Motivation

(nonsmooth optimization)
Many nonsmooth problems take the form

\[
\text{minimize } \phi(x) := f(x) + r(x)
\]
Regularized / Composite Objectives

Many nonsmooth problems take the form

\[
\text{minimize } \phi(x) := f(x) + r(x)
\]
Many nonsmooth problems take the form

\[
\text{minimize } \phi(x) := f(x) + r(x)
\]

\[
f \in \bigcup + r \in \bigvee
\]

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

Lasso, L1-LS, compressed sensing
Regularized / Composite Objectives

Many nonsmooth problems take the form

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

$$f \in \bigcup + \quad r \in \bigvee$$

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

Lasso, L1-LS, compressed sensing

Example: $$f(x) : \text{Logistic loss, and } r(x) = \lambda \|x\|_1$$

L1-Logistic regression, sparse LR
Composite objective minimization

\[
\text{minimize } \phi(x) := f(x) + r(x)
\]

**subgradient:**
\[
x^{k+1} = x^k - \eta_k g^k, \quad g^k \in \partial \phi(x^k)
\]
Composite objective minimization

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

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

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

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

subgradient: $x^{k+1} = x^k - \eta_k g^k, \quad g^k \in \partial \phi(x^k)$

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

Nesterov: exploit smoothness of $f$ to beat lower bound!
Proximal gradient method

Optimality conditions

\[ 0 \in \nabla f(x^*) + \partial r(x^*) \]
Proximal gradient method

Optimality conditions

\[
0 \in \nabla f(x^*) + \partial r(x^*) \\
0 \in \alpha \nabla f(x^*) + \alpha \partial r(x^*)
\]
Proximal gradient method

Optimality conditions

\[
0 \in \nabla f(x^*) + \partial r(x^*)
\]

\[
0 \in \alpha \nabla f(x^*) + \alpha \partial r(x^*)
\]

\[
x^* \in \alpha \nabla f(x^*) + (I + \alpha \partial r)(x^*)
\]
Proximal gradient method

Optimality conditions

\[ 0 \in \nabla f(x^*) + \partial r(x^*) \]
\[ 0 \in \alpha \nabla f(x^*) + \alpha \partial r(x^*) \]
\[ x^* \in \alpha \nabla f(x^*) + (I + \alpha \partial r)(x^*) \]
\[ x^* - \alpha \nabla f(x^*) \in (I + \alpha \partial r)(x^*) \]
Proximal gradient method

Optimality conditions

\[ 0 \in \nabla f(x^*) + \partial r(x^*) \]
\[ 0 \in \alpha \nabla f(x^*) + \alpha \partial r(x^*) \]
\[ x^* \in \alpha \nabla f(x^*) + (I + \alpha \partial r)(x^*) \]
\[ x^* - \alpha \nabla f(x^*) \in (I + \alpha \partial r)(x^*) \]
\[ x^* = (I + \alpha \partial r)^{-1}(x^* - \alpha \nabla f(x^*)) \]
Proximal gradient method

Optimality conditions

\[ 0 \in \nabla f(x^*) + \partial r(x^*) \]
\[ 0 \in \alpha \nabla f(x^*) + \alpha \partial r(x^*) \]
\[ x^* \in \alpha \nabla f(x^*) + (I + \alpha \partial r)(x^*) \]
\[ x^* - \alpha \nabla f(x^*) \in (I + \alpha \partial r)(x^*) \]
\[ x^* = (I + \alpha \partial r)^{-1}(x^* - \alpha \nabla f(x^*)) \]
\[ x^* = \text{prox}_{\alpha r}(x^* - \alpha \nabla f(x^*)) \]
Proximal gradient method

Optimality conditions

\[ 0 \in \nabla f(x^*) + \partial r(x^*) \]
\[ 0 \in \alpha \nabla f(x^*) + \alpha \partial r(x^*) \]
\[ x^* \in \alpha \nabla f(x^*) + (I + \alpha \partial r)(x^*) \]
\[ x^* - \alpha \nabla f(x^*) \in (I + \alpha \partial r)(x^*) \]
\[ x^* = (I + \alpha \partial r)^{-1}(x^* - \alpha \nabla f(x^*)) \]
\[ x^* = \text{prox}_{\alpha r}(x^* - \alpha \nabla f(x^*)) \]

Above fixed-point eqn suggests iteration

\[ x_{k+1} = \text{prox}_{\alpha_k r}(x_k - \alpha_k \nabla f(x_k)) \]

This method converges as \( O(1/k) \) for convex \( f \in C^1_L \)!
Prox operators
From projections to proximity

Let $1_{\mathcal{X}}$ be the *indicator function* for closed, cvx $\mathcal{X}$. 
From projections to proximity

Let $\mathbf{1}_\mathcal{X}$ be the *indicator function* for closed, cvx $\mathcal{X}$. Recall *orthogonal projection* $P_\mathcal{X}(y)$

$$P_\mathcal{X}(y) := \text{argmin} \quad \frac{1}{2} \|x - y\|_2^2 \quad \text{s.t.} \quad x \in \mathcal{X}.$$
From projections to proximity

