Convex Optimization
(EE227A: UC Berkeley)

Lecture 16
(Proximal methods)
14 March, 2013

Suvrit Sra
Organizational

♦ HW3 will be released later today on bSpace
♦ Midterm to be out sometime on 18th
♦ HW2 solutions to be out before midterm released
♦ 19th March — review session to recap important material
♦ 21st March, 2013 — midterm due beginning of class.
Revisiting Gradient Projection

\[
\min f(x) \quad x \in \mathcal{X}
\]

**Gradient projection**

\[
x^{k+1} = P(x^k - \alpha_k \nabla f(x^k))
\]

where \(P\) denotes orthogonal projection onto \(\mathcal{X}\).
Gradient projection

\[ x^{k+1} = P(x^k - \alpha_k \nabla f(x^k)) \]

where \( P \) denotes orthogonal projection onto \( \mathcal{X} \).

- Mimic unconstrained case proof
- Hinges on firm nonexpansivity of \( P \)
- Also key: stationarity property \( x^* = P(x^* - \alpha \nabla f(x^*)) \)
Lemma If $x^*$ is optimal for problem, then $x^* = P(x^* - \alpha \nabla f(x^*))$

- Denote $g^* \equiv \nabla f(x^*)$ as before.
- **Optimality condition:** $\langle g^*, x - x^* \rangle \geq 0$ for all $x \in X$.
- **Optimality for proj:** $z = Py \implies \langle z - y, x - z \rangle \geq 0 \ \forall x \in X$.
- Plug $z \leftarrow x^*$, and $y \leftarrow x^* - \alpha g^*$,
  $$\langle x^* - y, x - x^* \rangle \geq 0 \implies \langle x^* - x^* + \alpha g^*, x - x^* \rangle \geq 0$$
  $$\implies \langle \alpha g^*, x - x^* \rangle \geq 0$$
  $$\implies \langle g^*, x - x^* \rangle \geq 0$$
  $$\implies x^* \text{ is optimal.}$$
Now we show that $\|x^{k+1} - x^*\|_2 \leq \|x^k - x^*\|_2$.

**Shorthand:** $u \equiv x^{k+1}$, $x \equiv x^k$, $g \equiv \nabla f(x^k)$
Gradient projection – convergence

Now we show that \( \|x^{k+1} - x^*\|_2 \leq \|x^k - x^*\|_2 \)

**Shorthand:** \( u \equiv x^{k+1}, \ x \equiv x^k, \ g \equiv \nabla f(x^k) \)

\[
\|u - x^*\|_2^2 = \|P(x - \alpha g) - P(x^* - \alpha g^*)\|_2^2 \\
\leq \|x - x^* - \alpha(g - g^*)\|_2^2 \\
\leq \|x - x^*\|_2^2 + \alpha^2 \|g - g^*\|_2^2 - 2\alpha \langle g - g^*, x - x^* \rangle \\
\leq \|x - x^*\|_2^2 + \alpha^2 \|g - g^*\|_2^2 - \frac{2\alpha}{L} \|g - g^*\|_2^2 \\
= \|x - x^*\|_2^2 + \alpha(\alpha - \frac{2}{L}) \|g - g^*\|_2^2 \\
= r_k^2 - \frac{1}{L} \|g - g^*\|_2^2 \quad (\text{if } \alpha = 1/L).
\]

Thus, we have in particular, \( r_{k+1} \leq r_k \leq r_0 \)
Now we show that $f(x^{k+1}) \leq f(x^k) - \frac{L}{2} \|\Box\|^2$
Gradient projection -- convergence

Now we show that \( f(x^{k+1}) \leq f(x^k) - \frac{L}{2} \| u - x \|_2^2 \)

\[
f(u) \leq f(x) + \langle g, u - x \rangle + \frac{L}{2} \| u - x \|_2^2
= f(x) + \langle g, P(x - \alpha g) - Px \rangle + \frac{L}{2} \| u - x \|_2^2
\]

