's method (aka heavy-ball) for $f \in S^1_{L,\mu}$

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k) + \beta_k (x^k - x^{k-1})$$

- **Converges** (locally, i.e., for $\|x^0 - x^*\|_2 \leq \epsilon$) as

$$\|x^k - x^*\|_2^2 \leq \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}\right)^{2k} \|x^0 - x^*\|_2^2,$$

for $\alpha_k = \frac{4}{(\sqrt{L}+\sqrt{\mu})^2}$ and $\beta_k = \left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^2$

Nesterov's optimal gradient method

$\min_x f(x)$, where $S^1_{L,\mu}$ with $\mu \geq 0$

Nesterov's optimal gradient method

$\min_x f(x)$, where $S^1_{L,\mu}$ with $\mu \geq 0$

1. Choose $x^0 \in \mathbb{R}^n$, $\alpha_0 \in (0, 1)$
2. Let $y^0 \leftarrow x^0$; set $q = \mu/L$

Nesterov's optimal gradient method

$\min_x f(x)$, where $S^1_{L,\mu}$ with $\mu \geq 0$

1. Choose $x^0 \in \mathbb{R}^n$, $\alpha_0 \in (0,1)$
2. Let $y^0 \leftarrow x^0$; set $q = \mu/L$
3. k-th iteration ($k \geq 0$):
   a). Compute $f(y^k)$ and $\nabla f(y^k)$; update primary solution

$$x^{k+1} = y^k - \tfrac{1}{L}\nabla f(y^k)$$

Nesterov's optimal gradient method

$\min_x f(x)$, where $S^1_{L,\mu}$ with $\mu \geq 0$

1. Choose $x^0 \in \mathbb{R}^n$, $\alpha_0 \in (0,1)$
2. Let $y^0 \leftarrow x^0$; set $q = \mu/L$
3. k-th iteration ($k \geq 0$):
   a). Compute $f(y^k)$ and $\nabla f(y^k)$; update primary solution

$$x^{k+1} = y^k - \frac{1}{L}\nabla f(y^k)$$

   b). Compute stepsize α_{k+1} by solving

$$\alpha_{k+1}^2 = (1-\alpha_{k+1})\alpha_k^2 + q\alpha_{k+1}$$

Nesterov's optimal gradient method

$\min_x f(x)$, where $S^1_{L,\mu}$ with $\mu \geq 0$

1. Choose $x^0 \in \mathbb{R}^n$, $\alpha_0 \in (0,1)$
2. Let $y^0 \leftarrow x^0$; set $q = \mu/L$
3. k-th iteration ($k \geq 0$):

   a). Compute $f(y^k)$ and $\nabla f(y^k)$; update primary solution

   $$x^{k+1} = y^k - \tfrac{1}{L}\nabla f(y^k)$$

   b). Compute stepsize α_{k+1} by solving

   $$\alpha_{k+1}^2 = (1-\alpha_{k+1})\alpha_k^2 + q\alpha_{k+1}$$

   c). Set $\beta_k = \alpha_k(1-\alpha_k)/(\alpha_k^2 + \alpha_{k+1})$

   d). Update secondary solution

   $$y^{k+1} = x^{k+1} + \beta_k(x^{k+1} - x^k)$$

Optimal gradient method – rate

Theorem Let $\{x^k\}$ be sequence generated by above algorithm. If $\alpha_0 \geq \sqrt{\mu/L}$, then

$$f(x^k) - f(x^*) \leq c_1 \min\left\{\left(1 - \sqrt{\frac{\mu}{L}}\right)^k, \frac{4L}{(2\sqrt{L} + c_2 k)^2}\right\},$$

where constants c_1, c_2 depend on α_0, L, μ.

Proof: Somewhat involved; see notes.

