Stochastic and incremental methods

(Optml++ Meeting 3)

Suvrit Sra
Massachusetts Institute of Technology

OPTML++, Fall 2015

Outline

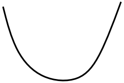
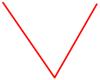
– Lect 1: Recap on convexity
– Lect 1: Recap on duality, optimality
– Lect 2: First-order optimization algorithms
– Today: Operator splitting
– Next: Stochastic and incremental methods

Composite objectives

Composite objectives

Frequently nonsmooth problems take the form

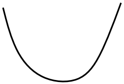
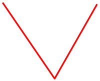
$$\text{minimize } f(x) := \ell(x) + r(x)$$

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

Composite objectives

Frequently nonsmooth problems take the form

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

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

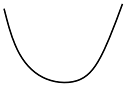
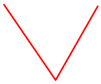
Example: $\ell(x) = \frac{1}{2}\|Ax - b\|^2$ and $r(x) = \lambda\|x\|_1$

Lasso, L1-LS, compressed sensing

Composite objectives

Frequently nonsmooth problems take the form

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

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

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

Lasso, L1-LS, compressed sensing

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

L1-Logistic regression, sparse LR

Composite objective minimization

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

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

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!

Proximal Gradient Method

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

Projected gradient

$$x \leftarrow P_{\mathcal{X}}(x - \alpha \nabla f(x))$$

Proximal Gradient Method

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

Projected gradient

$$x \leftarrow P_{\mathcal{X}}(x - \alpha \nabla f(x))$$

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

Proximal gradient

$$x \leftarrow \text{prox}_{\alpha h}(x - \alpha \nabla f(x))$$

$\text{prox}_{\alpha h}$ denotes **Euclidean** proximity operator for h

Proximal Gradient Method

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

Projected gradient

$$x \leftarrow P_{\mathcal{X}}(x - \alpha \nabla f(x))$$

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

Proximal gradient

$$x \leftarrow \text{prox}_{\alpha h}(x - \alpha \nabla f(x))$$

$\text{prox}_{\alpha h}$ denotes **Euclidean** proximity operator for h

Non-Euclidean prox-operators also studied

Proximity operator

Projection

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

Proximity operator

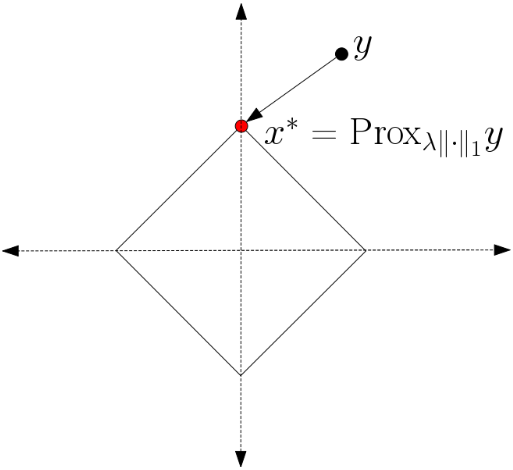
Projection

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

Proximity: Replace $\mathbb{1}_{\mathcal{X}}$ by a closed 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



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

Proximity operators

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

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

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: either $x = 0$ or $x \neq 0$

Aka: Soft-thresholding operator

Where does it come from?

Lemma $x^* = \text{prox}_{\alpha h}(x^* - \alpha \nabla f(x^*)), \forall \alpha > 0$

Where does it come from?

Lemma $x^* = \text{prox}_{\alpha h}(x^* - \alpha\nabla f(x^*)), \forall\alpha > 0$

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

Where does it come from?

Lemma $x^* = \text{prox}_{\alpha h}(x^* - \alpha \nabla f(x^*)), \forall \alpha > 0$

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

Where does it come from?

Lemma $x^* = \text{prox}_{\alpha h}(x^* - \alpha \nabla f(x^*)), \forall \alpha > 0$

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

Where does it come from?

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

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

Where does it come from?