Recall that \( \| Pa - Pb \|_2^2 \leq \langle Pa - Pb, a - b \rangle \). Thus,

\[
\| P(x - \alpha g) - Px \|_2^2 \leq \langle P(x - \alpha g) - Px, x - \alpha g - x \rangle
= -\alpha \langle g, P(x - \alpha g) - Px \rangle
\]

\[
\implies -\alpha^{-1} \| u - x \|_2^2 \leq \langle g, P(x - \alpha g) - Px \rangle
\]

Which implies that

\[
f(u) \leq f(x) + \left( \frac{L}{2} - \frac{1}{\alpha} \right) \| u - x \|_2^2
= f(x) - \frac{L}{2} \| P(x - \alpha g) - x \|_2^2.
\]
\[ f^k \geq f^{k+1} + \frac{L}{2} \| P(x^k - \alpha g^k) - x^k \|_2^2 \]

\[ \implies f^0 - f^* \geq f^{k+1} - f^* + \frac{L}{2} \sum_{i=0}^{k} \| P(x^i - \alpha g^i) - x^i \|_2^2. \]
Gradient projection – convergence

\[ f^k \geq f^{k+1} + \frac{L}{2} \| P(x^k - \alpha g^k) - x^k \|_2^2 \]

\[ \implies f^0 - f^* \geq f^{k+1} - f^* + \frac{L}{2} \sum_{i=0}^{k} \| P(x^i - \alpha g^i) - x^i \|_2^2. \]

- Since lhs is finite, and \( f^{k+1} \geq f^* \) letting \( k \to \infty \) implies that
Gradient projection – convergence

\[ f^k \geq f^{k+1} + \frac{L}{2} \| P(x^k - \alpha g^k) - x^k \|^2 \]

\[ \implies f^0 - f^* \geq f^{k+1} - f^* + \frac{L}{2} \sum_{i=0}^{k} \| P(x^i - \alpha g^i) - x^i \|^2. \]

➤ Since lhs is finite, and \( f^{k+1} \geq f^* \) letting \( k \to \infty \) implies that

\[ \lim_{k \to \infty} \| P(x^k - \alpha g^k) - x^k \|_2 = 0. \]
Gradient projection – convergence

\[ f^k \geq f^{k+1} + \frac{L}{2} \| P(x^k - \alpha g^k) - x^k \|^2 \]

\[ \implies f^0 - f^* \geq f^{k+1} - f^* + \frac{L}{2} \sum_{i=0}^{k} \| P(x^i - \alpha g^i) - x^i \|^2. \]

- Since lhs is finite, and \( f^{k+1} \geq f^* \) letting \( k \to \infty \) implies that
  \[ \lim_{k \to \infty} \| P(x^k - \alpha g^k) - x^k \|_2 = 0. \]

- This is nothing but optimality condition!
Gradient projection – convergence

\[ f^k \geq f^{k+1} + \frac{L}{2} \| P(x^k - \alpha g^k) - x^k \|_2^2 \]

\[ \implies f^0 - f^* \geq f^{k+1} - f^* + \frac{L}{2} \sum_{i=0}^{k} \| P(x^i - \alpha g^i) - x^i \|_2^2. \]

- Since lhs is finite, and \( f^{k+1} \geq f^* \) letting \( k \to \infty \) implies that
  \[ \lim_{k \to \infty} \| P(x^k - \alpha g^k) - x^k \|_2 = 0. \]

- This is nothing but optimality condition!
- So far, we did not use convexity!
- **Rate of convergence** \( O(1/k) \) using convexity
  (some more ideas needed though; see notes)
Proximal gradients – convergence

Proximal residual

\[ \lim_{k \to \infty} \| \text{prox}_{\alpha r}(x^k - \alpha g^k) - x^k \|_2 = 0. \]

