Optimization for Machine Learning

Lecture 11: Proximal methods

6.881: MIT

Suvrit Sra
Massachusetts Institute of Technology

30 Mar, 2021

Motivation

(nonsmooth optimization)

Regularized / Composite Objectives

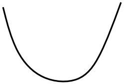
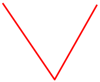
Many nonsmooth problems take the form

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

Regularized / Composite Objectives

Many nonsmooth problems take the form

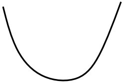
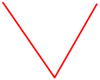
$$\boxed{\text{minimize } \phi(x) := f(x) + \color{red}{r(x)}}$$

$f \in$  $+ \quad r \in$ 

Regularized / Composite Objectives

Many nonsmooth problems take the form

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

$f \in$  $+ \quad r \in$ 

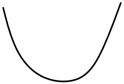
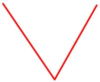
Example: $f(x) = \frac{1}{2}\|Ax - b\|^2$ and $r(x) = \lambda\|x\|_1$

Lasso, L1-LS, compressed sensing

Regularized / Composite Objectives

Many nonsmooth problems take the form

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

$f \in$  $+\ r \in$ 

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

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

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

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

Composite objective minimization

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

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

Composite objective minimization

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

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

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

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

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

Nesterov: **exploit** smoothness of f to beat lower bound!

Proximal gradient method

Optimality conditions

$$0 \quad \in \quad \nabla f(x^*) + \partial r(x^*)$$

Proximal gradient method

Optimality conditions

$$
\begin{aligned}
0 &\in \nabla f(x^*) + \partial r(x^*) \\
0 &\in \alpha \nabla f(x^*) + \alpha \partial r(x^*)
\end{aligned}
$$

Proximal gradient method

Optimality conditions

$$\begin{aligned} 0 &\in \nabla f(x^*) + \partial r(x^*) \\ 0 &\in \alpha\nabla f(x^*) + \alpha\partial r(x^*) \\ x^* &\in \alpha\nabla f(x^*) + (I + \alpha\partial r)(x^*) \end{aligned}$$

Proximal gradient method

Optimality conditions

$$\begin{aligned} 0 &\in \nabla f(x^*) + \partial r(x^*) \\ 0 &\in \alpha\nabla f(x^*) + \alpha\partial r(x^*) \\ x^* &\in \alpha\nabla f(x^*) + (I + \alpha\partial r)(x^*) \\ x^* - \alpha\nabla f(x^*) &\in (I + \alpha\partial r)(x^*) \end{aligned}$$

Proximal gradient method

Optimality conditions

$$\begin{aligned}
0 &\in \nabla f(x^*) + \partial r(x^*) \\
0 &\in \alpha\nabla f(x^*) + \alpha\partial r(x^*) \\
x^* &\in \alpha\nabla f(x^*) + (I + \alpha\partial r)(x^*) \\
x^* - \alpha\nabla f(x^*) &\in (I + \alpha\partial r)(x^*) \\
x^* &= (I + \alpha\partial r)^{-1}(x^* - \alpha\nabla f(x^*))
\end{aligned}$$

Proximal gradient method

Optimality conditions

$$
\begin{aligned}
0 &\in \nabla f(x^*) + \partial r(x^*) \\
0 &\in \alpha\nabla f(x^*) + \alpha\partial r(x^*) \\
x^* &\in \alpha\nabla f(x^*) + (I + \alpha\partial r)(x^*) \\
x^* - \alpha\nabla f(x^*) &\in (I + \alpha\partial r)(x^*) \\
x^* &= (I + \alpha\partial r)^{-1}(x^* - \alpha\nabla f(x^*)) \\
x^* &= \operatorname{prox}_{\alpha r}(x^* - \alpha\nabla f(x^*))
\end{aligned}
$$

Proximal gradient method

Optimality conditions

$$
\begin{aligned}
0 &\in \nabla f(x^*) + \partial r(x^*) \\
0 &\in \alpha \nabla f(x^*) + \alpha \partial r(x^*) \\
x^* &\in \alpha \nabla f(x^*) + (I + \alpha \partial r)(x^*) \\
x^* - \alpha \nabla f(x^*) &\in (I + \alpha \partial r)(x^*) \\
x^* &= (I + \alpha \partial r)^{-1}(x^* - \alpha \nabla f(x^*)) \\
x^* &= \mathrm{prox}_{\alpha r}(x^* - \alpha \nabla f(x^*))
\end{aligned}
$$

Above fixed-point eqn suggests iteration

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

$$\boxed{\text{This method converges as } O(1/k) \text{ for convex } f \in C_L^1!}$$

Prox operators

From projections to proximity

Let $\mathbb{1}_{\mathcal{X}}$ be the *indicator function* for closed, cvx $\mathcal{X}$.