Let $1_{\mathcal{X}}$ be the *indicator function* for closed, cvx $\mathcal{X}$. Recall *orthogonal projection* $P_{\mathcal{X}}(y)$

\[ P_{\mathcal{X}}(y) := \arg\min_{x} \frac{1}{2} \|x - y\|^2 \quad \text{s.t. } x \in \mathcal{X}. \]

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

\[ P_{\mathcal{X}}(y) := \arg\min_{x \in \mathbb{R}^n} \frac{1}{2} \|x - y\|^2 + 1_{\mathcal{X}}(x). \]
From projections to proximity

Let $\mathbb{1}_\mathcal{X}$ be the *indicator function* for closed, cvx $\mathcal{X}$. Recall **orthogonal projection** $P_{\mathcal{X}}(y)$

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

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

$$P_{\mathcal{X}}(y) := \arg\min_{x \in \mathbb{R}^n} \frac{1}{2} \|x - y\|_2^2 + \mathbb{1}_\mathcal{X}(x).$$

**Proximity**: Replace $\mathbb{1}_\mathcal{X}$ by some convex function!

$$\text{prox}_r(y) := \arg\min_{x \in \mathbb{R}^n} \frac{1}{2} \|x - y\|_2^2 + r(x)$$
Proximity operator

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

Unique solution due to strong convexity. Observe that:

\[
\begin{align*}
0 & \in x - y + \partial r(x) \\
y & \in (\text{Id} + \partial r)(x) \\
x & = (\text{Id} + \partial r)^{-1}(y) \\
x & = \text{prox}_r(y).
\end{align*}
\]
Exercise: Let $r(x) = \|x\|_1$. Solve $\text{prox}_{\lambda r}(y)$.

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

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

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

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

Aka: Soft-thresholding operator
Proximity operators

- $\text{prox}_r$ has several important and nice properties

Theorem.
The operator $\text{prox}_r$ is firmly nonexpansive (FNE)

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

Exercise: Prove the above property.

Corollary.
The operator $\text{prox}_r$ is nonexpansive

Proof: apply Cauchy-Schwarz to FNE.
Proximity operators

- proxᵣ has several important and nice properties

**Theorem.** The operator $\text{prox}_r$ is **firmly nonexpansive** (FNE)

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

**Exercise:** Prove the above property.
Proximity operators

- \( \text{prox}_r \) has several important and nice properties

**Theorem.** The operator \( \text{prox}_r \) is **firmly nonexpansive** (FNE)

\[
\|\text{prox}_r x - \text{prox}_r y\|_2^2 \leq \langle \text{prox}_r x - \text{prox}_r y, x - y \rangle
\]

**Exercise:** Prove the above property.

**Corollary.** The operator \( \text{prox}_r \) is **nonexpansive**

**Proof:** apply Cauchy-Schwarz to FNE.
Let $C$ be a closed, convex set. From first-order optimality conditions $\langle \nabla f(x^*), x - x^* \rangle \geq 0 \forall x \in C$. Thus,

$\langle y - P_C(y), x - P_C(y) \rangle \leq 0, \quad \forall x \in C.$
Let $C$ be a closed, convex set. From first-order optimality conditions $\langle \nabla f(x^*), x - x^* \rangle \geq 0 \ \forall x \in C$. Thus,

$$\langle y - P_C(y), x - P_C(y) \rangle \leq 0, \quad \forall \ x \in C.$$  

Using the above inequality, for two points $x_1, x_2$ we obtain

$$\langle x_1 - P_C(x_1), P_C(x_2) - P_C(x_1) \rangle \leq 0$$
Let $C$ be a closed, convex set. From first-order optimality conditions $\langle \nabla f(x^*), x - x^* \rangle \geq 0 \ \forall x \in C$. Thus,

$$\langle y - P_C(y), x - P_C(y) \rangle \leq 0, \ \forall x \in C.$$ 

Using the above inequality, for two points $x_1, x_2$ we obtain

$$\langle x_1 - P_C(x_1), P_C(x_2) - P_C(x_1) \rangle \leq 0$$
$$\langle x_2 - P_C(x_2), P_C(x_1) - P_C(x_2) \rangle \leq 0$$
Let \( C \) be a closed, convex set. From first-order optimality conditions \( \langle \nabla f(x^*), x - x^* \rangle \geq 0 \ \forall x \in C \). Thus,

\[
\langle y - P_C(y), x - P_C(y) \rangle \leq 0, \quad \forall x \in C.
\]

Using the above inequality, for two points \( x_1, x_2 \) we obtain

\[
\langle x_1 - P_C(x_1), P_C(x_2) - P_C(x_1) \rangle \leq 0
\]

\[
\langle x_2 - P_C(x_2), P_C(x_1) - P_C(x_2) \rangle \leq 0
\]

\[
\langle P_C(x_1) - P_C(x_2), x_2 - x_1 + P_C(x_1) - P_C(x_2) \rangle \leq 0.
\]
Let $C$ be a closed, convex set. From first-order optimality conditions $\langle \nabla f(x^*), x - x^* \rangle \geq 0 \ \forall x \in C$. Thus,

$$\langle y - P_C(y), x - P_C(y) \rangle \leq 0, \ \forall x \in C.$$ 

Using the above inequality, for two points $x_1, x_2$ we obtain

$$\langle x_1 - P_C(x_1), P_C(x_2) - P_C(x_1) \rangle \leq 0$$
$$\langle x_2 - P_C(x_2), P_C(x_1) - P_C(x_2) \rangle \leq 0$$
$$\langle P_C(x_1) - P_C(x_2), x_2 - x_1 + P_C(x_1) - P_C(x_2) \rangle \leq 0.$$ 

$$\|P_C(x_1) - P_C(x_2)\|_2^2 \leq \langle P_C(x_1) - P_C(x_2), x_1 - x_2 \rangle$$
Consequences of FNE