Strongly convex case – simplification

If $\mu > 0$, select $\alpha_0 = \sqrt{\mu/L}$. The two main steps get simplified:

1. Set $\beta_k = \alpha_k(1-\alpha_k)/(\alpha_k^2 + \alpha_{k+1})$
2. $y^{k+1} = x^{k+1} + \beta_k(x^{k+1} - x^k)$

$$\alpha_k = \sqrt{\frac{\mu}{L}} \qquad \beta_k = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}, \qquad k \geq 0.$$

Strongly convex case – simplification

If $\mu > 0$, select $\alpha_0 = \sqrt{\mu/L}$. The two main steps get simplified:

1. Set $\beta_k = \alpha_k(1-\alpha_k)/(\alpha_k^2 + \alpha_{k+1})$
2. $y^{k+1} = x^{k+1} + \beta_k(x^{k+1} - x^k)$

$$\alpha_k = \sqrt{\frac{\mu}{L}} \qquad \beta_k = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}, \qquad k \geq 0.$$

Optimal method simplifies to

1. Choose $y^0 = x^0 \in \mathbb{R}^n$
2. k-th iteration ($k \geq 0$):

Strongly convex case – simplification

If $\mu > 0$, select $\alpha_0 = \sqrt{\mu/L}$. The two main steps get simplified:

1. Set $\beta_k = \alpha_k(1-\alpha_k)/(\alpha_k^2 + \alpha_{k+1})$
2. $y^{k+1} = x^{k+1} + \beta_k(x^{k+1} - x^k)$

$$\alpha_k = \sqrt{\frac{\mu}{L}} \qquad \beta_k = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}, \qquad k \geq 0.$$

Optimal method simplifies to

1. Choose $y^0 = x^0 \in \mathbb{R}^n$
2. k-th iteration ($k \geq 0$):
   a). $x^{k+1} = y^k - \frac{1}{L}\nabla f(y^k)$
   b). $y^{k+1} = x^{k+1} + \beta(x^{k+1} - x^k)$

Strongly convex case – simplification

If $\mu > 0$, select $\alpha_0 = \sqrt{\mu/L}$. The two main steps get simplified:

1. Set $\beta_k = \alpha_k(1-\alpha_k)/(\alpha_k^2 + \alpha_{k+1})$
2. $y^{k+1} = x^{k+1} + \beta_k(x^{k+1} - x^k)$

$$\alpha_k = \sqrt{\tfrac{\mu}{L}} \qquad \beta_k = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}, \qquad k \geq 0.$$

Optimal method simplifies to

1. Choose $y^0 = x^0 \in \mathbb{R}^n$
2. k-th iteration ($k \geq 0$):
   a). $x^{k+1} = y^k - \frac{1}{L}\nabla f(y^k)$
   b). $y^{k+1} = x^{k+1} + \beta(x^{k+1} - x^k)$

Notice similarity to Polyak's method!

Summary so far

- Convex $f(x)$ with $\|\partial f\| \leq G$ – **subgradient method**
- Differentiable $f \in C_L^1$ using gradient methods
- Rate of convergence for smooth convex problems
- Faster rate of convergence for smooth, strongly convex
- Constrained gradient methods – Frank-Wolfe method
- Constrained gradient methods – gradient projection
- Nesterov's optimal gradient methods (smooth)

Summary so far

- Convex $f(x)$ with $\|\partial f\| \leq G$ – **subgradient method**
- Differentiable $f \in C_L^1$ using gradient methods
- Rate of convergence for smooth convex problems
- Faster rate of convergence for smooth, strongly convex
- Constrained gradient methods – Frank-Wolfe method
- Constrained gradient methods – gradient projection
- Nesterov's optimal gradient methods (smooth)
- Gap between lower and upper bounds
- $O(1/\sqrt{t})$ convex (subgradient method);
- $O(1/t^2)$ for C_L^1; linear for smoooth, strongly convex