From projections to proximity

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

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

From projections to proximity

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

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

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

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

From projections to proximity

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

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

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

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

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

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

Proximity operator

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

Proximity operator

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

Unique solution due to strong convexity. Observe that:

$$\begin{aligned} 0 &\in x - y + \partial r(x) \\ y &\in (\mathrm{Id} + \partial r)(x) \\ x &= (\mathrm{Id} + \partial r)^{-1}(y) \\ x &= \mathrm{prox}_r(y). \end{aligned}$$

Exercise: proximity operator for ℓ_1

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 $x \neq 0$

Aka: Soft-thresholding operator

Proximity operators

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

Proximity operators

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

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

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

Exercise: Prove the above property.

Proximity operators

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

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

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

Exercise: Prove the above property.

Corollary. The operator prox_r is **nonexpansive**

Proof: apply Cauchy-Schwarz to FNE.

Details: FNE for projections

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,

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

Details: FNE for projections

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

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

Details: FNE for projections

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

$$\begin{aligned} \langle x_1 - P_C(x_1), P_C(x_2) - P_C(x_1) \rangle &\leq 0 \\ \langle x_2 - P_C(x_2), P_C(x_1) - P_C(x_2) \rangle &\leq 0 \end{aligned}$$

Details: FNE for projections

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

$$\begin{aligned}
\langle x_1 - P_C(x_1), P_C(x_2) - P_C(x_1) \rangle &\leq 0 \\
\langle x_2 - P_C(x_2), P_C(x_1) - P_C(x_2) \rangle &\leq 0 \\
\langle P_C(x_1) - P_C(x_2), x_2 - x_1 + P_C(x_1) - P_C(x_2) \rangle &\leq 0.
\end{aligned}$$

Details: FNE for projections

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

$$\begin{aligned}
\langle x_1 - P_C(x_1), P_C(x_2) - P_C(x_1) \rangle &\leq 0 \\
\langle x_2 - P_C(x_2), P_C(x_1) - P_C(x_2) \rangle &\leq 0 \\
\langle P_C(x_1) - P_C(x_2), x_2 - x_1 + P_C(x_1) - P_C(x_2) \rangle &\leq 0.
\end{aligned}$$

$$\boxed{\|P_C(x_1) - P_C(x_2)\|_2^2 \leq \langle P_C(x_1) - P_C(x_2), x_1 - x_2 \rangle}$$

Consequences of FNE

Projected gradient method

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

Proximal gradient method

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

Same convergence theory goes through!

Consequences of FNE

Projected gradient method

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

Proximal gradient method

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

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

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

Consequences of FNE

Projected gradient method

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

Proximal gradient method

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

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

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

Krasnoselskii-Mann theorem: If a FNE map on a closed convex set has a fixed-point, then the iteration $x_{k+1} \leftarrow (1-\alpha_k)\mathrm{Id} + \alpha_k F(x_k)$ converges to it for $\alpha_k \in [0,1]$ provided $\sum_k \alpha_k(1-\alpha_k) = \infty$ for any starting point x_0.
See: C. Byrne (2003). *"A unified treatment of some iterative algorithms in signal processing and image reconstruction."*

Exercise: Moreau Decomposition

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

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

Exercise: Moreau Decomposition

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

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

Proof sketch:

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

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

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

Proximal-Gradient

$$\min f(x) + h(x)$$

Why does prox-grad method work?

$$x_{k+1} = \operatorname{prox}_{\alpha_k h}(x_k - \alpha_k \nabla f(x_k))$$

Why does prox-grad method work?

