Advanced Optimization

(10-801: CMU)

Lecture 18

Proximal methods, Monotone operators

24 Mar, 2014

Suvrit Sra

Proximal Gradient

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

Projected gradient

$$x \leftarrow \Pi(x - \alpha \nabla f(x))$$

Π denotes **orthogonal** projection onto $\mathcal{X}$.

Proximal Gradient

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

Projected gradient

$$x \leftarrow \Pi(x - \alpha \nabla f(x))$$

Π denotes **orthogonal** projection onto $\mathcal{X}$.

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

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

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

Projected gradient

$$x \leftarrow \Pi(x - \alpha \nabla f(x))$$

Π denotes **orthogonal** projection onto $\mathcal{X}$.

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

Proximal gradient

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

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

NOTE: non-orthogonal, non-Euclidean versions also exist

Where does it come from?

Shorthand: $P \equiv \text{prox}$.

Lemma If x^* is optimal, then $x^* = P_{\alpha h}(x^* - \alpha \nabla f(x^*))$, $\forall \alpha > 0$

Where does it come from?

Shorthand: $P \equiv \mathrm{prox}$.

Lemma If x^* is optimal, then $x^* = P_{\alpha h}(x^* - \alpha \nabla f(x^*))$, $\forall \alpha > 0$

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

Where does it come from?

Shorthand: $P \equiv \text{prox}$.

Lemma If x^* is optimal, then $x^* = P_{\alpha h}(x^* - \alpha \nabla f(x^*))$, $\forall \alpha > 0$

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

Where does it come from?

Shorthand: $P \equiv \text{prox}$.

Lemma If x^* is optimal, then $x^* = P_{\alpha h}(x^* - \alpha \nabla f(x^*))$, $\forall \alpha > 0$

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

Where does it come from?

Shorthand: $P \equiv \text{prox}$.

Lemma If x^* is optimal, then $x^* = P_{\alpha h}(x^* - \alpha \nabla f(x^*))$, $\forall \alpha > 0$

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

Where does it come from?

Shorthand: $P \equiv \mathrm{prox}$.

Lemma If x^* is optimal, then $x^* = P_{\alpha h}(x^* - \alpha \nabla f(x^*))$, $\forall \alpha > 0$

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

Where does it come from?

Shorthand: $P \equiv \text{prox}$.

Lemma If x^* is optimal, then $x^* = P_{\alpha h}(x^* - \alpha \nabla f(x^*))$, $\forall \alpha > 0$

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

Where does it come from?

Shorthand: $P \equiv \text{prox}$.

Lemma If x^* is optimal, then $x^* = P_{\alpha h}(x^* - \alpha \nabla f(x^*))$, $\forall \alpha > 0$

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

Above fixed-point eqn suggests iteration

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

Why does it work?

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?

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

Gradient mapping: the "gradient-like object"

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

Why does 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)

Convergence analysis

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

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

Convergence analysis

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

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

- ♣ Gradient vectors of closeby points are close to each other
- ♣ Objective function has "bounded curvature"
- ♣ Speed at which gradient varies is bounded

Convergence analysis

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

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

- ♣ Gradient vectors of closeby points are close to each other
- ♣ Objective function has "bounded curvature"
- ♣ Speed at which gradient varies is bounded

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

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

Convergence analysis

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

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

- ♣ Gradient vectors of closeby points are close to each other
- ♣ Objective function has "bounded curvature"
- ♣ Speed at which gradient varies is bounded

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

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

For convex f, compare with

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle.$$

Descent lemma

Proof. Since $f \in C_L^1$, by Taylor's theorem, for the vector $z_t = x + t(y - x)$ we have

$$f(y) = f(x) + \int_0^1 \langle \nabla f(z_t), \, y - x \rangle dt.$$

Descent lemma

Proof. Since $f \in C_L^1$, by Taylor's theorem, for the vector $z_t = x + t(y - x)$ we have

$$f(y) = f(x) + \int_0^1 \langle \nabla f(z_t),\, y - x \rangle dt.$$

Add and subtract $\langle \nabla f(x),\, y - x \rangle$ on rhs we have

$$f(y) - f(x) - \langle \nabla f(x),\, y - x \rangle \quad = \quad \int_0^1 \langle \nabla f(z_t) - \nabla f(x),\, y - x \rangle dt$$