Nonsmooth optimization

- Unconstrained problem: $\min f(x)$, where $x \in \mathbb{R}^n$
- f convex on $\mathbb{R}^n$, and Lipschitz cont. on bounded set

$$|f(x) - f(y)| \leq L\|x - y\|_2, \qquad x, y \in \mathcal{X}.$$

Nonsmooth optimization

- Unconstrained problem: $\min f(x)$, where $x \in \mathbb{R}^n$
- f convex on $\mathbb{R}^n$, and Lipschitz cont. on bounded set

$$|f(x) - f(y)| \leq L\|x - y\|_2, \qquad x, y \in \mathcal{X}.$$

- At each step, we access to $f(x)$ and $g \in \partial f(x)$

Nonsmooth optimization

- Unconstrained problem: $\min f(x)$, where $x \in \mathbb{R}^n$
- f convex on $\mathbb{R}^n$, and Lipschitz cont. on bounded set

$$|f(x) - f(y)| \le L\|x - y\|_2, \qquad x, y \in \mathcal{X}.$$

- At each step, we access to $f(x)$ and $g \in \partial f(x)$
- Find $x^k \in \mathbb{R}^n$ such that $f(x^k) - f^* \le \epsilon$

Nonsmooth optimization

- Unconstrained problem: $\min f(x)$, where $x \in \mathbb{R}^n$
- f convex on $\mathbb{R}^n$, and Lipschitz cont. on bounded set

$$|f(x) - f(y)| \leq L\|x - y\|_2, \qquad x, y \in \mathcal{X}.$$

- At each step, we access to $f(x)$ and $g \in \partial f(x)$
- Find $x^k \in \mathbb{R}^n$ such that $f(x^k) - f^* \leq \epsilon$
- **First-order methods**: $x^k \in x^0 + \text{span}\left\{g^0, \ldots, g^{k-1}\right\}$

Nonsmooth optimization

EXAMPLE

- Let $\phi(x) = |x|$ for $x \in \mathbb{R}$

Nonsmooth optimization

EXAMPLE

- Let $\phi(x) = |x|$ for $x \in \mathbb{R}$
- Subgradient method $x^{k+1} = x^k - \alpha_k g^k$, where $g^k \in \partial |x^k|$.

Nonsmooth optimization

EXAMPLE

- Let $\phi(x) = |x|$ for $x \in \mathbb{R}$
- Subgradient method $x^{k+1} = x^k - \alpha_k g^k$, where $g^k \in \partial |x^k|$.
- If $x^0 = 1$ and $\alpha_k = \frac{1}{\sqrt{k+1}} + \frac{1}{\sqrt{k+2}}$ (this stepsize is known to be optimal), then $|x^k| = \frac{1}{\sqrt{k+1}}$

Nonsmooth optimization

EXAMPLE

- Let $\phi(x) = |x|$ for $x \in \mathbb{R}$
- Subgradient method $x^{k+1} = x^k - \alpha_k g^k$, where $g^k \in \partial |x^k|$.
- If $x^0 = 1$ and $\alpha_k = \frac{1}{\sqrt{k+1}} + \frac{1}{\sqrt{k+2}}$ (this stepsize is known to be optimal), then $|x^k| = \frac{1}{\sqrt{k+1}}$
- Thus, $O(\frac{1}{\epsilon^2})$ iterations are needed to obtain ϵ-accuracy.

Nonsmooth optimization

EXAMPLE

- Let $\phi(x) = |x|$ for $x \in \mathbb{R}$
- Subgradient method $x^{k+1} = x^k - \alpha_k g^k$, where $g^k \in \partial |x^k|$.
- If $x^0 = 1$ and $\alpha_k = \frac{1}{\sqrt{k+1}} + \frac{1}{\sqrt{k+2}}$ (this stepsize is known to be optimal), then $|x^k| = \frac{1}{\sqrt{k+1}}$
- Thus, $O(\frac{1}{\epsilon^2})$ iterations are needed to obtain ϵ-accuracy.
- This behavior typical for the subgradient method which exhibits $O(1/\sqrt{k})$ convergence in general

Nonsmooth optimization

EXAMPLE

- Let $\phi(x) = |x|$ for $x \in \mathbb{R}$
- Subgradient method $x^{k+1} = x^k - \alpha_k g^k$, where $g^k \in \partial|x^k|$.
- If $x^0 = 1$ and $\alpha_k = \frac{1}{\sqrt{k+1}} + \frac{1}{\sqrt{k+2}}$ (this stepsize is known to be optimal), then $|x^k| = \frac{1}{\sqrt{k+1}}$
- Thus, $O(\frac{1}{\epsilon^2})$ iterations are needed to obtain ϵ-accuracy.
- This behavior typical for the subgradient method which exhibits $O(1/\sqrt{k})$ convergence in general

Can we do better in general?