$$\begin{aligned} x_{k+1} &= \operatorname{prox}_{\alpha_k h}(x_k - \alpha_k \nabla f(x_k)) \\ x_{k+1} &= x_k - \alpha_k G_{\alpha_k}(x_k). \end{aligned}$$

Why does prox-grad method work?

$$\begin{aligned} x_{k+1} &= \operatorname{prox}_{\alpha_k h}(x_k - \alpha_k \nabla f(x_k)) \\ x_{k+1} &= x_k - \alpha_k G_{\alpha_k}(x_k). \end{aligned}$$

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?

$$
\begin{aligned}
x_{k+1} &= \operatorname{prox}_{\alpha_k h}(x_k - \alpha_k \nabla f(x_k)) \\
x_{k+1} &= x_k - \alpha_k G_{\alpha_k}(x_k).
\end{aligned}
$$

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_α analogous to ∇f
- If x locally optimal, then $G_\alpha(x) = 0$ (nonconvex f)

Convergence analysis

Assumption: Lipschitz continuous gradient; denoted $f \in C_L^1$

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

Lemma (Descent). Let $f \in C_L^1$. Then,

$$f(y) \le f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{L}{2}\|y - x\|_2^2$$

Descent lemma – corollary

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

Descent lemma – corollary

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

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

Descent lemma – corollary

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

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

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

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

Descent lemma – corollary

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

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

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

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

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

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

Exercise: Prove! (*Hint:* f, h are cvx, and $G_\alpha(x) - \nabla f(x) \in \partial h(y)$)

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

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

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

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

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:

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

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

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

Exercise: Argue convergence via this inequality.

Plug $z = x^*$ in
$f(y) + h(y) \le f(z) + h(z) + \langle G_\alpha(x), x - z\rangle - \frac{\alpha}{2}\|G_\alpha(x)\|_2^2$, to obtain progress in terms of iterates:

$$\begin{aligned}
\phi(x') - \phi^* &\le \langle G_\alpha(x), x - x^*\rangle - \tfrac{\alpha}{2}\|G_\alpha(x)\|_2^2 \\
&= \frac{1}{2\alpha}\left[2\langle \alpha G_\alpha(x), x - x^*\rangle - \|\alpha G_\alpha(x)\|_2^2\right] \\
&= \frac{1}{2\alpha}\left[\|x - x^*\|_2^2 - \|x - x^* - \alpha G_\alpha(x)\|_2^2\right]
\end{aligned}$$

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

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

Exercise: Argue convergence via this inequality.
Plug $z = x^*$ in
$f(y) + h(y) \le f(z) + h(z) + \langle G_\alpha(x),\, x - z\rangle - \frac{\alpha}{2}\|G_\alpha(x)\|_2^2$, to obtain progress in terms of iterates:

$$\begin{aligned}
\phi(x') - \phi^* &\le \langle G_\alpha(x),\, x - x^*\rangle - \tfrac{\alpha}{2}\|G_\alpha(x)\|_2^2 \\
&= \frac{1}{2\alpha}\left[2\langle \alpha G_\alpha(x),\, x - x^*\rangle - \|\alpha G_\alpha(x)\|_2^2\right] \\
&= \frac{1}{2\alpha}\left[\|x - x^*\|_2^2 - \|x - x^* - \alpha G_\alpha(x)\|_2^2\right] \\
&= \frac{1}{2\alpha}\left[\|x - x^*\|_2^2 - \|x' - x^*\|_2^2\right].
\end{aligned}$$

Convergence rate

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

Convergence rate

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

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

Convergence rate

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

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

Convergence rate

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

$$\begin{aligned}
\sum\nolimits_{i=1}^{k+1} (\phi(x_i) - \phi^*) &\leq \tfrac{L}{2} \sum\nolimits_{i=1}^{k+1} \left[\|x_k - x*\|_2^2 - \|x_{i+1} - x*\|_2^2\right] \\
&= \tfrac{L}{2} \left[\|x_1 - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2\right] \\
&\leq \tfrac{L}{2} \|x_1 - x^*\|_2^2.
\end{aligned}$$

Convergence rate

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

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

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.

