Introduction to large-scale optimization
(Lecture 2)

Suvrit Sra
Massachusetts Institute of Technology

Microsoft Research India
Machine Learning Summer School, June 2015

Course materials

- http://suvrit.de/teach/msr2015/
- 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
- Some related courses:
  - EE227A, Spring 2013, (UC Berkeley)
  - 10-801, Spring 2014 (CMU)
  - EE364a,b (Boyd, Stanford)
  - EE236b,c (Vandenberghe, UCLA)
- NIPS, ICML, UAI, AISTATS, SIOPT, Math. Prog.

Outline

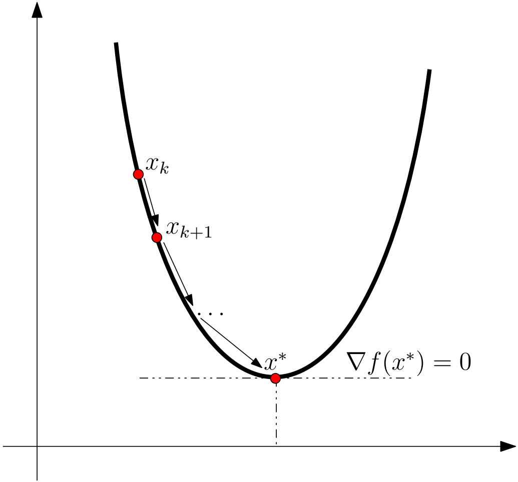
- Recap on convexity
- Recap on duality, optimality
- First-order optimization algorithms
- Proximal methods, operator splitting
- Incremental methods
- High-level view of parallel, distributed
- Some words on nonconvex