**Proof:** Essentially mimics gradient projection case (care needed).
Proximal gradients – convergence

Proximal residual

\[ \lim_{k \to \infty} \| \text{prox}_{\alpha r}(x^k - \alpha g^k) - x^k\|_2 = 0. \]

**Proof:** Essentially mimics gradient projection case (care needed).

- Rate of convergence using convexity
- Analysis slightly more complicated (see notes)
Optimal methods

- Gradient method converges as $O(1/k)$
- Optimal gradient method attains $O(1/k^2)$
Optimal methods

- Gradient method converges as $O(1/k)$
- Optimal gradient method attains $O(1/k^2)$
- Gradient-projection method converges as $O(1/k)$
- Optimal version of gradient-projection $O(1/k^2)$
Optimal methods

- Gradient method converges as $O(1/k)$
- Optimal gradient method attains $O(1/k^2)$
- Gradient-projection method converges as $O(1/k)$
- Optimal version of gradient-projection $O(1/k^2)$
- Similar situation for strongly convex smooth problems
Optimal methods

- Gradient method converges as $O(1/k)$
- Optimal gradient method attains $O(1/k^2)$
- Gradient-projection method converges as $O(1/k)$
- Optimal version of gradient-projection $O(1/k^2)$
- Similar situation for strongly convex smooth problems
- Proximal-grad: converges as $O(1/k)$ for $C^1_L \text{cvx}$
- Proximal-grad: nonoptimal linear rate for $S^1_{L,\mu}$
Optimal methods

- Gradient method converges as $O(1/k)$
- Optimal gradient method attains $O(1/k^2)$
- Gradient-projection method converges as $O(1/k)$
- Optimal version of gradient-projection $O(1/k^2)$
- Similar situation for strongly convex smooth problems
- Proximal-gradients: converges as $O(1/k)$ for $C^{1}_L$ cvx
- Proximal-gradients: nonoptimal linear rate for $S^{1}_{L,\mu}$

Can we obtain optimal proximal-gradient method?
Optimal Prox-grad – FISTA

\[
\min \ell(x) + r(x)
\]

1. Set \( x^0 \in \mathbb{R}^n \); Let \( z^0 = x^0 \), \( t_0 = 1 \).
Optimal Prox-grad – FISTA

\[
\min \ell(x) + r(x)
\]

1. Set \(x^0 \in \mathbb{R}^n\); Let \(z^0 = x^0, t_0 = 1\)
2. \(k\)-th step \((k \geq 0)\)

Remark: Achieves \(O(1/k^2)\) optimal rate (assuming Lipschitzness).

Observe: Compare with optimal gradient method (very similar)

More details in notes
Optimal Prox-grad – FISTA

$$\min \ell(x) + r(x)$$

1. Set $x^0 \in \mathbb{R}^n$; Let $z^0 = x^0$, $t_0 = 1$
2. $k$-th step ($k \geq 0$)
   - $x^{k+1} = \text{prox}_{\alpha_k r}(y^k - \alpha_k \nabla \ell(y^k))$
   - $t_{k+1} = (1 + \sqrt{4t_k^2 + 1})/2$
   - $\lambda_k = (t_{k+1} + t_k - 1)/t_{k+1}$
   - $y^{k+1} = x^k + \lambda_k(x^{k+1} - x^k)$

**Remark:** Achieves $O(1/k^2)$ optimal rate (assuming Lipschitzness).

**Observe:** Compare with optimal gradient method (very similar)
Optimal Prox-grad – FISTA

\[
\min \ l(x) + r(x)
\]

1. Set \( x^0 \in \mathbb{R}^n \); Let \( z^0 = x^0 \), \( t_0 = 1 \)

2. \( k \)-th step (\( k \geq 0 \))
   - \( x^{k+1} = \text{prox}_{\alpha_k r}(y^k - \alpha_k \nabla \ell(y^k)) \)
   - \( t_{k+1} = (1 + \sqrt{4t_k^2 + 1})/2 \)
   - \( \lambda_k = (t_{k+1} + t_k - 1)/t_{k+1} \)
   - \( y^{k+1} = x^k + \lambda_k(x^{k+1} - x^k) \)