Projected gradient method
\[ x^{k+1} = P_X (x^k - \alpha_k \nabla f(x^k)) \]

Proximal gradient method
\[ x^{k+1} = \text{prox}_{\alpha_k r} (x^k - \alpha_k \nabla f(x^k)) \]

Same convergence theory goes through!
Consequences of FNE

Projected gradient method

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

Proximal gradient method

\[ x^{k+1} = \text{prox}_{\alpha_k r}(x^k - \alpha_k \nabla f(x^k)) \]

Same convergence theory goes through!

**Exercise:** Extend proof of proj-grad convergence to prox-grad.

**Hint:** First show that at \( x^* \), the fixed-point equation holds

\[ x^* = \text{prox}_{\alpha r}(x^* - \alpha \nabla f(x^*)) \quad \alpha > 0 \]
Consequences of FNE

Projected gradient method
\[ x^{k+1} = P_X(x^k - \alpha_k \nabla f(x^k)) \]

Proximal gradient method
\[ x^{k+1} = \text{prox}_{\alpha_k r}(x^k - \alpha_k \nabla f(x^k)) \]

Same convergence theory goes through!

Exercise: Extend proof of proj-grad convergence to prox-grad.

Hint: First show that at \( x^* \), the fixed-point equation holds
\[ x^* = \text{prox}_{\alpha r}(x^* - \alpha \nabla f(x^*)), \quad \alpha > 0 \]

Krasnoselskii-Mann theorem: If a FNE map on a closed convex set has a fixed-point, then the iteration \( x_{k+1} \leftarrow (1 - \alpha_k) \text{Id} + \alpha_k F(x_k) \) converges to it for \( \alpha_k \in [0, 1] \) provided \( \sum_k \alpha_k(1 - \alpha_k) = \infty \) for any starting point \( x_0 \).

Exercise: Moreau Decomposition

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

**Exercise:** *Moreau decomposition:* $y = \text{prox}_r y + \text{prox}_{r^*} y$
Exercise: Moreau Decomposition

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

**Exercise:** Moreau decomposition: $y = \text{prox}_r y + \text{prox}_{r^*} y$

**Proof sketch:**
- Consider $\min \frac{1}{2} \| x - y \|_2^2 + r(x)$
- Introduce new variable $z = x$, to get
  
  $$\text{prox}_r y := \frac{1}{2} \| x - y \|_2^2 + r(z), \text{ s.t. } x = z$$

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

\[ \min f(x) + h(x) \]
Why does prox-grad method work?

\[ x_{k+1} = \text{prox}_{\alpha_k h}(x_k - \alpha_k \nabla f(x_k)) \]
Why does prox-grad method work?

\[ x_{k+1} = \text{prox}_{\alpha_k h}(x_k - \alpha_k \nabla f(x_k)) \]

\[ x_{k+1} = x_k - \alpha_k G_{\alpha_k}(x_k). \]
Why does prox-grad method work?

\[
x_{k+1} = \text{prox}_{\alpha_k h}(x_k - \alpha_k \nabla f(x_k))
\]
\[
x_{k+1} = x_k - \alpha_k G_{\alpha_k}(x_k).
\]

Gradient mapping: the “gradient-like object”

\[
G_{\alpha}(x) = \frac{1}{\alpha}(x - P_{\alpha h}(x - \alpha \nabla f(x)))
\]
Why does prox-grad method work?

\[ x_{k+1} = \text{prox}_{\alpha_k h}(x_k - \alpha_k \nabla f(x_k)) \]
\[ x_{k+1} = x_k - \alpha_k G_{\alpha_k}(x_k). \]

**Gradient mapping: the “gradient-like object”**

\[ G_{\alpha}(x) = \frac{1}{\alpha}(x - P_{\alpha h}(x - \alpha \nabla f(x))) \]

- Observe that \( G_{\alpha}(x) = 0 \) if and only if \( x \) is optimal
- So \( G_{\alpha} \) analogous to \( \nabla f \)
- If \( x \) locally optimal, then \( G_{\alpha}(x) = 0 \) (nonconvex \( f \))
Convergence analysis

**Assumption:** Lipschitz continuous gradient; denoted $f \in C^1_L$

$$\| \nabla f(x) - \nabla f(y) \|_2 \leq L \| x - y \|_2$$

**Lemma** (Descent). Let $f \in C^1_L$. Then,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \| y - x \|_2^2$$
Let $y = x - \alpha G_\alpha(x)$, then
Descent lemma – corollary

Let \( y = x - \alpha G_\alpha(x) \), then

\[
f(y) \leq f(x) - \alpha \langle \nabla f(x), G_\alpha(x) \rangle + \frac{\alpha^2 L}{2} \|G_\alpha(x)\|^2_2.
\]
Descent lemma – corollary