Descent methods

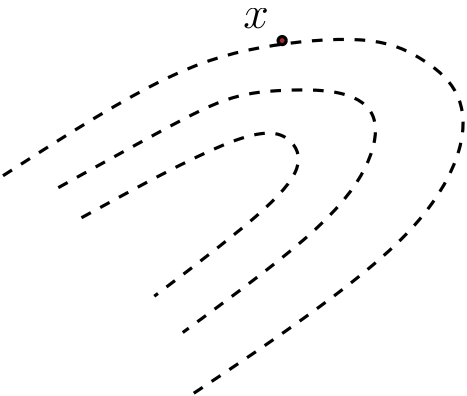
$$\boxed{\min_x \quad f(x)}$$

Descent methods

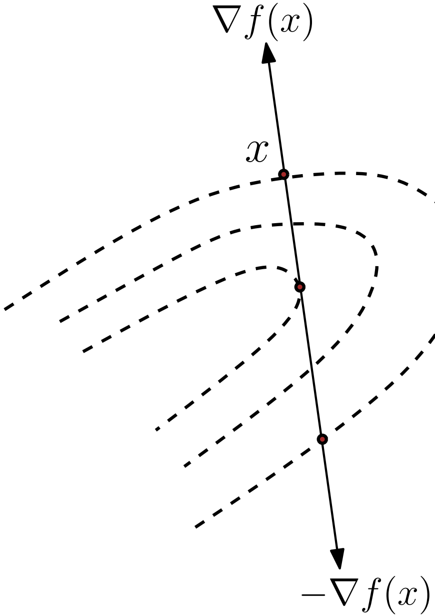
$$\min_x \quad f(x)$$



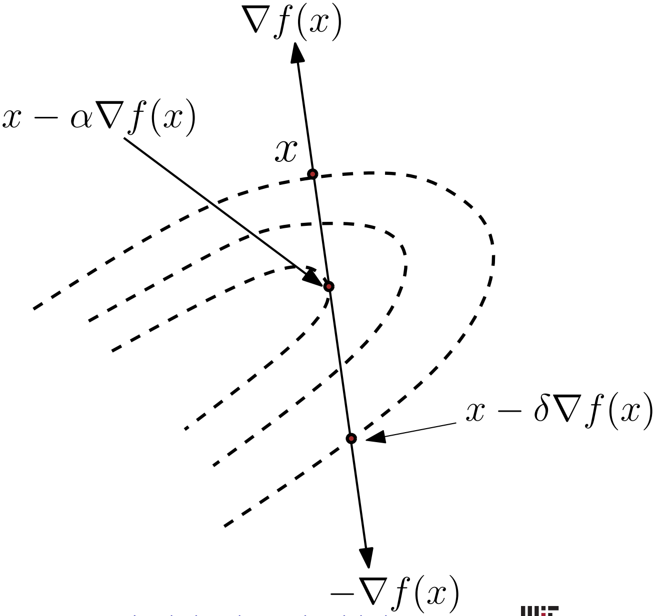
Descent methods



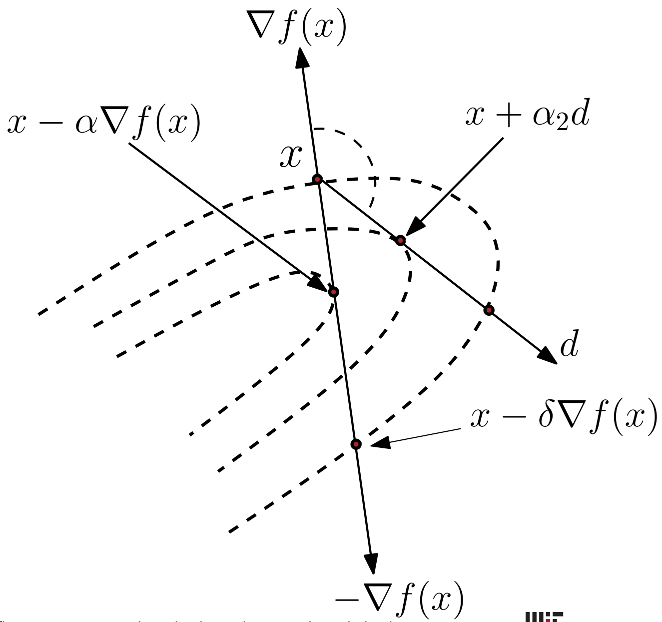
Descent methods



Descent methods



Descent methods



Algorithm

1. Start with some guess x^0;
2. For each $k = 0, 1, \ldots$
   - $x^{k+1} \leftarrow x^k + \alpha_k d^k$
   - Check when to stop (e.g., if $\nabla f(x^{k+1}) = 0$)

Gradient methods

$$\boxed{x^{k+1} = x^k + \alpha_k d^k, \quad k = 0, 1, \ldots}$$

- **stepsize** $\alpha_k \geq 0$, usually ensures $f(x^{k+1}) < f(x^k)$

Gradient methods

$$\boxed{x^{k+1} = x^k + \alpha_k d^k, \quad k = 0, 1, \ldots}$$

- **stepsize** $\alpha_k \geq 0$, usually ensures $f(x^{k+1}) < f(x^k)$
- **Descent direction** d^k satisfies

$$\langle \nabla f(x^k), d^k \rangle < 0$$

Gradient methods

$$\boxed{x^{k+1} = x^k + \alpha_k d^k, \quad k = 0, 1, \ldots}$$

- **stepsize** $\alpha_k \geq 0$, usually ensures $f(x^{k+1}) < f(x^k)$
- **Descent direction** d^k satisfies

$$\langle \nabla f(x^k), d^k \rangle < 0$$

Numerous ways to select α_k and d^k

Gradient methods

$$\boxed{x^{k+1} = x^k + \alpha_k d^k, \quad k = 0, 1, \ldots}$$

- **stepsize** $\alpha_k \geq 0$, usually ensures $f(x^{k+1}) < f(x^k)$
- **Descent direction** d^k satisfies

$$\langle \nabla f(x^k), d^k \rangle < 0$$

Numerous ways to select α_k and d^k

Usually methods **seek monotonic descent**

$$f(x^{k+1}) < f(x^k)$$

Gradient methods – direction

$$\boxed{x^{k+1} = x^k + \alpha_k d^k, \quad k = 0, 1, \ldots}$$

- Different choices of direction d^k
  - **Scaled gradient:** $d^k = -D^k \nabla f(x^k)$, $D^k \succ 0$
  - **Newton's method:** $(D^k = [\nabla^2 f(x^k)]^{-1})$
  - **Quasi-Newton:** $D^k \approx [\nabla^2 f(x^k)]^{-1}$
  - **Steepest descent:** $D^k = \boldsymbol{I}$
  - **Diagonally scaled:** D^k diagonal with $D^k_{ii} \approx \left(\frac{\partial^2 f(x^k)}{(\partial x_i)^2}\right)^{-1}$
  - **Discretized Newton:** $D^k = [H(x^k)]^{-1}$, H via finite-diff.

