Optimization for Machine Learning

(Problems; Algorithms - C)

Suvrit Sra

Massachusetts Institute of Technology

PKU Summer School on Data Science (July 2017)

Course materials

- *http://suvrit.de/teaching.html*
- Some references:
  - *Introductory lectures on convex optimization* – Nesterov
  - *Convex optimization* – Boyd & Vandenberghe
  - *Nonlinear programming* – Bertsekas
  - *Convex Analysis* – Rockafellar
  - *Fundamentals of convex analysis* – Urruty, Lemaréchal
  - *Lectures on modern convex optimization* – Nemirovski
  - *Optimization for Machine Learning* – Sra, Nowozin, Wright
  - *Theory of Convex Optimization for Machine Learning* – Bubeck
  - *NIPS 2016 Optimization Tutorial* – Bach, Sra
- Some related courses:
  - EE227A, Spring 2013, (Sra, UC Berkeley)
  - 10-801, Spring 2014 (Sra, CMU)
  - EE364a,b (Boyd, Stanford)
  - EE236b,c (Vandenberghe, UCLA)
- Venues: NIPS, ICML, UAI, AISTATS, SIOPT, Math. Prog.

Lecture Plan

- Introduction (3 lectures)
- Problems and algorithms (5 lectures)
- Non-convex optimization, perspectives (2 lectures)

Nonsmooth convergence rates

Theorem. (Nesterov.) Let $\mathcal{B} = \{x \mid \|x - x^0\|_2 \le D\}$. Assume, $x^* \in \mathcal{B}$. There exists a convex function f in $C_L^0(\mathcal{B})$ (with $L > 0$), such that for $0 \le k \le n-1$, the lower-bound

$$f(x^k) - f(x^*) \ge \frac{LD}{2(1+\sqrt{k+1})},$$

holds for **any algorithm** that generates x^k by linearly combining the previous iterates and subgradients.

Exercise: So design problems where we can do better!

Composite problems

Composite objectives

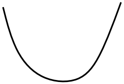
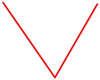
Frequently ML problems take the regularized form

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

Composite objectives

Frequently ML problems take the regularized form

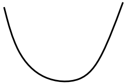
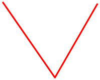
$$\text{minimize}\; f(x) := \ell(x) + r(x)$$

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

Composite objectives

Frequently ML problems take the regularized form

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

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

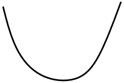
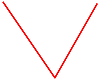
Example: $\ell(x) = \frac{1}{2}\|Ax - b\|^2$ and $r(x) = \lambda\|x\|_1$

Lasso, L1-LS, compressed sensing

Composite objectives

Frequently ML problems take the regularized 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

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

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

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

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 (sub)gradient

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

Proximal Gradient Method

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

Projected (sub)gradient

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

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

Proximal gradient

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

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

Proximal Gradient Method

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

Projected (sub)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 proximity operator for h

Why? If we can compute $\text{prox}_h(x)$ easily, prox-grad converges as fast gradient methods for smooth problems!

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

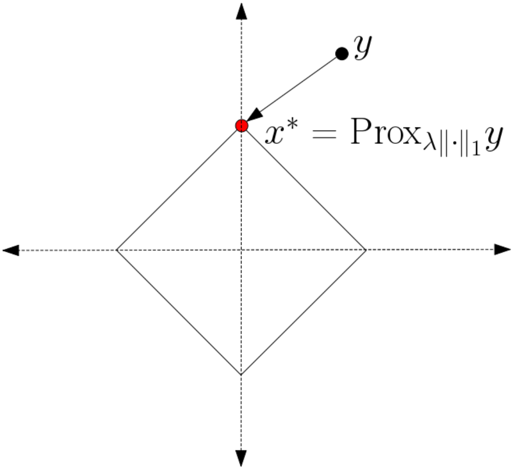
Projection

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

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

Exercise: 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 $x \neq 0$

Exercise: **Moreau decomposition** $y = \text{prox}_h y + \text{prox}_{h^*} y$ (notice analogy to $V = S + S^\perp$ in linear algebra)

How to cook-up prox-grad?

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

How to cook-up prox-grad?

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

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

How to cook-up prox-grad?

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

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

How to cook-up prox-grad?

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

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

How to cook-up prox-grad?

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

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

How to cook-up prox-grad?

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

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

How to cook-up prox-grad?

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

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

How to cook-up prox-grad?

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

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

Above fixed-point eqn suggests iteration

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

Convergence*

Proximal-gradient works, why?

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

Proximal-gradient works, why?

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

Gradient mapping: the "gradient-like object"

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

Proximal-gradient works, why?

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

Gradient mapping: the "gradient-like object"

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

- 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 $f \in C^1_L$, 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 $f \in C^1_L$, 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 $f \in C^1_L$, 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 $f \in C^1_L$, 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 $f \in C^1_L$, 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

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

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

Descent lemma – corollary

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

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

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

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

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

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

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

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

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

Exercise: Prove! (hint: f, h are convex, $G_\alpha(x) - \nabla f(x) \in \partial h(y)$)

Convergence analysis

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

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

Exercise: Why this inequality suffices to show convergence.

Convergence analysis

We've actually shown $x' = x - \alpha G_\alpha(x)$ 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: Why this inequality suffices to show convergence.
Use $z = x^*$ 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 $x' = x - \alpha G_\alpha(x)$ 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: Why this inequality suffices to show convergence.
Use $z = x^*$ in corollary to obtain progress in terms of iterates:

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

Convergence analysis

We've actually shown $x' = x - \alpha G_\alpha(x)$ 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: Why this inequality suffices to show convergence.
Use $z = x^*$ in corollary to obtain progress in terms of iterates:

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

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

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

Convergence rate

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

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

Convergence rate

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

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

Convergence rate

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

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

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

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

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

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

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

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

Proximal splitting methods

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

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

Proximal splitting methods

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

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

Example:

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

Proximal splitting methods

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

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

Example:

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

- 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!

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

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

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!

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

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

- 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!

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

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

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

Douglas-Rachford splitting

Reflection operator

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

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

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

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

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

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

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

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.

Proximal methods – cornucopia

- Douglas Rachford splitting
- ADMM (special case of DR on dual)
- Proximal-Dykstra
- Proximal methods for $f_1 + f_2 + \cdots + f_n$
- Peaceman-Rachford
- Proximal quasi-Newton, Newton
- Many other variation...

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.

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

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

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

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