Nonsmooth optimization

Nope!

Nonsmooth optimization

Nope!

Theorem (Nesterov.) Let $\mathcal{B} = \{x \mid \|x - x^0\|_2 \leq D\}$. Assume, $x^* \in \mathcal{B}$. There exists a convex function f in $C_L^0(\mathcal{B})$ (with $L > 0$), such that for $0 \leq k \leq n-1$, the lower-bound

$$f(x^k) - f(x^*) \geq \frac{LD}{2(1+\sqrt{k+1})},$$

holds for **any algorithm** that generates x^k by linearly combining the previous iterates and subgradients.

Nope!

Theorem (Nesterov.) Let $\mathcal{B} = \{x \mid \|x - x^0\|_2 \le D\}$. Assume, $x^* \in \mathcal{B}$. There exists a convex function f in $C_L^0(\mathcal{B})$ (with $L > 0$), such that for $0 \le k \le n-1$, the lower-bound

$$f(x^k) - f(x^*) \ge \frac{LD}{2(1+\sqrt{k+1})},$$

holds for **any algorithm** that generates x^k by linearly combining the previous iterates and subgradients.

Should we give up?

Nonsmooth optimization

Nope!

Theorem (Nesterov.) Let $\mathcal{B} = \{x \mid \|x - x^0\|_2 \leq D\}$. Assume, $x^* \in \mathcal{B}$. There exists a convex function f in $C_L^0(\mathcal{B})$ (with $L > 0$), such that for $0 \leq k \leq n-1$, the lower-bound

$$f(x^k) - f(x^*) \geq \frac{LD}{2(1+\sqrt{k+1})},$$

holds for **any algorithm** that generates x^k by linearly combining the previous iterates and subgradients.

Should we give up? **No!**

Nonsmooth optimization

Nope!

Theorem (Nesterov.) Let $\mathcal{B} = \left\{x \mid \|x - x^0\|_2 \le D\right\}$. Assume, $x^* \in \mathcal{B}$. There exists a convex function f in $C_L^0(\mathcal{B})$ (with $L > 0$), such that for $0 \le k \le n-1$, the lower-bound

$$f(x^k) - f(x^*) \ge \frac{LD}{2(1+\sqrt{k+1})},$$

holds for **any algorithm** that generates x^k by linearly combining the previous iterates and subgradients.

Should we give up? **No!** Several possibilities remain!

Nonsmooth optimization

- Blackbox too pessimistic

Nonsmooth optimization

- Blackbox too pessimistic
- Nesterov's breakthroughs
  - Excessive gap technique
  - Composite objective minimization

Nonsmooth optimization

- Blackbox too pessimistic
- Nesterov's breakthroughs
  - Excessive gap technique
  - Composite objective minimization
- Nemirovski's workshorse of general convex optimization
  - Mirror-descent, NERML
  - Mirror-prox

Nonsmooth optimization

- Blackbox too pessimistic
- Nesterov's breakthroughs
  - Excessive gap technique
  - Composite objective minimization
- Nemirovski's workshorse of general convex optimization
  - Mirror-descent, NERML
  - Mirror-prox
- Other techniques, problem classes, etc.