Gradient methods – direction

$$x^{k+1} = x^k + \alpha_k d^k, \quad k = 0, 1, \ldots$$

- Different choices of direction d^k
  - **Scaled gradient:** $d^k = -D^k \nabla f(x^k)$, $D^k \succ 0$
  - **Newton's method:** $(D^k = [\nabla^2 f(x^k)]^{-1})$
  - **Quasi-Newton:** $D^k \approx [\nabla^2 f(x^k)]^{-1}$
  - **Steepest descent:** $D^k = \boldsymbol{I}$
  - **Diagonally scaled:** D^k diagonal with $D^k_{ii} \approx \left(\frac{\partial^2 f(x^k)}{(\partial x_i)^2}\right)^{-1}$
  - **Discretized Newton:** $D^k = [H(x^k)]^{-1}$, H via finite-diff.
  - …

Exercise: Verify that $\langle \nabla f(x^k), d^k \rangle < 0$ for above choices

Gradient methods – stepsize

- **Exact:** $\alpha_k := \underset{\alpha \geq 0}{\operatorname{argmin}}\, f(x^k + \alpha d^k)$

Gradient methods – stepsize

- **Exact:** $\alpha_k := \underset{\alpha \geq 0}{\operatorname{argmin}}\, f(x^k + \alpha d^k)$
- **Limited min:** $\alpha_k = \underset{0 \leq \alpha \leq s}{\operatorname{argmin}}\, f(x^k + \alpha d^k)$

Gradient methods – stepsize

- **Exact:** $\alpha_k := \underset{\alpha \geq 0}{\operatorname{argmin}}\, f(x^k + \alpha d^k)$
- **Limited min:** $\alpha_k = \underset{0 \leq \alpha \leq s}{\operatorname{argmin}}\, f(x^k + \alpha d^k)$
- **Armijo-rule**. Given **fixed** scalars, s, β, σ with $0 < \beta < 1$ and $0 < \sigma < 1$ (chosen experimentally). Set

$$\alpha_k = \beta^{m_k} s,$$

where we **try** $\beta^m s$ for $m = 0, 1, \ldots$ until **sufficient descent**

$$f(x^k) - f(x + \beta^m s d^k) \geq -\sigma \beta^m s \langle \nabla f(x^k), d^k \rangle$$

- **Constant:** $\alpha_k = 1/L$ (for suitable value of L)
- **Diminishing:** $\alpha_k \to 0$ but $\sum_k \alpha_k = \infty$.

Convergence

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

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

Convergence

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

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,

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

Theorem Let $f \in C_L^1$ and $\{x^k\}$ be sequence generated as above, with $\alpha_k = 1/L$. Then, $f(x^{k+1}) - f(x^*) = O(1/k)$.

Linear convergence

Assumption: **Strong convexity**; denote $f \in S^1_{L,\mu}$

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

- Setting $\alpha_k = 2/(\mu + L)$ yields linear rate ($\mu > 0$)

Strongly convex – linear rate

Theorem. If $f \in S^1_{L,\mu}$, $0 < \alpha < 2/(L+\mu)$, then the gradient method generates a sequence $\{x^k\}$ that satisfies

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

Moreover, if $\alpha = 2/(L+\mu)$ then

$$f(x^k) - f^* \leq \frac{L}{2}\left(\frac{\kappa - 1}{\kappa + 1}\right)^{2k} \|x^0 - x^*\|_2^2,$$

where $\kappa = L/\mu$ is the **condition number**.

Gradient methods – lower bounds

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

Theorem Lower bound I (Nesterov) For any $x^0 \in \mathbb{R}^n$, and $1 \leq k \leq \frac{1}{2}(n-1)$, there is a smooth f, s.t.

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

Gradient methods – lower bounds

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

Theorem Lower bound I (Nesterov) For any $x^0 \in \mathbb{R}^n$, and $1 \leq k \leq \frac{1}{2}(n-1)$, there is a smooth f, s.t.

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

Theorem Lower bound II (Nesterov). For class of smooth, strongly convex, i.e., $S^\infty_{L,\mu}$ ($\mu > 0$, $\kappa > 1$)

$$f(x^k) - f(x^*) \geq \frac{\mu}{2}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2k} \|x^0 - x^*\|_2^2.$$

Nonsmooth opt.