Let $y = x - \alpha G_\alpha(x)$, then

$$f(y) \leq f(x) - \alpha \langle \nabla f(x), G_\alpha(x) \rangle + \frac{\alpha^2 L}{2} \|G_\alpha(x)\|_2^2.$$  

**Corollary.** So if $0 \leq \alpha \leq 1/L$, we have

$$f(y) \leq f(x) - \alpha \langle \nabla f(x), G_\alpha(x) \rangle + \frac{\alpha}{2} \|G_\alpha(x)\|_2^2.$$
Let $y = x - \alpha G_\alpha(x)$, then

$$f(y) \leq f(x) - \alpha \langle \nabla f(x), G_\alpha(x) \rangle + \frac{\alpha^2 L}{2} \| G_\alpha(x) \|^2.$$

**Corollary.** So if $0 \leq \alpha \leq 1/L$, we have

$$f(y) \leq f(x) - \alpha \langle \nabla f(x), G_\alpha(x) \rangle + \frac{\alpha}{2} \| G_\alpha(x) \|^2.$$

**Lemma** Let $y = x - \alpha G_\alpha(x)$. Then, for any $z$ we have

$$f(y) + h(y) \leq f(z) + h(z) + \langle G_\alpha(x), x - z \rangle - \frac{\alpha}{2} \| G_\alpha(x) \|^2.$$

**Exercise:** Prove! (*Hint: $f, h$ are cvx, and $G_\alpha(x) - \nabla f(x) \in \partial h(y)$*)
We’ve actually shown that $x' \leftarrow x - \alpha G_\alpha(x)$ is a descent method. Write $\phi = f + h$; plug in $z = x$ to obtain

$$
\phi(x') \leq \phi(x) - \frac{\alpha}{2} \|G_\alpha(x)\|_2^2.
$$

**Exercise:** Argue convergence via this inequality.
Convergence analysis

We’ve actually shown that $x' \leftarrow x - \alpha G_\alpha(x)$ is a descent method. Write $\phi = f + h$; plug in $z = x$ to obtain

$$\phi(x') \leq \phi(x) - \frac{\alpha}{2} \|G_\alpha(x)\|^2_2.$$ 

**Exercise:** Argue convergence via this inequality.

Plug $z = x^*$ in

$$f(y) + h(y) \leq f(z) + h(z) + \langle G_\alpha(x), x - z \rangle - \frac{\alpha}{2} \|G_\alpha(x)\|^2_2,$$

to obtain progress in terms of iterates:

$$\phi(x') - \phi^* \leq \langle G_\alpha(x), x - x^* \rangle - \frac{\alpha}{2} \|G_\alpha(x)\|^2_2.$$
Convergence analysis

We’ve actually shown that $x' \leftarrow x - \alpha G_\alpha(x)$ is a descent method. Write $\phi = f + h$; plug in $z = x$ to obtain

$$\phi(x') \leq \phi(x) - \frac{\alpha}{2} \|G_\alpha(x)\|_2^2.$$  

**Exercise:** Argue convergence via this inequality.

Plug $z = x^*$ in

$$f(y) + h(y) \leq f(z) + h(z) + \langle G_\alpha(x), x - z \rangle - \frac{\alpha}{2} \|G_\alpha(x)\|_2^2,$$  

to obtain progress in terms of iterates:

$$\phi(x') - \phi^* \leq \langle G_\alpha(x), x - x^* \rangle - \frac{\alpha}{2} \|G_\alpha(x)\|_2^2$$

$$= \frac{1}{2\alpha} \left[ 2\langle \alpha G_\alpha(x), x - x^* \rangle - \|\alpha G_\alpha(x)\|_2^2 \right]$$

$$= \frac{1}{2\alpha} \left[ \|x - x^*\|_2^2 - \|x - x^* - \alpha G_\alpha(x)\|_2^2 \right]$$
We’ve actually shown that $x' \leftarrow x - \alpha G_\alpha(x)$ is a descent method. Write $\phi = f + h$; plug in $z = x$ to obtain

$$\phi(x') \leq \phi(x) - \frac{\alpha}{2} \|G_\alpha(x)\|_2^2.$$ 

**Exercise:** Argue convergence via this inequality. Plug $z = x^*$ in

$$f(y) + h(y) \leq f(z) + h(z) + \langle G_\alpha(x), x - z \rangle - \frac{\alpha}{2} \|G_\alpha(x)\|_2^2,$$ 

to obtain progress in terms of iterates:

$$\phi(x') - \phi^* \leq \langle G_\alpha(x), x - x^* \rangle - \frac{\alpha}{2} \|G_\alpha(x)\|_2^2$$

$$= \frac{1}{2\alpha} \left[ 2\langle \alpha G_\alpha(x), x - x^* \rangle - \|\alpha G_\alpha(x)\|_2^2 \right]$$

$$= \frac{1}{2\alpha} \left[ \|x - x^*\|_2^2 - \|x - x^* - \alpha G_\alpha(x)\|_2^2 \right]$$

$$= \frac{1}{2\alpha} \left[ \|x - x^*\|_2^2 - \|x' - x^*\|_2^2 \right].$$
Convergence rate