Remark: Achieves \( O(1/k^2) \) optimal rate (assuming Lipschitzness).
Observe: Compare with optimal gradient method (very similar)

More details in notes
Monotone operators
Think of $\partial f$ as a set-valued map

$$\partial f = x \Rightarrow \partial f(x).$$
Think of $\partial f$ as a **set-valued map**

$$\partial f = x \Rightarrow \partial f(x).$$

**Relation** $R$ is a subset of $\mathbb{R}^n \times \mathbb{R}^n$

- **Empty relation:** $\emptyset$
- **Identity:** $I := \{(x, x) \mid x \in \mathbb{R}^n\}$
- **Zero:** $0 := \{(x, 0) \mid x \in \mathbb{R}^n\}$
- **Subdifferential:** $\partial f := \{(x, g) \mid x \in \mathbb{R}^n, g \in \partial f(x)\}$
Think of $\partial f$ as a **set-valued map**

$$\partial f = x \Rightarrow \partial f(x).$$

**Relation** $R$ is a subset of $\mathbb{R}^n \times \mathbb{R}^n$

- **Empty relation**: $\emptyset$
- **Identity**: $I := \{(x, x) \mid x \in \mathbb{R}^n\}$
- **Zero**: $0 := \{(x, 0) \mid x \in \mathbb{R}^n\}$
- **Subdifferential**: $\partial f := \{(x, g) \mid x \in \mathbb{R}^n, g \in \partial f(x)\}$
- We write $R(x)$ to mean $\{y \mid (x, y) \in R\}$.
- **Example**: $\partial f(x) = \{g \mid (x, g) \in \partial f\}$
Goal: solve generalized equation $0 \in R(x)$

That is, find $x \in \mathbb{R}^n$ such that $(x, 0) \in R$
Generalized equations

- **Goal:** solve *generalized equation* $0 \in R(x)$
- That is, find $x \in \mathbb{R}^n$ such that $(x, 0) \in R$
- **Example:** Say $R \equiv \partial f$, then goal

$$0 \in R(x) = \partial f(x),$$

means we want to find an $x$ that minimizes $f$. 
Operations with relations

- **Inverse:** $R^{-1} := \{(y, x) \mid (x, y) \in R\}$
Operations with relations

- **Inverse**: \( R^{-1} := \{(y, x) \mid (x, y) \in R\} \)
- **Addition**: \( R + S := \{(x, y + z) \mid (x, y) \in R, (x, z) \in S\} \)
- **Example**: \( I + R := \{(x, x + y) \mid (x, y) \in R\} \)

- **Scaling**: \( \lambda R := \{(x, \lambda y) \mid (x, y) \in R\} \)
- **Resolvent**: For relation \( R \) with parameter \( \lambda \in \mathbb{R} \), \( S := (I + \lambda R)^{-1} \)
Operations with relations

- **Inverse:** \( R^{-1} := \{(y, x) \mid (x, y) \in R\} \)
- **Addition:** \( R + S := \{(x, y + z) \mid (x, y) \in R, (x, z) \in S\} \)
- **Example:** \( I + R := \{(x, x + y) \mid (x, y) \in R\} \)
- **Scaling:** \( \lambda R = \{(x, \lambda y) \mid (x, y) \in R\} \)
Operations with relations