Descent lemma

Proof. Since $f \in C_L^1$, by Taylor's theorem, for the vector $z_t = x + t(y-x)$ we have

$$f(y) = f(x) + \int_0^1 \langle \nabla f(z_t),\, y-x \rangle dt.$$

Add and subtract $\langle \nabla f(x),\, y-x \rangle$ on rhs we have

$$\begin{aligned} f(y) - f(x) - \langle \nabla f(x),\, y-x \rangle &= \int_0^1 \langle \nabla f(z_t) - \nabla f(x),\, y-x \rangle dt \\ |f(y) - f(x) - \langle \nabla f(x),\, y-x \rangle| &= \left| \int_0^1 \langle \nabla f(z_t) - \nabla f(x),\, y-x \rangle dt \right| \end{aligned}$$

Descent lemma

Proof. Since $f \in C_L^1$, by Taylor's theorem, for the vector $z_t = x + t(y - x)$ we have

$$f(y) = f(x) + \int_0^1 \langle \nabla f(z_t), y - x \rangle dt.$$

Add and subtract $\langle \nabla f(x), y - x \rangle$ on rhs we have

$$\begin{aligned} f(y) - f(x) - \langle \nabla f(x), y - x \rangle &= \int_0^1 \langle \nabla f(z_t) - \nabla f(x), y - x \rangle dt \\ |f(y) - f(x) - \langle \nabla f(x), y - x \rangle| &= \left| \int_0^1 \langle \nabla f(z_t) - \nabla f(x), y - x \rangle dt \right| \\ &\leq \int_0^1 |\langle \nabla f(z_t) - \nabla f(x), y - x \rangle| dt \end{aligned}$$

Descent lemma

Proof. Since $f \in C_L^1$, by Taylor's theorem, for the vector $z_t = x + t(y - x)$ we have

$$f(y) = f(x) + \int_0^1 \langle \nabla f(z_t), y - x\rangle dt.$$

Add and subtract $\langle \nabla f(x), y - x\rangle$ on rhs we have

$$\begin{aligned} f(y) - f(x) - \langle \nabla f(x), y - x\rangle &= \int_0^1 \langle \nabla f(z_t) - \nabla f(x), y - x\rangle dt \\ |f(y) - f(x) - \langle \nabla f(x), y - x\rangle| &= \left|\int_0^1 \langle \nabla f(z_t) - \nabla f(x), y - x\rangle dt\right| \\ &\leq \int_0^1 |\langle \nabla f(z_t) - \nabla f(x), y - x\rangle| dt \\ &\leq \int_0^1 \|\nabla f(z_t) - \nabla f(x)\|_2 \cdot \|y - x\|_2 dt \end{aligned}$$

Descent lemma

Proof. Since $f \in C_L^1$, by Taylor's theorem, for the vector $z_t = x + t(y-x)$ we have

$$f(y) = f(x) + \int_0^1 \langle \nabla f(z_t),\, y - x\rangle dt.$$

Add and subtract $\langle \nabla f(x),\, y-x\rangle$ on rhs we have

$$\begin{aligned}
f(y) - f(x) - \langle \nabla f(x),\, y-x\rangle &= \int_0^1 \langle \nabla f(z_t) - \nabla f(x),\, y-x\rangle dt \\
|f(y) - f(x) - \langle \nabla f(x),\, y-x\rangle| &= \left|\int_0^1 \langle \nabla f(z_t) - \nabla f(x),\, y-x\rangle dt\right| \\
&\leq \int_0^1 |\langle \nabla f(z_t) - \nabla f(x),\, y-x\rangle| dt \\
&\leq \int_0^1 \|\nabla f(z_t) - \nabla f(x)\|_2 \cdot \|y-x\|_2 dt \\
&\leq L \int_0^1 t\|x-y\|_2^2 dt
\end{aligned}$$

Descent lemma

Proof. Since $f \in C_L^1$, by Taylor's theorem, for the vector $z_t = x + t(y - x)$ we have

$$f(y) = f(x) + \int_0^1 \langle \nabla f(z_t), y - x \rangle dt.$$

Add and subtract $\langle \nabla f(x), y - x \rangle$ on rhs we have

$$\begin{aligned}
f(y) - f(x) - \langle \nabla f(x), y - x \rangle &= \int_0^1 \langle \nabla f(z_t) - \nabla f(x), y - x \rangle dt \\
|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| &= \left| \int_0^1 \langle \nabla f(z_t) - \nabla f(x), y - x \rangle dt \right| \\
&\leq \int_0^1 |\langle \nabla f(z_t) - \nabla f(x), y - x \rangle| dt \\
&\leq \int_0^1 \|\nabla f(z_t) - \nabla f(x)\|_2 \cdot \|y - x\|_2 dt \\
&\leq L \int_0^1 t \|x - y\|_2^2 dt \\
&= \frac{L}{2} \|x - y\|_2^2.
\end{aligned}$$