Set $x \leftarrow x_k$, $x' \leftarrow x_{k+1}$, and $\alpha = 1/L$. Then add
Set \( x \leftarrow x_k, \ x' \leftarrow x_{k+1} \), and \( \alpha = 1/L \). Then add

\[
\sum_{i=1}^{k+1} (\phi(x_i) - \phi^*) \leq \frac{L}{2} \sum_{i=1}^{k+1} \left[ \| x_k - x^* \|_2^2 - \| x_{i+1} - x^* \|_2^2 \right]
\]
Convergence rate

Set \( x \leftarrow x_k, \ x' \leftarrow x_{k+1} \), and \( \alpha = 1/L \). Then add

\[
\sum_{i=1}^{k+1} (\phi(x_i) - \phi^*) \leq \frac{L}{2} \sum_{i=1}^{k+1} \left[ \|x_k - x^*\|^2_2 - \|x_{i+1} - x^*\|^2_2 \right]
\]
\[
= \frac{L}{2} \left[ \|x_1 - x^*\|^2_2 - \|x_{k+1} - x^*\|^2_2 \right]
\]

This is the well-known \( O\left(\frac{1}{k}\right) \) rate for proximal-gradient.

But for \( \mathcal{C}_1 \)-convex functions, optimal rate is \( O\left(\frac{1}{k^2}\right) \)!
Set \( x \leftarrow x_k, x' \leftarrow x_{k+1}, \) and \( \alpha = 1/L. \) Then add

\[
\sum_{i=1}^{k+1} (\phi(x_i) - \phi^*) \leq \frac{L}{2} \sum_{i=1}^{k+1} \left[ \|x_k - x^*\|_2 - \|x_{i+1} - x^*\|_2 \right]
\]

\[
= \frac{L}{2} \left[ \|x_1 - x^*\|_2 - \|x_{k+1} - x^*\|_2 \right]
\]

\[
\leq \frac{L}{2} \|x_1 - x^*\|_2^2.
\]

This is the well-known \( O\left(\frac{1}{k}\right) \) rate for proximal-gradient. But for \( C_1 \) convex functions, optimal rate is \( O\left(\frac{1}{k^2}\right) \)!
Convergence rate

Set $x \leftarrow x_k$, $x' \leftarrow x_{k+1}$, and $\alpha = 1/L$. Then add

$$
\sum_{i=1}^{k+1} (\phi(x_i) - \phi^*) \leq \frac{L}{2} \sum_{i=1}^{k+1} \left[ \|x_k - x^*\|^2_2 - \|x_{i+1} - x^*\|^2_2 \right]
= \frac{L}{2} \left[ \|x_1 - x^*\|^2_2 - \|x_{k+1} - x^*\|^2_2 \right]
\leq \frac{L}{2} \|x_1 - x^*\|^2_2.
$$

Since $\phi(x_k)$ is a decreasing sequence, it follows that

$$
\phi(x_{k+1}) - \phi^* \leq \frac{1}{k+1} \sum_{i=1}^{k+1} (\phi(x_i) - \phi^*) \leq \frac{L}{2(k+1)} \|x_1 - x^*\|^2_2.
$$

This is the well-known $O(1/k)$ rate for proximal-gradient.
Set $x \leftarrow x_k$, $x' \leftarrow x_{k+1}$, and $\alpha = 1/L$. Then add

$$\sum_{i=1}^{k+1} (\phi(x_i) - \phi^*) \leq \frac{L}{2} \sum_{i=1}^{k+1} \left[ \|x_k - x^*\|_2^2 - \|x_{i+1} - x^*\|_2^2 \right]$$

$$= \frac{L}{2} \left[ \|x_1 - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2 \right]$$

$$\leq \frac{L}{2} \|x_1 - x^*\|_2^2.$$

Since $\phi(x_k)$ is a decreasing sequence, it follows that

$$\phi(x_{k+1}) - \phi^* \leq \frac{1}{k+1} \sum_{i=1}^{k+1} (\phi(x_i) - \phi^*) \leq \frac{L}{2(k+1)} \|x_1 - x^*\|_2^2.$$

This is the well-known $O(1/k)$ rate for proximal-gradient. But for $C^1_L$ convex functions, optimal rate is $O(1/k^2)$!
Accelerated Proximal Gradient

Let $x_0 = y_0 \in \text{dom } h$. For $k \geq 1$:

\[
    x_k = \text{prox}_{\alpha_k h}(y_{k-1} - \alpha_k \nabla f(y_{k-1}))
\]

\[
    y_k = x_k + \frac{k - 1}{k + 2} (x_k - x_{k-1}).
\]


Exercise: Prove this claim!
Accelerated Proximal Gradient

Let $x_0 = y_0 \in \text{dom } h$. For $k \geq 1$:

$$x_k = \text{prox}_{\alpha_k h}(y_{k-1} - \alpha_k \nabla f(y_{k-1}))$$

$$y_k = x_k + \frac{k - 1}{k + 2} (x_k - x_{k-1}).$$


- Uses extra “memory” for interpolation
- Same computational cost as ordinary prox-grad
- Convergence rate theoretically optimal

Exercise: Prove this claim!