- **Inverse:** \( R^{-1} := \{(y, x) \mid (x, y) \in R\} \)
- **Addition:** \( R + S := \{(x, y + z) \mid (x, y) \in R, (x, z) \in S\} \)
- **Example:** \( I + R := \{(x, x + y) \mid (x, y) \in R\} \)
- **Scaling:** \( \lambda R = \{(x, \lambda y) \mid (x, y) \in R\} \)
- **Resolvent:** For relation \( R \) with parameter \( \lambda \in \mathbb{R} \)

\[ S := (I + \lambda R)^{-1} \]
Operations with relations

- **Inverse:** \( R^{-1} := \{(y, x) \mid (x, y) \in R\} \)
- **Addition:** \( R + S := \{(x, y + z) \mid (x, y) \in R, (x, z) \in S\} \)
- **Example:** \( I + R := \{(x, x + y) \mid (x, y) \in R\} \)
- **Scaling:** \( \lambda R = \{(x, \lambda y) \mid (x, y) \in R\} \)
- **Resolvent:** For relation \( R \) with parameter \( \lambda \in \mathbb{R} \)

\[ S := (I + \lambda R)^{-1} \]

- \( I + \lambda R = \{(x, x + \lambda y) \mid (x, y) \in R\} \)
Operations with relations

- **Inverse:** $R^{-1} := \{(y, x) \mid (x, y) \in R\}$

- **Addition:** $R + S := \{(x, y + z) \mid (x, y) \in R, (x, z) \in S\}$

- **Example:** $I + R := \{(x, x + y) \mid (x, y) \in R\}$

- **Scaling:** $\lambda R = \{(x, \lambda y) \mid (x, y) \in R\}$

- **Resolvent:** For relation $R$ with parameter $\lambda \in \mathbb{R}$
  
  $$S := (I + \lambda R)^{-1}$$

- $I + \lambda R = \{(x, x + \lambda y) \mid (x, y) \in R\}$

- $S = \{(x + \lambda y, x) \mid (x, y) \in R\}$
Operations with relations

- **Inverse:** \( R^{-1} := \{(y, x) \mid (x, y) \in R\} \)
- **Addition:** \( R + S := \{(x, y + z) \mid (x, y) \in R, (x, z) \in S\} \)
- **Example:** \( I + R := \{(x, x + y) \mid (x, y) \in R\} \)
- **Scaling:** \( \lambda R = \{(x, \lambda y) \mid (x, y) \in R\} \)
- **Resolvent:** For relation \( R \) with parameter \( \lambda \in \mathbb{R} \)

\[
S := (I + \lambda R)^{-1}
\]

- \( I + \lambda R = \{(x, x + \lambda y) \mid (x, y) \in R\} \)
- \( S = \{(x + \lambda y, x) \mid (x, y) \in R\} \)
- If \( \lambda \neq 0 \), shorthand \((x \leftarrow v, y \leftarrow (u - v)/\lambda)\)

\[
S := \{(u, v) \mid (u - v)/\lambda \in R(v)\}
\]
**Def.** The set valued operator $R \subset \mathbb{R}^n \times \mathbb{R}^n$ is called **monotone** if

$$\langle R(x) - R(y), x - y \rangle \geq 0, \quad x, y \in \mathbb{R}^n.$$
Def. The set valued operator $R \subset \mathbb{R}^n \times \mathbb{R}^n$ is called monotone if

$$\langle R(x) - R(y), x - y \rangle \geq 0, \quad x, y \in \mathbb{R}^n.$$ 

Examples:

- Any positive semidefinite matrix $\langle Ax - Ay, x - y \rangle \geq 0$
Monotone operators

**Def.** The set valued operator \( R \subset \mathbb{R}^n \times \mathbb{R}^n \) is called **monotone** if

\[
\langle R(x) - R(y), x - y \rangle \geq 0, \quad x, y \in \mathbb{R}^n.
\]

**Examples:**

- Any positive semidefinite matrix \( \langle Ax - Ay, x - y \rangle \geq 0 \)
- The subdifferential \( \partial f \) of a convex function (verify!)
Def. The set valued operator $R \subset \mathbb{R}^n \times \mathbb{R}^n$ is called **monotone** if

$$\langle R(x) - R(y), x - y \rangle \geq 0, \quad x, y \in \mathbb{R}^n.$$ 

Examples:

- Any positive semidefinite matrix $\langle Ax - Ay, x - y \rangle \geq 0$
- The subdifferential $\partial f$ of a convex function (verify!)
- Any monotonically nondecreasing function $T : \mathbb{R} \to \mathbb{R}$
**Def.** The set valued operator $R \subset \mathbb{R}^n \times \mathbb{R}^n$ is called **monotone** if

$$\langle R(x) - R(y), x - y \rangle \geq 0, \quad x, y \in \mathbb{R}^n.$$  

**Examples:**

- Any positive semidefinite matrix $\langle Ax - Ay, x - y \rangle \geq 0$
- The subdifferential $\partial f$ of a convex function (verify!)
- Any monotonically nondecreasing function $T : \mathbb{R} \to \mathbb{R}$
- Projection and proximity operators (recall firm nonexpansivity)

Generalize notion of monotonicity to vector world
Monotone operators

Exercise: Prove $\alpha R$ monotone if $R$ monotone and $\alpha \geq 0$

Exercise: Prove $R^{-1}$ monotone, if $R$ is monotone

Exercise: If $R$, $S$ monotone, and $\alpha \geq 0$, then $R + \alpha S$ is monotone.
Monotone operators

Exercise: Prove $\alpha R$ monotone if $R$ monotone and $\alpha \geq 0$

Exercise: Prove $R^{-1}$ monotone, if $R$ is monotone

Exercise: If $R$, $S$ monotone, and $\alpha \geq 0$, then $R + \alpha S$ is monotone.

Corollary: Resolvent operator of monotone operator is monotone.
Importance of resolvents

Solve generalized equation

\[ 0 \in R(x) \]
Importance of resolvents

Solve generalized equation

\[ 0 \in R(x) \]

**Theorem** The solutions to the generalized equation coincide with points that satisfy the resolvent equation \( x = (I + \alpha R)^{-1}(x) \)
Importance of resolvents

Solve generalized equation

\[ 0 \in R(x) \]

**Theorem**  The solutions to the generalized equation coincide with points that satisfy the resolvent equation \( x = (I + \alpha R)^{-1}(x) \)

**Proof:**

\[ 0 \in R(x) \]
Solve generalized equation

\[ 0 \in R(x) \]

**Theorem** The solutions to the generalized equation coincide with points that satisfy the resolvent equation \( x = (I + \alpha R)^{-1}(x) \)

**Proof:**

\[ 0 \in R(x) \iff 0 \in \alpha R(x) \]
Importance of resolvents

Solve generalized equation

\[
0 \in R(x)
\]

**Theorem** The solutions to the generalized equation coincide with points that satisfy the resolvent equation \( x = (I + \alpha R)^{-1}(x) \)

**Proof:**

\[
0 \in R(x) \iff 0 \in \alpha R(x) \iff x \in (I + \alpha R)(x)
\]
Importance of resolvents

Solve generalized equation

\[ 0 \in R(x) \]

**Theorem** The solutions to the generalized equation coincide with points that satisfy the resolvent equation \( x = (I + \alpha R)^{-1}(x) \)

**Proof:**

\[ 0 \in R(x) \leftrightarrow 0 \in \alpha R(x) \leftrightarrow x \in (I + \alpha R)(x) \leftrightarrow x = (I + \alpha R)^{-1}(x) \]
Proximity operator as resolvent

**Theorem** Let $f$ be a closed convex function, and $\lambda > 0$, then

$$(I + \lambda \partial f)^{-1}(y) = \text{prox}_{\lambda f}(y).$$
**Theorem** Let $f$ be a closed convex function, and $\lambda > 0$, then

$$(I + \lambda \partial f)^{-1}(y) = \text{prox}_{\lambda f}(y).$$

- Suppose $(I + \lambda \partial f)^{-1}$ is single valued ($\partial f$ is monotone)
Theorem Let $f$ be a closed convex function, and $\lambda > 0$, then
\[
(I + \lambda \partial f)^{-1}(y) = \text{prox}_{\lambda f}(y).
\]