Subgradient method

$$\boxed{\begin{array}{c} x^{k+1} = x^k - \alpha_k g^k \\ \text{where } g^k \in \partial f(x^k) \text{ is } \textbf{any} \text{ subgradient} \end{array}}$$

Subgradient method

$$x^{k+1} = x^k - \alpha_k g^k$$

where $g^k \in \partial f(x^k)$ is **any** subgradient

Stepsize $\alpha_k > 0$ must be chosen

Subgradient method

$$x^{k+1} = x^k - \alpha_k g^k$$

where $g^k \in \partial f(x^k)$ is **any** subgradient

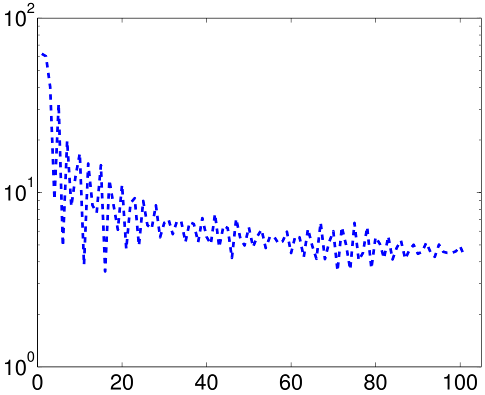
Stepsize $\alpha_k > 0$ must be chosen

- Method generates sequence $\{x^k\}_{k \geq 0}$
- Does this sequence converge to an optimal solution x^*?
- If yes, then how fast?
- What if have constraints: $x \in \mathcal{X}$?

Example

$$\min \quad \tfrac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1$$

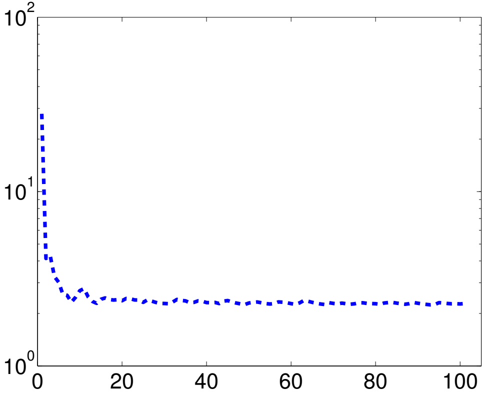
$$x^{k+1} = x^k - \alpha_k(A^T(Ax^k - b) + \lambda\,\mathrm{sgn}(x^k))$$



Example

$$\min \quad \tfrac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1$$

$$x^{k+1} = x^k - \alpha_k(A^T(Ax^k - b) + \lambda\,\mathrm{sgn}(x^k))$$



(More careful implementation)

Subgradient method – stepsizes

- **Constant** Set $\alpha_k = \alpha > 0$, for $k \geq 0$
- **Scaled constant** $\alpha_k = \alpha / \|g^k\|_2 \quad (\|x^{k+1} - x^k\|_2 = \alpha)$

Subgradient method – stepsizes

- **Constant** Set $\alpha_k = \alpha > 0$, for $k \geq 0$
- **Scaled constant** $\alpha_k = \alpha/\|g^k\|_2$ $(\|x^{k+1} - x^k\|_2 = \alpha)$
- **Square summable but not summable**

$$\sum_k \alpha_k^2 < \infty, \qquad \sum_k \alpha_k = \infty$$

- **Diminishing scalar**

$$\lim_k \alpha_k = 0, \qquad \sum_k \alpha_k = \infty$$

- **Adaptive stepsizes** (not covered)

Not a descent method!
Work with best f^k so far: $f^k_{\min} := \min_{0 \leq i \leq k} f^i$

Exercise

Support vector machines

- Let $\mathcal{D} := \{(x_i, y_i) \mid x_i \in \mathbb{R}^n, y_i \in \{\pm 1\}\}$
- We wish to find $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$ such that

$$\min_{w,b} \quad \tfrac{1}{2}\|w\|_2^2 + C\sum\nolimits_{i=1}^{m} \max[0, 1 - y_i(w^T x_i + b)]$$

- Derive and implement a subgradient method
- Plot evolution of objective function
- Experiment with different values of $C > 0$
- Plot and keep track of $f_{\min}^k := \min_{0 \le t \le k} f(x^t)$

Nonsmooth complexity

- Let $\phi(x) = |x|$ for $x \in \mathbb{R}$

Nonsmooth complexity

- Let $\phi(x) = |x|$ for $x \in \mathbb{R}$
- Subgradient method $x^{k+1} = x^k - \alpha_k g^k$, where $g^k \in \partial|x^k|$.