Lemma $x^* = \text{prox}_{\alpha h}(x^* - \alpha\nabla f(x^*)), \forall\alpha > 0$

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

Where does it come from?

Lemma $x^* = \text{prox}_{\alpha h}(x^* - \alpha \nabla f(x^*)), \forall \alpha > 0$

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

Where does it come from?

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

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

Above fixed-point eqn suggests iteration

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

Why does it 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 it work?

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

Gradient mapping: the "gradient-like object"

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

Why does it 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"

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

- Our lemma shows: $G_\alpha(x) = 0$ if and only if x is optimal
- So G_α analogous to ∇f
- If x locally optimal, then $G_\alpha(x) = 0$ (nonconvex f)
- Analysis yields $O(1/k)$ convergence

Faster methods

Optimal gradient methods

♠ Efficiency estimates for the gradient method:

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

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

Optimal gradient methods

♠ Efficiency estimates for the gradient method:

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

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

♠ **Lower complexity bounds**:

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

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

Optimal gradient methods

- Subgradient method upper and lower bounds

$$f(x^k) - f(x^*) \leq O(1/\sqrt{k})$$
$$f(x^k) - f(x^*) \geq \frac{LD}{2(1+\sqrt{k+1})}.$$

- Composite objective problems: proximal gradient gives same bounds as gradient methods.

Optimal gradient method – rate

Theorem Let $\{x^k\}$ be sequence generated by OptGrad. 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.

Optimal Proximal Gradient

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

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).

Optimal Proximal Gradient

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

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

Optimal Proximal Gradient

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

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

$$\phi(x^k) - \phi^* \leq \frac{2L}{(k+1)^2} \|x^0 - x^*\|_2^2.$$

The operator view

Set-valued mappings

Think of ∂f as a **set-valued map**

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

Set-valued mappings

Think of ∂f as a **set-valued map**

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

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

Set-valued mappings

Think of ∂f as a **set-valued map**

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

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**: $\partial f := \{(x, g) \mid x \in \mathbb{R}^n, g \in \partial f(x)\}$

Set-valued mappings

Think of ∂f as a **set-valued map**

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

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 will write $R(x)$ to mean $\{y \mid (x, y) \in R\}$.
- Example: $\partial f(x) = \{g \mid (x, g) \in \partial f\}$

Why this notation?

- **Goal:** solve *generalized equation* $0 \in R(x)$
- That is, find $x \in \mathbb{R}^n$ such that $(x, 0) \in R$

Why this notation?

- **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

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

  means we want to find an x that minimizes f.
- Helps succinctly write / analyze problems and algorithms

Working with operators

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

Working with operators

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

Working with operators

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

Working with operators

- **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 $\lambda \in \mathbb{R}$

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

Working with operators

- **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 $\lambda \in \mathbb{R}$

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

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

Working with operators

- **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 $\lambda \in \mathbb{R}$

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

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

Which operators are "easier"?

Which operators are "easier"?

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

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

Which operators are "easier"?

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 $\langle Ax - Ay, x - y \rangle \geq 0$

Which operators are "easier"?

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 ∂f of a convex function (verify!)

Which operators are "easier"?

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 ∂f of a convex function (verify!)
- Any monotonically nondecreasing function $T : \mathbb{R} \to \mathbb{R}$

Which operators are "easier"?

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 ∂f of a convex function (verify!)
- Any monotonically nondecreasing function $T : \mathbb{R} \to \mathbb{R}$
- Projection and proximity operators

Which operators are "easier"?