Suvrit Sra (suvrit@mit.edu) 6.881 Optimization for Machine Learning (03/30/21; Lecture 11)
Accelerated Proximal Gradient

Let $x_0 = y_0 \in \text{dom } h$. For $k \geq 1$:

\[
x_k = \text{prox}_{\alpha_k h}(y_{k-1} - \alpha_k \nabla f(y_{k-1}))
\]

\[
y_k = x_k + \frac{k - 1}{k + 2} (x_k - x_{k-1}).
\]


- Uses extra “memory” for interpolation
- Same computational cost as ordinary prox-grad
- Convergence rate theoretically optimal

\[
\phi(x_k) - \phi^* \leq \frac{2L}{(k + 1)^2} \| x_0 - x^* \|_2^2.
\]

Exercise: Prove this claim!
Proximal Splitting
Proximal splitting methods

\[ \ell(x) + f(x) + h(x) \]

- Direct use of prox-grad not easy
- Requires computation of: \( \text{prox}_{\lambda(f+h)} \) (i.e., \( (I + \lambda(\partial f + \partial h))^{-1} \))
Proximal splitting methods

\[ \ell(x) + f(x) + h(x) \]

- Direct use of prox-grad not easy
- Requires computation of: \( \text{prox}_\lambda(f+h) \) (i.e., \( (I + \lambda(\partial f + \partial h))^{-1} \))

Example:

\[
\min \left\{ \frac{1}{2} \|x - y\|_2^2 + \lambda \|x\|_2 + \mu \sum_{i=1}^{n-1} |x_{i+1} - x_i| \right\}.
\]

\[ f(x) \quad h(x) \]
Proximal splitting methods

\[ \ell(x) + f(x) + h(x) \]

- Direct use of prox-grad not easy
- Requires computation of: \( \text{prox}_{\lambda(f+h)} \) (i.e., \( (I + \lambda(\partial f + \partial h))^{-1} \))

**Example:**

\[
\min \quad \frac{1}{2} \| x - y \|_2^2 + \lambda \| x \|_2 + \mu \sum_{i=1}^{n-1} |x_{i+1} - x_i| \\
\begin{align*}
\text{prox}_f(x) + \text{prox}_h(x)
\end{align*}
\]

- But good feature: \( \text{prox}_f \) and \( \text{prox}_h \) separately easier
- Can we exploit that?
If \((I + \partial f + \partial h)^{-1}\) hard, but \((I + \partial f)^{-1}\) and \((I + \partial h)^{-1}\) "easy"
If \((I + \partial f + \partial h)^{-1}\) hard, but \((I + \partial f)^{-1}\) and \((I + \partial h)^{-1}\) “easy”

Derive a fixed-point equation that “splits” the operators
If \((I + \partial f + \partial h)^{-1}\) hard, but \((I + \partial f)^{-1}\) and \((I + \partial h)^{-1}\) “easy”

Derive a fixed-point equation that “splits” the operators

Assume we are solving

\[
\min \ f(x) + h(x),
\]

where both \(f\) and \(h\) are convex but potentially nondifferentiable.

**Warning:** We implicitly assumed: \(\partial (f + h) = \partial f + \partial h\).
Proximal splitting – operator notation

- If \((I + \partial f + \partial h)^{-1}\) hard, but \((I + \partial f)^{-1}\) and \((I + \partial h)^{-1}\) “easy”
- Derive a fixed-point equation that “splits” the operators

Assume we are solving

\[
\min f(x) + h(x),
\]

where both \(f\) and \(h\) are convex but potentially nondifferentiable.

Warning: We implicitly assumed: \(\partial(f + h) = \partial f + \partial h\).

Intuitive thinking

Seeking a “nice” fixed-point equation

(inspiration \(x = \text{prox}_r(x - \alpha \nabla f)\))
Proximal splitting

\[ 0 \in \partial f(x) + \partial h(x) \]
Proximal splitting

\[ 0 \in \partial f(x) + \partial h(x) \]
\[ 2x \in (I + \partial f)(x) + (I + \partial h)(x) \]
Proximal splitting

\[ 0 \in \partial f(x) + \partial h(x) \]
\[ 2x \in (I + \partial f)(x) + (I + \partial h)(x) \]

Key idea of splitting: new variable!

\[ z \in (I + \partial h)(x) \implies x = \text{prox}_h(z) \]
Proximal splitting

\[ 0 \in \partial f(x) + \partial h(x) \]
\[ 2x \in (I + \partial f)(x) + (I + \partial h)(x) \]

**Key idea of splitting: new variable!**

\[ z \in (I + \partial h)(x) \implies x = \text{prox}_h(z) \]
\[ 2x - z \in (I + \partial f)(x) \]
Proximal splitting

\[ 0 \in \partial f(x) + \partial h(x) \]
\[ 2x \in (I + \partial f)(x) + (I + \partial h)(x) \]

Key idea of splitting: new variable!

\[ z \in (I + \partial h)(x) \implies x = \text{prox}_h(z) \]

\[ 2x - z \in (I + \partial f)(x) \implies x \in (I + \partial f)^{-1}(2x - z) \]
Proximal splitting

\[ 0 \in \partial f(x) + \partial h(x) \]
\[ 2x \in (I + \partial f)(x) + (I + \partial h)(x) \]

**Key idea of splitting: new variable!**

\[ z \in (I + \partial h)(x) \implies x = \text{prox}_h(z) \]

\[ 2x - z \in (I + \partial f)(x) \implies x \in (I + \partial f)^{-1}(2x - z) \]