Convergence rate

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

$$\begin{aligned}
\sum\nolimits_{i=1}^{k+1}(\phi(x_i) - \phi^*) &\leq \tfrac{L}{2}\sum\nolimits_{i=1}^{k+1}\left[\|x_k - x*\|_2^2 - \|x_{i+1} - x*\|_2^2\right] \\
&= \tfrac{L}{2}\left[\|x_1 - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2\right] \\
&\leq \tfrac{L}{2}\|x_1 - x^*\|_2^2.
\end{aligned}$$

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_L^1 convex functions, optimal rate is $O(1/k^2)$!

Accelerated Proximal Gradient

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

$$x_k = \operatorname{prox}_{\alpha_k h}(y_{k-1} - \alpha_k \nabla f(y_{k-1}))$$
$$y_k = x_k + \frac{k-1}{k+2}(x_k - x_{k-1}).$$

Framework due to: Nesterov (1983, 2004); also Beck, Teboulle (2009).
Simplified analysis: Tseng (2008).

Accelerated Proximal Gradient

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

$$\begin{aligned} x_k &= \operatorname{prox}_{\alpha_k h}(y_{k-1} - \alpha_k \nabla f(y_{k-1})) \\ y_k &= x_k + \frac{k-1}{k+2}(x_k - x_{k-1}). \end{aligned}$$

Framework due to: Nesterov (1983, 2004); also Beck, Teboulle (2009).
Simplified analysis: Tseng (2008).

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

Accelerated Proximal Gradient

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

$$x_k = \operatorname{prox}_{\alpha_k h}(y_{k-1} - \alpha_k \nabla f(y_{k-1}))$$
$$y_k = x_k + \frac{k-1}{k+2}(x_k - x_{k-1}).$$

Framework due to: Nesterov (1983, 2004); also Beck, Teboulle (2009).
Simplified analysis: Tseng (2008).

- 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

$$\boxed{\ell(x) + f(x) + h(x)}$$

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

Proximal splitting methods

$$\boxed{\ell(x) + f(x) + h(x)}$$

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

Example:

$$\min \quad \tfrac{1}{2}\|x - y\|_2^2 + \underbrace{\lambda\|x\|_2}_{f(x)} + \underbrace{\mu \sum_{i=1}^{n-1} |x_{i+1} - x_i|}_{h(x)}.$$

Proximal splitting methods

$$\boxed{\ell(x) + f(x) + h(x)}$$

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

Example:

$$\min \quad \tfrac{1}{2}\|x - y\|_2^2 + \underbrace{\lambda\|x\|_2}_{f(x)} + \underbrace{\mu \sum\nolimits_{i=1}^{n-1} |x_{i+1} - x_i|}_{h(x)}.$$

- But good feature: prox_f and prox_h separately easier
- Can we exploit that?

Proximal splitting – operator notation

- If $(I + \partial f + \partial h)^{-1}$ hard, but $(I + \partial f)^{-1}$ and $(I + \partial h)^{-1}$ "easy"

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

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

$$\min \quad f(x) + h(x),$$

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

$$\min \quad f(x) + h(x),$$

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 $x = \mathrm{prox}_r(x - \alpha \nabla f)$)

Proximal splitting

$$0 \quad \in \quad \partial f(x) + \partial h(x)$$

Proximal splitting

$$\begin{aligned} 0 &\in \partial f(x) + \partial h(x) \\ 2x &\in (I + \partial f)(x) + (I + \partial h)(x) \end{aligned}$$

Proximal splitting

$$\begin{aligned} 0 &\in \partial f(x) + \partial h(x) \\ 2x &\in (I + \partial f)(x) + (I + \partial h)(x) \end{aligned}$$

Key idea of splitting: new variable!

$$z \in (I + \partial h)(x) \implies x = \operatorname{prox}_h(z)$$

Proximal splitting

$$\begin{aligned} 0 &\in \partial f(x) + \partial h(x) \\ 2x &\in (I + \partial f)(x) + (I + \partial h)(x) \end{aligned}$$

Key idea of splitting: new variable!

$$z \in (I + \partial h)(x) \implies x = \mathrm{prox}_h(z)$$

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

Proximal splitting

$$
\begin{aligned}
0 &\in \partial f(x) + \partial h(x) \\
2x &\in (I + \partial f)(x) + (I + \partial h)(x)
\end{aligned}
$$