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 ∂f of a convex function (verify!)
- Any monotonically nondecreasing function $T : \mathbb{R} \to \mathbb{R}$
- Projection and proximity operators

Generalize notion of monotonicity to vectors

♠ Abstraction takes linear-algebra intuition to optimization

Importance of resolvent operators

Aim: solve generalized equation

$$0 \in R(x)$$

Importance of resolvent operators

Aim: 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 resolvent operators

Aim: 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:

$0 \in R(x)$

Importance of resolvent operators

Aim: 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:

$0 \in R(x) \leftrightarrow 0 \in \alpha R(x)$

Importance of resolvent operators

Aim: 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:

$$0 \in R(x) \leftrightarrow 0 \in \alpha R(x) \leftrightarrow x \in (I + \alpha R)(x)$$

Importance of resolvent operators

Aim: 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:

$$0 \in R(x) \leftrightarrow 0 \in \alpha R(x) \leftrightarrow x \in (I + \alpha R)(x) \leftrightarrow x = (I + \alpha R)^{-1}(x)$$

Rederiving proximal-gradient

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

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

Rederiving proximal-gradient

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

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

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

Rederiving proximal-gradient

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

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

- Suppose $(I + \lambda \partial h)^{-1}$ is single valued
- Then, $x = (I + \lambda \partial h)^{-1}(y) \implies y \in (I + \lambda \partial h)(x)$

Rederiving proximal-gradient

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

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

- Suppose $(I + \lambda\partial h)^{-1}$ is single valued
- Then, $x = (I + \lambda\partial h)^{-1}(y) \implies y \in (I + \lambda\partial h)(x)$
- That is, $y \in x + \lambda\partial h(x)$

Rederiving proximal-gradient

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

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

- Suppose $(I + \lambda\partial h)^{-1}$ is single valued
- Then, $x = (I + \lambda\partial h)^{-1}(y) \implies y \in (I + \lambda\partial h)(x)$
- That is, $y \in x + \lambda\partial h(x)$
- Equivalently, $x - y + \lambda\partial h(x) \ni 0$

Rederiving proximal-gradient

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

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

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

$$\text{prox}_{\lambda h}(y) \equiv y \mapsto \underset{x}{\text{argmin}}\ \tfrac{1}{2}\|x - y\|_2^2 + \lambda h(x)$$

More proximal splitting

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

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

More proximal splitting

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

- 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:

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

More proximal splitting

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

- 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:

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

- 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"
- Let us 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"
- Let us 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.

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

Proximal splitting

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

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!

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

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!

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

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

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

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

Douglas-Rachford method

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

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

Finally, z is on both sides of the eqn

Douglas-Rachford method

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

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

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

Douglas-Rachford method

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

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

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

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

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

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

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

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

$$z \leftarrow Tz$$
$$T = I + P_f(2P_h - I) - P_h$$

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.

Other methods

- ADMM (DR on dual: **nontrivial theorem**)
- Proximal-Dykstra
- Proximal methods for $f_1 + f_2 + \cdots + f_n$
- Peaceman-Rachford
- Proximal quasi-Newton, Newton
- Ultimately, proximal-point method
- ...

ADMM

Let us see separable objective with constraints

ADMM

Let us see separable objective with constraints

$$\begin{aligned} \min \quad & f(x) + g(z) \\ \text{s.t.} \quad & Ax + Bz = c. \end{aligned}$$

ADMM

Let us see separable objective with constraints

$$\begin{aligned} \min \quad & f(x) + g(z) \\ \text{s.t.} \quad & Ax + Bz = c. \end{aligned}$$

- Objective function separated into x and z variables
- The constraint prevents a trivial decoupling

ADMM

Let us see separable objective with constraints

$$\begin{aligned} \min \quad & f(x) + g(z) \\ \text{s.t.} \quad & Ax + Bz = c. \end{aligned}$$

- Objective function separated into x and z variables
- The constraint prevents a trivial decoupling
- Introduce **augmented lagrangian** (AL)

$$L_\rho(x, z, y) := f(x) + g(z) + y^T(Ax + Bz - c) + \frac{\rho}{2}\|Ax + Bz - c\|_2^2$$

ADMM

Let us see separable objective with constraints

$$\begin{aligned} \min \quad & f(x) + g(z) \\ \text{s.t.} \quad & Ax + Bz = c. \end{aligned}$$

- Objective function separated into x and z variables
- The constraint prevents a trivial decoupling
- Introduce **augmented lagrangian** (AL)

$$L_\rho(x, z, y) := f(x) + g(z) + y^T(Ax + Bz - c) + \frac{\rho}{2}\|Ax + Bz - c\|_2^2$$