Proximal splitting

Composite objectives

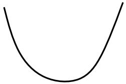
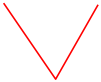
Frequently nonsmooth problems take the form

$$\boxed{\text{minimize } f(x) := \ell(x) + r(x)}$$

Composite objectives

Frequently nonsmooth problems take the form

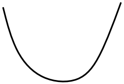
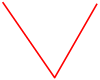
$$\boxed{\text{minimize } f(x) := \ell(x) + r(x)}$$

$\ell \in$  $+ \; r \in$ 

Composite objectives

Frequently nonsmooth problems take the form

$$\boxed{\text{minimize } f(x) := \ell(x) + r(x)}$$

$\ell \in$  $+ \; r \in$ 

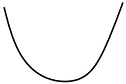
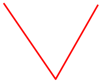
Example: $\ell(x) = \frac{1}{2}\|Ax - b\|^2$ and $r(x) = \lambda\|x\|_1$

Lasso, L1-LS, compressed sensing

Composite objectives

Frequently nonsmooth problems take the form

$$\boxed{\text{minimize } f(x) := \ell(x) + \color{red}{r(x)}}$$

$\ell \in$  $+\ r \in$ 

Example: $\ell(x) = \frac{1}{2}\|Ax - b\|^2$ and $r(x) = \lambda\|x\|_1$

$$\boxed{\text{Lasso, L1-LS, compressed sensing}}$$

Example: $\ell(x)$: Logistic loss, and $r(x) = \lambda\|x\|_1$

$$\boxed{\text{L1-Logistic regression, sparse LR}}$$

Composite objective minimization

$$\boxed{\text{minimize } f(x) := \ell(x) + r(x)}$$

$$\boxed{\textbf{subgradient: } x^{k+1} = x^k - \alpha^k g^k, \; g^k \in \partial f(x^k)}$$

Composite objective minimization

$$\text{minimize } f(x) := \ell(x) + r(x)$$

subgradient: $x^{k+1} = x^k - \alpha^k g^k$, $g^k \in \partial f(x^k)$

subgradient: converges slowly at rate $O(1/\sqrt{k})$

Composite objective minimization

minimize $f(x) := \ell(x) + r(x)$

subgradient: $x^{k+1} = x^k - \alpha^k g^k$, $g^k \in \partial f(x^k)$

subgradient: converges slowly at rate $O(1/\sqrt{k})$

but: f is *smooth* plus *nonsmooth*

we should **exploit:** smoothness of ℓ for better method!

Projections: another view

Let $\mathbb{I}_{\mathcal{X}}$ be the *indicator function* for closed, cvx $\mathcal{X}$, defined as:

$$\mathbb{I}_{\mathcal{X}}(x) := \begin{cases} 0 & \text{if } x \in \mathcal{X}, \\ \infty & \text{otherwise.} \end{cases}$$

Projections: another view

Let $\mathbb{I}_{\mathcal{X}}$ be the *indicator function* for closed, cvx $\mathcal{X}$, defined as:

$$\mathbb{I}_{\mathcal{X}}(x) := \begin{cases} 0 & \text{if } x \in \mathcal{X}, \\ \infty & \text{otherwise.} \end{cases}$$

Recall **orthogonal projection** $P_{\mathcal{X}}(y)$

$$P_{\mathcal{X}}(y) := \operatorname{argmin} \quad \tfrac{1}{2}\|x - y\|_2^2 \quad \text{s.t.} \quad x \in \mathcal{X}.$$

Projections: another view

Let $\mathbb{I}_{\mathcal{X}}$ be the *indicator function* for closed, cvx $\mathcal{X}$, defined as:

$$\mathbb{I}_{\mathcal{X}}(x) := \begin{cases} 0 & \text{if } x \in \mathcal{X}, \\ \infty & \text{otherwise.} \end{cases}$$

Recall **orthogonal projection** $P_{\mathcal{X}}(y)$

$$P_{\mathcal{X}}(y) := \operatorname{argmin} \quad \tfrac{1}{2}\|x-y\|_2^2 \quad \text{s.t.} \quad x \in \mathcal{X}.$$

Rewrite orthogonal projection $P_{\mathcal{X}}(y)$ as

$$P_{\mathcal{X}}(y) := \operatorname{argmin}_{x \in \mathbb{R}^n} \quad \tfrac{1}{2}\|x-y\|_2^2 + \mathbb{I}_{\mathcal{X}}(x).$$