Nonsmooth complexity

- Let $\phi(x) = |x|$ for $x \in \mathbb{R}$
- Subgradient method $x^{k+1} = x^k - \alpha_k g^k$, where $g^k \in \partial|x^k|$.
- If $x^0 = 1$ and $\alpha_k = \frac{1}{\sqrt{k+1}} + \frac{1}{\sqrt{k+2}}$ (this stepsize is known to be optimal), then $|x^k| = \frac{1}{\sqrt{k+1}}$

Nonsmooth complexity

- Let $\phi(x) = |x|$ for $x \in \mathbb{R}$
- Subgradient method $x^{k+1} = x^k - \alpha_k g^k$, where $g^k \in \partial|x^k|$.
- If $x^0 = 1$ and $\alpha_k = \frac{1}{\sqrt{k+1}} + \frac{1}{\sqrt{k+2}}$ (this stepsize is known to be optimal), then $|x^k| = \frac{1}{\sqrt{k+1}}$
- Thus, $O(\frac{1}{\epsilon^2})$ iterations are needed to obtain ϵ-accuracy.

Nonsmooth complexity

- Let $\phi(x) = |x|$ for $x \in \mathbb{R}$
- Subgradient method $x^{k+1} = x^k - \alpha_k g^k$, where $g^k \in \partial|x^k|$.
- If $x^0 = 1$ and $\alpha_k = \frac{1}{\sqrt{k+1}} + \frac{1}{\sqrt{k+2}}$ (this stepsize is known to be optimal), then $|x^k| = \frac{1}{\sqrt{k+1}}$
- Thus, $O(\frac{1}{\epsilon^2})$ iterations are needed to obtain ϵ-accuracy.
- This behavior typical for the subgradient method which exhibits $O(1/\sqrt{k})$ convergence in general

Nonsmooth complexity

- Let $\phi(x) = |x|$ for $x \in \mathbb{R}$
- Subgradient method $x^{k+1} = x^k - \alpha_k g^k$, where $g^k \in \partial|x^k|$.
- If $x^0 = 1$ and $\alpha_k = \frac{1}{\sqrt{k+1}} + \frac{1}{\sqrt{k+2}}$ (this stepsize is known to be optimal), then $|x^k| = \frac{1}{\sqrt{k+1}}$
- Thus, $O(\frac{1}{\epsilon^2})$ iterations are needed to obtain ϵ-accuracy.
- This behavior typical for the subgradient method which exhibits $O(1/\sqrt{k})$ convergence in general

$$\boxed{\text{Can we do better in general?}}$$

Nonsmooth complexity

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

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

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

Constrained problems

Constrained optimization

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

Constrained optimization

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

- Previously: $x^{t+1} = x^t - \alpha_t g^t$
- This could be infeasible!

Projected subgradient method

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

where $g^k \in \partial f(x^k)$ is any subgradient

- **Projection:** closest feasible point

$$P_{\mathcal{X}}(y) = \operatorname*{argmin}_{x \in \mathcal{X}} \|x - y\|^2$$

Projected subgradient method

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

where $g^k \in \partial f(x^k)$ is any subgradient

- **Projection:** closest feasible point

$$P_{\mathcal{X}}(y) = \operatorname*{argmin}_{x \in \mathcal{X}} \|x - y\|^2$$

- Great as long as projection is "easy"
- Same questions as before:
  - Does it converge?
  - For which stepsizes?
  - How fast?