Key idea of splitting: new variable!

$$z \in (I + \partial h)(x) \implies x = \mathrm{prox}_h(z)$$

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

Proximal splitting

$$\begin{aligned} 0 &\in \partial f(x) + \partial h(x) \\ 2x &\in (I + \partial f)(x) + (I + \partial h)(x) \end{aligned}$$

Key idea of splitting: new variable!

$$z \in (I + \partial h)(x) \implies x = \operatorname{prox}_h(z)$$

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

- Not a fixed-point equation yet

Proximal splitting

$$\begin{aligned} 0 &\in \partial f(x) + \partial h(x) \\ 2x &\in (I + \partial f)(x) + (I + \partial h)(x) \end{aligned}$$

Key idea of splitting: new variable!

$$z \in (I + \partial h)(x) \implies x = \mathrm{prox}_h(z)$$

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

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

Douglas-Rachford splitting

Reflection operator

$$R_h(z) := 2\,\mathrm{prox}_h(z) - z$$

Douglas-Rachford splitting

Reflection operator

$$R_h(z) := 2\,\mathrm{prox}_h(z) - z$$

Douglas-Rachford method

$$z \in (I + \partial h)(x), \quad x = \mathrm{prox}_h(z)$$

Douglas-Rachford splitting

Reflection operator

$$R_h(z) := 2\,\mathrm{prox}_h(z) - z$$

Douglas-Rachford method

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

Douglas-Rachford splitting

Reflection operator

$$R_h(z) := 2\,\mathrm{prox}_h(z) - z$$

Douglas-Rachford method

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

$$\begin{aligned} 0 &\in \partial f(x) + \partial g(x) \\ 2x &\in (I + \partial f)(x) + (I + \partial g)(x) \\ 2x - z &\in (I + \partial f)(x) \end{aligned}$$

Douglas-Rachford splitting

Reflection operator

$$R_h(z) := 2\,\mathrm{prox}_h(z) - z$$

Douglas-Rachford method

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

$$\begin{aligned}
0 &\in \partial f(x) + \partial g(x) \\
2x &\in (I + \partial f)(x) + (I + \partial g)(x) \\
2x - z &\in (I + \partial f)(x) \\
x &= \mathrm{prox}_f(R_h(z))
\end{aligned}$$

Douglas-Rachford splitting

Reflection operator

$$R_h(z) := 2\,\mathrm{prox}_h(z) - z$$

Douglas-Rachford method

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

$$\begin{aligned}
0 &\in \partial f(x) + \partial g(x) \\
2x &\in (I + \partial f)(x) + (I + \partial g)(x) \\
2x - z &\in (I + \partial f)(x) \\
x &= \mathrm{prox}_f(R_h(z)) \\
\text{but } R_h(z) &= 2x - z \implies \\
z &= 2x - R_h(z)
\end{aligned}$$

Douglas-Rachford splitting

Reflection operator

$$R_h(z) := 2\,\mathrm{prox}_h(z) - z$$

Douglas-Rachford method

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

$$\begin{aligned}
0 &\in \partial f(x) + \partial g(x) \\
2x &\in (I + \partial f)(x) + (I + \partial g)(x) \\
2x - z &\in (I + \partial f)(x) \\
x &= \mathrm{prox}_f(R_h(z)) \\
\text{but } R_h(z) &= 2x - z \implies \\
z &= 2x - R_h(z) \\
z &= 2\,\mathrm{prox}_f(R_h(z)) - R_h(z) =
\end{aligned}$$

Douglas-Rachford splitting

Reflection operator

$$R_h(z) := 2\,\mathrm{prox}_h(z) - z$$

Douglas-Rachford method

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

$$\begin{aligned}
0 &\in \partial f(x) + \partial g(x) \\
2x &\in (I + \partial f)(x) + (I + \partial g)(x) \\
2x - z &\in (I + \partial f)(x) \\
x &= \mathrm{prox}_f(R_h(z)) \\
\text{but } R_h(z) &= 2x - z \implies \\
z &= 2x - R_h(z) \\
z &= 2\,\mathrm{prox}_f(R_h(z)) - R_h(z) = R_f(R_h(z))
\end{aligned}$$

$$\boxed{\text{Finally, } z \text{ is on both sides of the eqn}}$$

Douglas-Rachford method

$$\boxed{0 \in \partial f(x) + \partial h(x) \Leftrightarrow \begin{cases} x = \mathrm{prox}_h(z) \\ z = R_f(R_h(z)) \end{cases}}$$