Generalizing projections – proximity

Projection

$$P_{\mathcal{X}}(y) := \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \quad \tfrac{1}{2}\|x - y\|_2^2 + \mathbb{I}_{\mathcal{X}}(x)$$

Generalizing projections – proximity

Projection

$$P_{\mathcal{X}}(y) := \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \quad \tfrac{1}{2}\|x - y\|_2^2 + \mathbb{I}_{\mathcal{X}}(x)$$

Proximity: Replace $\mathbb{I}_{\mathcal{X}}$ by some convex function!

$$\operatorname{prox}_r(y) := \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \quad \tfrac{1}{2}\|x - y\|_2^2 + r(x)$$

Generalizing projections – proximity

Projection

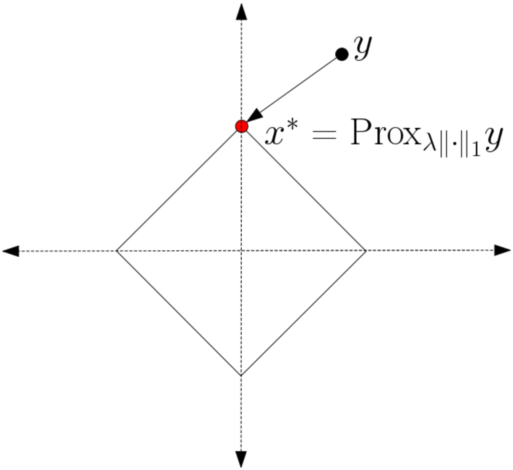
$$P_{\mathcal{X}}(y) := \operatorname*{argmin}_{x \in \mathbb{R}^n} \quad \tfrac{1}{2}\|x - y\|_2^2 + \mathbb{I}_{\mathcal{X}}(x)$$

Proximity: Replace $\mathbb{I}_{\mathcal{X}}$ by some convex function!

$$\operatorname{prox}_r(y) := \operatorname*{argmin}_{x \in \mathbb{R}^n} \quad \tfrac{1}{2}\|x - y\|_2^2 + r(x)$$

Def. $\operatorname{prox}_R : \mathbb{R}^n \to \mathbb{R}^n$ is called a **proximity operator**

Proximity operator



ℓ_1-norm ball of radius $\rho(\lambda)$

Proximity operators

Exercise: Let $r(x) = \|x\|_1$. Solve $\operatorname{prox}_{\lambda r}(y)$.

$$\min_{x \in \mathbb{R}^n} \quad \tfrac{1}{2}\|x - y\|_2^2 + \lambda \|x\|_1.$$

Hint 1: The above problem decomposes into n independent subproblems of the form

$$\min_{x \in \mathbb{R}} \quad \tfrac{1}{2}(x - y)^2 + \lambda |x|.$$

Hint 2: Consider the two cases separately: either $x = 0$ or $x \neq 0$

Proximity operators

Exercise: Let $r(x) = \|x\|_1$. Solve $\mathrm{prox}_{\lambda r}(y)$.

$$\min_{x \in \mathbb{R}^n} \quad \tfrac{1}{2}\|x - y\|_2^2 + \lambda\|x\|_1.$$

Hint 1: The above problem decomposes into n independent subproblems of the form

$$\min_{x \in \mathbb{R}} \quad \tfrac{1}{2}(x - y)^2 + \lambda|x|.$$

Hint 2: Consider the two cases separately: either $x = 0$ or $x \neq 0$

Aka: Soft-thresholding operator

Basics of proximal splitting

Recall **Gradient projection** for solving $\min_{\mathcal{X}} f(x)$ for $f \in C_L^1$:

$$x^{k+1} = P_{\mathcal{X}}(x^k - \alpha_k \nabla f(x^k))$$

Basics of proximal splitting

Recall **Gradient projection** for solving $\min_{\mathcal{X}} f(x)$ for $f \in C_L^1$:

$$x^{k+1} = P_{\mathcal{X}}(x^k - \alpha_k \nabla f(x^k))$$

Proximal gradient method solves $\min \ell(x) + r(x)$

$$x^{k+1} = \operatorname{prox}_{\alpha_k r}(x^k - \alpha_k \nabla f(x^k)).$$

Basics of proximal splitting

Recall **Gradient projection** for solving $\min_{\mathcal{X}} f(x)$ for $f \in C_L^1$:

$$x^{k+1} = P_{\mathcal{X}}(x^k - \alpha_k \nabla f(x^k))$$

Proximal gradient method solves $\min \ell(x) + r(x)$

$$x^{k+1} = \mathrm{prox}_{\alpha_k r}(x^k - \alpha_k \nabla f(x^k)).$$

- This method aka: **Forward-backward splitting** (FBS)
- "Forward step:" The gradient-descent step
- "Backward step:" The prox-operator

FBS – example

Lasso / L1-LS

$$\min \quad \tfrac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1.$$

FBS – example

Lasso / L1-LS

$$\min \quad \tfrac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1.$$

$$\begin{aligned} \operatorname{prox}_{\lambda\|x\|_1} y &= \operatorname{sgn}(y) \circ \max(|y| - \lambda, 0) \\ x^{k+1} &= \operatorname{prox}_{\alpha_k \lambda\|\cdot\|_1}(x^k - \alpha_k A^T(Ax^k - b)). \end{aligned}$$

so-called iterative soft-thresholding algorithm!

FBS – example

Lasso / L1-LS

$$\min \quad \tfrac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1.$$

$$\begin{aligned} \operatorname{prox}_{\lambda\|x\|_1} y &= \operatorname{sgn}(y) \circ \max(|y| - \lambda, 0) \\ x^{k+1} &= \operatorname{prox}_{\alpha_k \lambda\|\cdot\|_1}(x^k - \alpha_k A^T(Ax^k - b)). \end{aligned}$$

so-called iterative soft-thresholding algorithm!

Exercise: Try solving the problem:

$$\min \quad \tfrac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_2.$$

Exercise

Recall our older example: $\frac{1}{2}\|D^T x - b\|_2^2$. We solved its unconstrained and constrained versions so far. Now implement a Matlab script to solve

$$\min \quad \tfrac{1}{2}\|D^T x - b\|_2^2 + \lambda\|x\|_1.$$

Exercise

Recall our older example: $\frac{1}{2}\|D^T x - b\|_2^2$. We solved its unconstrained and constrained versions so far. Now implement a Matlab script to solve

$$\min \quad \tfrac{1}{2}\|D^T x - b\|_2^2 + \lambda\|x\|_1.$$

- Use FBS as shown above
- Try different values of $\lambda > 0$ in your code

Exercise

Recall our older example: $\frac{1}{2}\|D^T x - b\|_2^2$. We solved its unconstrained and constrained versions so far. Now implement a Matlab script to solve

$$\min \quad \tfrac{1}{2}\|D^T x - b\|_2^2 + \lambda\|x\|_1.$$

- ♠ Use FBS as shown above
- ♠ Try different values of $\lambda > 0$ in your code
- ♠ Use different choices of b

Exercise

Recall our older example: $\frac{1}{2}\|D^T x - b\|_2^2$. We solved its unconstrained and constrained versions so far. Now implement a Matlab script to solve

$$\min \quad \tfrac{1}{2}\|D^T x - b\|_2^2 + \lambda\|x\|_1.$$

- ♠ Use FBS as shown above
- ♠ Try different values of $\lambda > 0$ in your code
- ♠ Use different choices of b
- ♠ Show a choice of b for which $x = \mathbf{0}$ (zero vector) is optimal