- Suppose $(I + \lambda \partial f)^{-1}$ is single valued ($\partial f$ is monotone)

- Then, $x = (I + \lambda \partial f)^{-1}(y) \implies y \in (I + \lambda \partial f)(x)$
**Theorem** Let \( f \) be a closed convex function, and \( \lambda > 0 \), then

\[
(I + \lambda \partial f)^{-1}(y) = \text{prox}_{\lambda f}(y).
\]

- Suppose \((I + \lambda \partial f)^{-1}\) is single valued (\(\partial f\) is monotone)
- Then, \(x = (I + \lambda \partial f)^{-1}(y) \implies y \in (I + \lambda \partial f)(x)\)
- That is, \(y \in x + \lambda \partial f(x)\)
**Theorem** Let $f$ be a closed convex function, and $\lambda > 0$, then

$$(I + \lambda \partial f)^{-1}(y) = \text{prox}_{\lambda f}(y).$$

- Suppose $(I + \lambda \partial f)^{-1}$ is single valued ($\partial f$ is monotone)
- Then, $x = (I + \lambda \partial f)^{-1}(y) \implies y \in (I + \lambda \partial f)(x)$
- That is, $y \in x + \lambda \partial f(x)$
- Equivalently, $x - y + \lambda \partial f(x) \ni 0$
**Theorem** Let $f$ be a closed convex function, and $\lambda > 0$, then

$$(I + \lambda \partial f)^{-1}(y) = \text{prox}_{\lambda f}(y).$$

- Suppose $(I + \lambda \partial f)^{-1}$ is single valued ($\partial f$ is monotone)
- Then, $x = (I + \lambda \partial f)^{-1}(y) \implies y \in (I + \lambda \partial f)(x)$
- That is, $y \in x + \lambda \partial f(x)$
- Equivalently, $x - y + \lambda \partial f(x) \ni 0$
- Nothing other than optimality condition for prox-operator!

$$\text{prox}_{\lambda f}(y) \equiv y \mapsto \arg\min_{x} \frac{1}{2} \|x - y\|_{2}^{2} + \lambda f(x)$$
Deriving proximal-grad method

\[
\min \quad \ell(x) + \lambda r(x)
\]

Outline
Deriving proximal-grad method

\[
\min \ell(x) + \lambda r(x)
\]

Outline

\[
0 \in \nabla \ell(x) + \lambda \partial r(x)
\]
Deriving proximal-grad method

\[ \min \ell(x) + \lambda r(x) \]

Outline

\[ 0 \in \nabla \ell(x) + \lambda \partial r(x) \]
\[ x \in \nabla \ell(x) + (I + \lambda \partial r)(x) \]
Deriving proximal-grad method

\[
\min \quad \ell(x) + \lambda r(x)
\]

**Outline**

\[
0 \in \nabla \ell(x) + \lambda \partial r(x)
\]

\[
x \in \nabla \ell(x) + (I + \lambda \partial r)(x)
\]

\[
x - \nabla \ell(x) \in (I + \lambda \partial r)(x)
\]

\[
x = (I + \lambda \partial r)^{-1}(x - \nabla \ell(x))
\]
Deriving proximal-grad method

\[
\min \ell(x) + \lambda r(x)
\]

Outline

\[
0 \in \nabla \ell(x) + \lambda \partial r(x)
\]

\[
x \in \nabla \ell(x) + (I + \lambda \partial r)(x)
\]

\[
x - \nabla \ell(x) \in (I + \lambda \partial r)(x)
\]

\[
x = (I + \lambda \partial r)^{-1}(x - \nabla \ell(x))
\]

\[
x = \text{prox}_{\lambda r}(x - \nabla \ell(x))
\]
If both $f$, $g$ nonsmooth, ordinary splitting does not work!

How to solve it?
References

♠ S. Boyd. EE364B Lecture slides
♠ Yu. Nesterov. *Introductory Lectures on Convex Optimization*
♠ F. Dinuzzo. Lecture slides on large scale optimization