▶ Not a fixed-point equation yet
Proximal splitting

\[ 0 \in \partial f(x) + \partial h(x) \]
\[ 2x \in (I + \partial f)(x) + (I + \partial h)(x) \]

Key idea of splitting: new variable!

\[ z \in (I + \partial h)(x) \implies x = \text{prox}_h(z) \]

\[ 2x - z \in (I + \partial f)(x) \implies x \in (I + \partial f)^{-1}(2x - z) \]

- Not a fixed-point equation yet
- We need one more idea
Douglas-Rachford splitting

Reflection operator

\[ R_h(z) := 2 \text{prox}_h(z) - z \]
Douglas-Rachford splitting

Reflection operator

\[ R_h(z) := 2 \text{prox}_h(z) - z \]

Douglas-Rachford method

\[ z \in (I + \partial h)(x), \quad x = \text{prox}_h(z) \]
Douglas-Rachford splitting

Reflection operator

\[ R_h(z) := 2 \text{prox}_h(z) - z \]

Douglas-Rachford method

\[ z \in (I + \partial h)(x), \quad x = \text{prox}_h(z) \implies R_h(z) = 2x - z \]
Douglas-Rachford splitting

**Reflection operator**

\[ R_h(z) := 2 \text{prox}_h(z) - z \]

**Douglas-Rachford method**

\[ z \in (I + \partial h)(x), \quad x = \text{prox}_h(z) \implies R_h(z) = 2x - z \]

\[ 0 \in \partial f(x) + \partial g(x) \]

\[ 2x \in (I + \partial f)(x) + (I + \partial g)(x) \]

\[ 2x - z \in (I + \partial f)(x) \]
Douglas-Rachford splitting

Reflection operator

\[ R_h(z) := 2 \text{prox}_h(z) - z \]

Douglas-Rachford method

\[ z \in (I + \partial h)(x), \quad x = \text{prox}_h(z) \implies R_h(z) = 2x - z \]

\[ 0 \in \partial f(x) + \partial g(x) \]

\[ 2x \in (I + \partial f)(x) + (I + \partial g)(x) \]

\[ 2x - z \in (I + \partial f)(x) \]

\[ x = \text{prox}_f(R_h(z)) \]
Douglas-Rachford splitting

Reflection operator

\[ R_h(z) := 2 \text{prox}_h(z) - z \]

Douglas-Rachford method

\[ z \in (I + \partial h)(x), \quad x = \text{prox}_h(z) \implies R_h(z) = 2x - z \]

\[ 0 \in \partial f(x) + \partial g(x) \]

\[ 2x \in (I + \partial f)(x) + (I + \partial g)(x) \]

\[ 2x - z \in (I + \partial f)(x) \]

\[ x = \text{prox}_f(R_h(z)) \]

but \[ R_h(z) = 2x - z \implies \]

\[ z = 2x - R_h(z) \]
Douglas-Rachford splitting

Reflection operator

\[ R_h(z) := 2 \text{prox}_h(z) - z \]

Douglas-Rachford method

\[ z \in (I + \partial h)(x), \quad x = \text{prox}_h(z) \implies R_h(z) = 2x - z \]

\[ 0 \in \partial f(x) + \partial g(x) \]

\[ 2x \in (I + \partial f)(x) + (I + \partial g)(x) \]

\[ 2x - z \in (I + \partial f)(x) \]

\[ x = \text{prox}_f(R_h(z)) \]

but \[ R_h(z) = 2x - z \implies \]

\[ z = 2x - R_h(z) \]

\[ z = 2 \text{prox}_f(R_h(z)) - R_h(z) = \]
Douglas-Rachford splitting

Reflection operator

\[ R_h(z) := 2 \text{prox}_h(z) - z \]

Douglas-Rachford method

\[ z \in (I + \partial h)(x), \quad x = \text{prox}_h(z) \implies R_h(z) = 2x - z \]

\[ 0 \in \partial f(x) + \partial g(x) \]

\[ 2x \in (I + \partial f)(x) + (I + \partial g)(x) \]

\[ 2x - z \in (I + \partial f)(x) \]

\[ x = \text{prox}_f(R_h(z)) \]

but \[ R_h(z) = 2x - z \implies \]

\[ z = 2x - R_h(z) \]

\[ z = 2 \text{prox}_f(R_h(z)) - R_h(z) = R_f(R_h(z)) \]

Finally, \( z \) is on both sides of the eqn
Douglas-Rachford method

\[
0 \in \partial f(x) + \partial h(x) \iff \begin{cases} 
    x = \text{prox}_h(z) \\
    z = R_f(R_h(z)) 
\end{cases}
\]

**DR method:** given \( z_0 \), iterate for \( k \geq 0 \)

\[
\begin{align*}
    x_k &= \text{prox}_h(z_k) \\
    v_k &= \text{prox}_f(2x_k - z_k) \\
    z_{k+1} &= z_k + \gamma_k(v_k - x_k)
\end{align*}
\]
Douglas-Rachford method

\[ 0 \in \partial f(x) + \partial h(x) \iff \begin{cases} x = \text{prox}_h(z) \\ z = R_f(R_h(z)) \end{cases} \]