DR method: given z_0, iterate for $k \geq 0$

$$\begin{aligned} x_k &= \mathrm{prox}_h(z_k) \\ v_k &= \mathrm{prox}_f(2x_k - z_k) \\ z_{k+1} &= z_k + \gamma_k(v_k - x_k) \end{aligned}$$

Douglas-Rachford method

$$0 \in \partial f(x) + \partial h(x) \Leftrightarrow \begin{cases} x = \operatorname{prox}_h(z) \\ z = R_f(R_h(z)) \end{cases}$$

DR method: given z_0, iterate for $k \geq 0$

$$\begin{aligned} x_k &= \operatorname{prox}_h(z_k) \\ v_k &= \operatorname{prox}_f(2x_k - z_k) \\ z_{k+1} &= z_k + \gamma_k(v_k - x_k) \end{aligned}$$

Theorem. If $f + h$ admits minimizers, and (γ_k) satisfy

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

then the DR-iterates v_k and x_k converge to a minimizer.

Douglas-Rachford method

For $\gamma_k = 1$, we have

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

Douglas-Rachford method

For $\gamma_k = 1$, we have

$$z_{k+1} = z_k + v_k - x_k$$
$$z_{k+1} = z_k + \mathrm{prox}_f(2\,\mathrm{prox}_h(z_k) - z_k) - \mathrm{prox}_h(z_k)$$

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

$$\boxed{\begin{gathered} z \leftarrow Tz \\ T = I + P_f(2P_h - I) - P_h \end{gathered}}$$

Douglas-Rachford method

For $\gamma_k = 1$, we have

$$z_{k+1} = z_k + v_k - x_k$$
$$z_{k+1} = z_k + \mathrm{prox}_f(2\,\mathrm{prox}_h(z_k) - z_k) - \mathrm{prox}_h(z_k)$$

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

$$\boxed{\begin{gathered} z \leftarrow Tz \\ T = I + P_f(2P_h - I) - P_h \end{gathered}}$$

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

Exercise: Prove this claim.

Best approximation problem

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

Best approximation problem

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

$$\boxed{\text{Can we use DR?}}$$

Best approximation problem

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

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

$$\min \quad \|a - b\|_2 \qquad a \in A, b \in B.$$

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

Exercise

Best approximation problem

$$\min_x \quad d_A^2(x) + d_B^2(x),$$

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!)

Exercise

Best approximation problem

$$\min_x \quad d_A^2(x) + d_B^2(x),$$

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

$$z_{k+1} = \tfrac{1}{2}(P_A P_B + I) z_k, \qquad k \geq 0.$$

Exercise: * Convergence rate of above method?

Three operator splitting

$$\min_x \quad f(x) + g(x) + h(x)$$

Not so easy for DR-splitting for general f.

Three operator splitting

$$\min_x \quad f(x) + g(x) + h(x)$$

Not so easy for DR-splitting for general f.

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

$$\begin{aligned} z^k &= \operatorname{prox}_{\gamma h}(y^k) \\ x^k &= \operatorname{prox}_{\gamma g}(2z^k - y^k - \gamma \nabla f(z^k)) \\ y^{k+1} &= y^k + x^k - z^k \end{aligned}$$

Three operator splitting

$$\min_x \quad f(x) + g(x) + h(x)$$

Not so easy for DR-splitting for general f.

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

$$\begin{aligned} z^k &= \operatorname{prox}_{\gamma h}(y^k) \\ x^k &= \operatorname{prox}_{\gamma g}(2z^k - y^k - \gamma \nabla f(z^k)) \\ y^{k+1} &= y^k + x^k - z^k \end{aligned}$$

Operator notation

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

where $J_{\gamma h}$ denotes the operator $\operatorname{prox}_{\gamma h}$.