- Now, a Gauss-Seidel style update to the AL

ADMM

Let us see separable objective with constraints

$$\begin{aligned} \min \quad & f(x) + g(z) \\ \text{s.t.} \quad & Ax + Bz = c. \end{aligned}$$

- Objective function separated into x and z variables
- The constraint prevents a trivial decoupling
- Introduce **augmented lagrangian** (AL)

$$L_\rho(x, z, y) := f(x) + g(z) + y^T(Ax + Bz - c) + \frac{\rho}{2}\|Ax + Bz - c\|_2^2$$

- Now, a Gauss-Seidel style update to the AL

$$x_{k+1} = \operatorname{argmin}_x L_\rho(x, z_k, y_k)$$

ADMM

Let us see separable objective with constraints

$$\begin{aligned} \min \quad & f(x) + g(z) \\ \text{s.t.} \quad & Ax + Bz = c. \end{aligned}$$

- Objective function separated into x and z variables
- The constraint prevents a trivial decoupling
- Introduce **augmented lagrangian** (AL)

$$L_\rho(x, z, y) := f(x) + g(z) + y^T(Ax + Bz - c) + \frac{\rho}{2}\|Ax + Bz - c\|_2^2$$

- Now, a Gauss-Seidel style update to the AL

$$\begin{aligned} x_{k+1} &= \operatorname{argmin}_x L_\rho(x, z_k, y_k) \\ z_{k+1} &= \operatorname{argmin}_z L_\rho(x_{k+1}, z, y_k) \end{aligned}$$

ADMM

Let us see separable objective with constraints

$$\begin{aligned} \min \quad & f(x) + g(z) \\ \text{s.t.} \quad & Ax + Bz = c. \end{aligned}$$

- Objective function separated into x and z variables
- The constraint prevents a trivial decoupling
- Introduce **augmented lagrangian** (AL)

$$L_\rho(x, z, y) := f(x) + g(z) + y^T(Ax + Bz - c) + \frac{\rho}{2}\|Ax + Bz - c\|_2^2$$

- Now, a Gauss-Seidel style update to the AL

$$\begin{aligned} x_{k+1} &= \operatorname{argmin}_x L_\rho(x, z_k, y_k) \\ z_{k+1} &= \operatorname{argmin}_z L_\rho(x_{k+1}, z, y_k) \\ y_{k+1} &= y_k + \rho(Ax_{k+1} + Bz_{k+1} - c) \end{aligned}$$

ADMM – scaled version

- The AL is

$$L_\rho(x, z, y) := f(x) + g(z) + y^T(Ax + Bz - c) + \frac{\rho}{2}\|Ax + Bz - c\|_2^2$$

ADMM – scaled version

- The AL is

$$L_\rho(x,z,y) := f(x)+g(z)+y^T(Ax+Bz-c)+\tfrac{\rho}{2}\|Ax+Bz-c\|_2^2$$

- Combine linear and quadratic terms in L_ρ, so we have

$$L_\rho(x,z,y) = f(x)+g(z)+\tfrac{\rho}{2}\|Ax+Bz-c+d\|_2^2 + \text{constants}$$

  where we use $d_k = (1/\rho)y_k$ as a new variable.
- **Exercise:** Verify above algebra.

ADMM – scaled version

- The AL is

$$L_\rho(x,z,y) := f(x)+g(z)+y^T(Ax+Bz-c)+\tfrac{\rho}{2}\|Ax+Bz-c\|_2^2$$

- Combine linear and quadratic terms in L_ρ, so we have

$$L_\rho(x,z,y) = f(x)+g(z)+\tfrac{\rho}{2}\|Ax+Bz-c+d\|_2^2 + \text{constants}$$

where we use $d_k = (1/\rho)y_k$ as a new variable.

- **Exercise:** Verify above algebra.