Exercise

Recall our older example: $\frac{1}{2}\|D^T x - b\|_2^2$. We solved its unconstrained and constrained versions so far. Now implement a Matlab script to solve

$$\min \quad \tfrac{1}{2}\|D^T x - b\|_2^2 + \lambda\|x\|_1.$$

- ♠ Use FBS as shown above
- ♠ Try different values of $\lambda > 0$ in your code
- ♠ Use different choices of b
- ♠ Show a choice of b for which $x = \mathbf{0}$ (zero vector) is optimal
- ♠ Experiment with different stepsizes (try $\alpha_k = 1/4$ and if that does not "work" try smaller values that might work).
- ♠ Do not expect monotonic descent

Exercise

Recall our older example: $\frac{1}{2}\|D^T x - b\|_2^2$. We solved its unconstrained and constrained versions so far. Now implement a Matlab script to solve

$$\min \quad \frac{1}{2}\|D^T x - b\|_2^2 + \lambda\|x\|_1.$$

- ♠ Use FBS as shown above
- ♠ Try different values of $\lambda > 0$ in your code
- ♠ Use different choices of b
- ♠ Show a choice of b for which $x = \mathbf{0}$ (zero vector) is optimal
- ♠ Experiment with different stepsizes (try $\alpha_k = 1/4$ and if that does not "work" try smaller values that might work).
- ♠ Do not expect monotonic descent
- ♠ Compare with versions of the subgradient method

Proximity operators

Proximity operators

- prox_r has several nice properties
- Read / Skim the paper: "Proximal Splitting Methods in Signal Processing", by Combettes and Pesquet (2010).

Theorem The operator prox_r is **firmly nonexpansive** (FNE)

$$\|\operatorname{prox}_r x - \operatorname{prox}_r y\|_2^2 \leq \langle \operatorname{prox}_r x - \operatorname{prox}_r y,\, x - y \rangle$$

Proof: (blackboard)

Corollary. The operator prox_r is **nonexpansive**

Proof: apply Cauchy-Schwarz to FNE.

Consequence of FNE

Gradient projection

$$x^{k+1} = P_{\mathcal{X}}(x^k - \alpha_k \nabla f(x^k))$$

Proximal gradients / FBS

$$x^{k+1} = \mathrm{prox}_{\alpha_k r}(x^k - \alpha_k \nabla f(x^k))$$

Same convergence theory goes through!

Exercise: Try extending proof of gradient-projection convergence to convergence for FBS.
Hint: First show that at x^*, the fixed-point equation

$$x^* = \mathrm{prox}_{\alpha r}(x^* - \alpha \nabla f(x^*)), \qquad \alpha > 0$$

Moreau Decomposition

- **Aim:** Compute $\operatorname{prox}_r y$
- Sometimes it is easier to compute $\operatorname{prox}_{r^*} y$

$$r^*(u) := \sup_x u^T x - r(x)$$

- **Moreau decomposition:** $y = \operatorname{prox}_R y + \operatorname{prox}_{R^*} y$

Moreau decomposition

Proof sketch:

- Consider $\min \frac{1}{2}\|x-y\|_2^2 + r(x)$

Moreau decomposition

Proof sketch:

- Consider $\min \frac{1}{2}\|x-y\|_2^2 + r(x)$
- Introduce new variable $z = x$, to get

$$\operatorname{prox}_r y := \tfrac{1}{2}\|x-y\|_2^2 + r(z), \text{ s.t. } x = z$$

Moreau decomposition

Proof sketch:

- Consider $\min \frac{1}{2}\|x-y\|_2^2 + r(x)$
- Introduce new variable $z = x$, to get

$$\operatorname{prox}_r y := \frac{1}{2}\|x-y\|_2^2 + r(z), \text{ s.t. } x = z$$

- Derive *Lagrangian dual* for this

Moreau decomposition

Proof sketch:

- Consider $\min \frac{1}{2}\|x-y\|_2^2 + r(x)$
- Introduce new variable $z = x$, to get

$$\operatorname{prox}_r y := \frac{1}{2}\|x-y\|_2^2 + r(z), \text{ s.t. } x = z$$

- Derive *Lagrangian dual* for this
- Simplify, and conclude!