**DR method:** given \( z_0 \), iterate for \( k \geq 0 \)

\[
\begin{align*}
x_k &= \text{prox}_h(z_k) \\
v_k &= \text{prox}_f(2x_k - z_k) \\
z_{k+1} &= z_k + \gamma_k(v_k - x_k)
\end{align*}
\]

**Theorem.** If \( f + h \) admits minimizers, and \( (\gamma_k) \) satisfy

\[ \gamma_k \in [0, 2], \quad \sum_k \gamma_k(2 - \gamma_k) = \infty, \]

then the DR-iterates \( v_k \) and \( x_k \) converge to a minimizer.
Douglas-Rachford method

For \( \gamma_k = 1 \), we have

\[
    z_{k+1} = z_k + \nu_k - x_k
\]

\[
    z_{k+1} = z_k + \text{prox}_f(2 \text{prox}_h(z_k) - z_k) - \text{prox}_h(z_k)
\]
Douglas-Rachford method

For $\gamma_k = 1$, we have

$$z_{k+1} = z_k + v_k - x_k$$

$$z_{k+1} = z_k + \text{prox}_f(2 \text{prox}_h(z_k) - z_k) - \text{prox}_h(z_k)$$

Dropping superscripts, writing $P \equiv \text{prox}$, we have

$$z \leftarrow Tz$$

$$T = I + P_f(2P_h - I) - P_h$$
Douglas-Rachford method

For $\gamma_k = 1$, we have

\[ z_{k+1} = z_k + v_k - x_k \]
\[ z_{k+1} = z_k + \text{prox}_f(2 \text{prox}_h(z_k) - z_k) - \text{prox}_h(z_k) \]

Dropping superscripts, writing $P \equiv \text{prox}$, we have

\[
\begin{align*}
   z & \leftarrow Tz \\
   T & = I + P_f(2P_h - I) - P_h
\end{align*}
\]

**Lemma** DR can be written as: $z \leftarrow \frac{1}{2}(R_fR_h + I)z$, where $R_f$ denotes the reflection operator $2P_f - I$ (similarly $R_h$).

**Exercise:** Prove this claim.
Best approximation problem

\[ \min \delta_A(x) + \delta_B(x) \quad \text{where} \quad A \cap B = \emptyset. \]
Best approximation problem

\[ \min \delta_A(x) + \delta_B(x) \quad \text{where } A \cap B = \emptyset. \]

Can we use DR?
**Best approximation problem**

\[
\min \quad \delta_A(x) + \delta_B(x) \quad \text{where } A \cap B = \emptyset.
\]

**Can we use DR?**

Using a clever analysis of Bauschke & Combettes (2004), DR can still be applied! However, it generates diverging iterates that can be “projected back” to obtain a solution to

\[
\min \quad \|a - b\|_2 \quad a \in A, b \in B.
\]

See: Jegelka, Bach, Sra (NIPS 2013) for an example.
Exercise

Best approximation problem

$$\min_x d^2_A(x) + d^2_B(x),$$

where $d_A(x) := \inf \{\|z - x\|_2 \mid z \in A\}$ is the distance function.

**Exercise:** Show that $R_{d_A} = P_A$ (i.e., projection onto $A$!)
Exercise

**Best approximation problem**

\[
\min_x \ d_A^2(x) + d_B^2(x),
\]

where \(d_A(x) := \inf \{\|z - x\|_2 \mid z \in A\}\) is the distance function.

**Exercise:** Show that \(R_{d_A} = P_A\) (i.e., projection onto \(A\)!

Thus, DR for solving above problem becomes

\[
z_{k+1} = \frac{1}{2}(P_A P_B + I)z_k, \quad k \geq 0.
\]

**Exercise:** Convergence rate of above method?
Three operator splitting

\[
\min_x f(x) + g(x) + h(x)
\]

Not so easy for DR-splitting for general \( f \).
Three operator splitting

\[
\min_x f(x) + g(x) + h(x)
\]

Not so easy for DR-splitting for general \( f \).

1. Initialize \( y^0 \in \mathbb{R}^n \)
2. For \( k \geq 0 \), iterate:

\[
\begin{align*}
    z^k &= \text{prox}_{\gamma h}(y^k) \\
    x^k &= \text{prox}_{\gamma g}(2z^k - y^k - \gamma \nabla f(z^k)) \\
    y^{k+1} &= y^k + x^k - z^k
\end{align*}
\]
Three operator splitting

\[ \min_x f(x) + g(x) + h(x) \]

Not so easy for DR-splitting for general \( f \).

1. Initialize \( y^0 \in \mathbb{R}^n \)
2. For \( k \geq 0 \), iterate:
   \[
   z^k = \text{prox}_{\gamma h}(y^k) \\
   x^k = \text{prox}_{\gamma g}(2z^k - y^k - \gamma \nabla f(z^k)) \\
   y^{k+1} = y^k + x^k - z^k
   \]

Operator notation

\[
y^{k+1} \leftarrow [\text{Id} - J_{\gamma h} + J_{\gamma g} \circ (2J_{\gamma h} - \text{Id} - \gamma \nabla f \circ J_{\gamma h})](y^k),
\]

where \( J_{\gamma h} \) denotes the operator \( \text{prox}_{\gamma h} \).