Bounds $f(y)$ around x with quadratic functions

Descent lemma – corollary

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

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

Descent lemma – corollary

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

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

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

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

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

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

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

Descent lemma – corollary

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

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

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

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

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

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

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

Exer: Prove! (use convexity of f, h, and $G_\alpha(x) - \nabla f(x) \in \partial h(y)$)

Convergence analysis

We've actually shown that $x' = x - \alpha G_\alpha(x)$ 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 why this inequality suffices to show convergence.

Convergence analysis

We've actually shown that $x' = x - \alpha G_\alpha(x)$ 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 why this inequality suffices to show convergence.
Plug in $z = x^*$ in corollary 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 $x' = x - \alpha G_\alpha(x)$ 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 why this inequality suffices to show convergence. Plug in $z = x^*$ in corollary 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 $x' = x - \alpha G_\alpha(x)$ 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 why this inequality suffices to show convergence. Plug in $z = x^*$ in corollary 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

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

Convergence rate

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

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

Convergence rate

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

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

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.

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.
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 $k \geq 1$:

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

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

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

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

Simplified proof in lecture notes.

Monotone operators

Why is proximity called an "operator"?

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

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

Why is proximity called an “operator”?

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

Why is proximity called an "operator"?

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

Why is proximity called an "operator"?

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, $y \in x + \lambda \partial h(x)$

Why is proximity called an "operator"?

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, $x - y + \lambda \partial h(x) \ni 0$

Why is proximity called an "operator"?

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, $x - y + \lambda\partial h(x) \ni 0$
- Nothing other than optimality condition for prox-operator

$$\mathrm{prox}_{\lambda h}(y) \equiv y \mapsto \operatorname*{argmin}_x \ \tfrac{1}{2}\|x - y\|_2^2 + \lambda h(x)$$

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 (recall firm nonexpansivity)

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 (recall firm nonexpansivity)

Generalize notion of monotonicity to vectors

♠ Abstraction helps take our linear-algebra intuition to optimization

Monotone operators – simple facts

Exercise: Prove λR monotone if R monotone and $\lambda \geq 0$
Exercise: Prove R^{-1} monotone, if R is monotone
Exercise: For monotone R, S and $\lambda \geq 0$, $R + \lambda S$ is monotone.

Monotone operators – simple facts

Exercise: Prove λR monotone if R monotone and $\lambda \geq 0$
Exercise: Prove R^{-1} monotone, if R is monotone
Exercise: For monotone R, S and $\lambda \geq 0$, $R + \lambda S$ is monotone.
Corollary: Resolvent operator of monotone operator is monotone.

$$R \text{ monotone} \quad \Longrightarrow (I + \lambda R)^{-1} \text{ is monotone.}$$

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

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

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

Rederiving proximal-gradient

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

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

Rederiving proximal-gradient

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

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

Rederiving proximal-gradient

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

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

Rederiving proximal-gradient

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

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

Rederiving proximal-gradient

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

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

Resolvent of subdifferential is prox operator

Proximal splitting methods

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

- 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

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

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

Example:

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

Proximal splitting methods

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

- 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"
- 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!

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

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!

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

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

- Not a fixed-point equation yet

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

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

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

Finally, z is on both sides of the eqn

Douglas-Rachford method

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

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

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

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

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

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

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

Dropping superscripts, writing $P \equiv \operatorname{prox}$, 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

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

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.

Best approximation problem

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

Best approximation problem

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

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 which 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.

Example

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.

Example

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 $R_{d_A} = \Pi_A$ (quite interesting!)

Example

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 $R_{d_A} = \Pi_A$ (quite interesting!)
Thus, DR for solving above problem becomes

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

Example

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 $R_{d_A} = \Pi_A$ (quite interesting!)
Thus, DR for solving above problem becomes

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

Exercise:* Convergence rate of above method?

References

- ♠ *DTU 2010 slides, Laurent El Ghaoui*
- ♠ *EE227A slides, S. Sra*
- ♠ *Introductory Lectures on Convex Optimization, Yu. Nesterov*
- ♠ *EE364B notes, Stephen Boyd*