Scaled ADMM

$$x_{k+1} = \operatorname{argmin}_x f(x)+\tfrac{\rho}{2}\|Ax+Bz_k-c+d_k\|_2^2$$

ADMM – scaled version

- The AL is

$$L_\rho(x,z,y) := f(x)+g(z)+y^T(Ax+Bz-c)+\tfrac{\rho}{2}\|Ax+Bz-c\|_2^2$$

- Combine linear and quadratic terms in L_ρ, so we have

$$L_\rho(x,z,y) = f(x)+g(z)+\tfrac{\rho}{2}\|Ax+Bz-c+d\|_2^2 + \text{constants}$$

where we use $d_k = (1/\rho)y_k$ as a new variable.
- **Exercise:** Verify above algebra.

Scaled ADMM

$$\begin{aligned} x_{k+1} &= \operatorname{argmin}_x f(x) + \tfrac{\rho}{2}\|Ax + Bz_k - c + d_k\|_2^2 \\ z_{k+1} &= \operatorname{argmin}_z g(z) + \tfrac{\rho}{2}\|Ax_{k+1} + Bz - c + d_k\|_2^2 \end{aligned}$$

ADMM – scaled version

- The AL is

$$L_\rho(x, z, y) := f(x) + g(z) + y^T(Ax + Bz - c) + \tfrac{\rho}{2}\|Ax + Bz - c\|_2^2$$

- Combine linear and quadratic terms in L_ρ, so we have

$$L_\rho(x, z, y) = f(x) + g(z) + \tfrac{\rho}{2}\|Ax + Bz - c + d\|_2^2 + \text{constants}$$

where we use $d_k = (1/\rho)y_k$ as a new variable.
- **Exercise:** Verify above algebra.

Scaled ADMM

$$\begin{aligned} x_{k+1} &= \operatorname{argmin}_x f(x) + \tfrac{\rho}{2}\|Ax + Bz_k - c + d_k\|_2^2 \\ z_{k+1} &= \operatorname{argmin}_z g(z) + \tfrac{\rho}{2}\|Ax_{k+1} + Bz - c + d_k\|_2^2 \\ d_{k+1} &= d_k + (Ax_{k+1} + Bz_{k+1} - c) \end{aligned}$$

ADMM – convergence

Theorem Say f, g are convex, and L_0 (ordinary Lagrangian) has a saddle-point. Then, ADMM converges, and *feasible iterates* $Ax_k + Bz_k - c \to 0$. Also, objective function approaches (primal) optimal value: $f(x_k) + g(z_k) \to f(x^*) + g(z^*)$

ADMM – convergence

Theorem Say f, g are convex, and L_0 (ordinary Lagrangian) has a saddle-point. Then, ADMM converges, and *feasible iterates* $Ax_k + Bz_k - c \to 0$. Also, objective function approaches (primal) optimal value: $f(x_k) + g(z_k) \to f(x^*) + g(z^*)$

Selecting ρ is still an art!

ADMM – constrained optimization

$$\min \quad f(x) \quad \text{s.t.} \quad x \in \mathcal{X}.$$

ADMM – constrained optimization

$$\min \quad f(x) \quad \text{s.t.} \quad x \in \mathcal{X}.$$

ADMM form

$$\min \quad f(x) + \mathbb{1}_{\mathcal{X}}(z)$$

$$\text{s.t.}\ x - z = 0.$$

ADMM – constrained optimization

$$\min \quad f(x) \quad \text{s.t.} \quad x \in \mathcal{X}.$$

ADMM form

$$\min \quad f(x) + \mathbb{1}_{\mathcal{X}}(z)$$
$$\text{s.t. } x - z = 0.$$

ADMM iterations (scaled)

$$\begin{aligned} x_{k+1} &= \operatorname{argmin} f(x) + \tfrac{\rho}{2}\|x - z_k + d_k\|_2^2 \\ z_{k+1} &= P_{\mathcal{X}}(x_{k+1} + d_k) \\ d_{k+1} &= d_k + (x_{k+1} - z_{k+1}) \end{aligned}$$

Notice x update is proximity operator of $f(x)$; z update is proximity operator of $\mathbb{1}_{\mathcal{X}}$.