Examples

$$\begin{aligned} \min \quad & \tfrac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1 \\ \text{s.t. } & x \in \mathcal{X} \end{aligned}$$

- **Nonnegativity** $x \geq 0$
  $P_{\mathcal{X}}(z) = [z]_+$
  Update step: $x^{k+1} = [x^k - \alpha_k(A^T(Ax^k - b) + \lambda\,\text{sgn}(x^k))]_+$

Examples

$$\begin{aligned} \min \quad & \tfrac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1 \\ \text{s.t. } & x \in \mathcal{X} \end{aligned}$$

- **Nonnegativity** $x \geq 0$
  $P_{\mathcal{X}}(z) = [z]_+$
  Update step: $x^{k+1} = [x^k - \alpha_k(A^T(Ax^k - b) + \lambda\,\mathrm{sgn}(x^k))]_+$
- ℓ_∞**-ball** $\|x\|_\infty \leq 1$
  Projection: $\min \|x - z\|^2$ s.t. $x \leq 1$ and $x \geq -1$

Examples

$$\begin{aligned} \min \quad & \tfrac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1 \\ \text{s.t. } & x \in \mathcal{X} \end{aligned}$$

- **Nonnegativity** $x \geq 0$
  $P_{\mathcal{X}}(z) = [z]_+$
  Update step: $x^{k+1} = [x^k - \alpha_k(A^T(Ax^k - b) + \lambda \operatorname{sgn}(x^k))]_+$
- ℓ_∞**-ball** $\|x\|_\infty \leq 1$
  Projection: $\min \|x - z\|^2$ s.t. $x \leq 1$ and $x \geq -1$
  this is separable, so do it coordinate-wise:
  $P_{\mathcal{X}}(z) = y$ where $y_i = \mathrm{sgn}(z_i) \min\{|z_i|, 1\}$

Examples

$$\begin{aligned} \min \quad & \tfrac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1 \\ \text{s.t. } & x \in \mathcal{X} \end{aligned}$$

- **Nonnegativity** $x \geq 0$
  $P_{\mathcal{X}}(z) = [z]_+$
  Update step: $x^{k+1} = [x^k - \alpha_k(A^T(Ax^k - b) + \lambda\,\mathrm{sgn}(x^k))]_+$
- ℓ_∞**-ball** $\|x\|_\infty \leq 1$
  Projection: $\min \|x - z\|^2$ s.t. $x \leq 1$ and $x \geq -1$
  this is separable, so do it coordinate-wise:
  $P_{\mathcal{X}}(z) = y$ where $y_i = \mathrm{sgn}(z_i)\min\{|z_i|, 1\}$
  Update step:

$$\begin{aligned} z^{k+1} &= x^k - \alpha_k(A^T(Ax^k - b) + \lambda\,\mathrm{sgn}(x^k)) \\ x_i^{k+1} &= \mathrm{sgn}(z_i^{k+1})\min\{|z_i^{k+1}|, 1\} \end{aligned}$$

Examples

- **Linear constraints** $Ax = b$ ($A \in \mathbb{R}^{n \times m}$ has rank n)

$$\begin{aligned} P_{\mathcal{X}}(y) &= y - A^\top (AA^\top)^{-1}(Ay - b) \\ &= (I - A^\top (A^\top A)^{-1} A)y + A^\top (AA^\top)^{-1} b \end{aligned}$$

Examples

- **Linear constraints** $Ax = b$ ($A \in \mathbb{R}^{n\times m}$ has rank n)

$$\begin{aligned} P_{\mathcal{X}}(y) &= y - A^\top(AA^\top)^{-1}(Ay - b) \\ &= (I - A^\top(A^\top A)^{-1}A)y + A^\top(AA^\top)^{-1}b \end{aligned}$$

Update step, using $Ax^t = b$:

$$\begin{aligned} x^{t+1} &= P_{\mathcal{X}}(x^t - \alpha_t g^t) \\ &= x^t - \alpha_t(I - A^\top(AA^\top)^{-1}A)g^t \end{aligned}$$

Examples

- **Linear constraints** $Ax = b$ ($A \in \mathbb{R}^{n\times m}$ has rank n)

$$\begin{aligned} P_{\mathcal{X}}(y) &= y - A^\top(AA^\top)^{-1}(Ay - b) \\ &= (I - A^\top(A^\top A)^{-1}A)y + A^\top(AA^\top)^{-1}b \end{aligned}$$

  Update step, using $Ax^t = b$:

$$\begin{aligned} x^{t+1} &= P_{\mathcal{X}}(x^t - \alpha_t g^t) \\ &= x^t - \alpha_t(I - A^\top(AA^\top)^{-1}A)g^t \end{aligned}$$

- **Simplex** $x^\top 1 = 1$ and $x \geq 0$
  more complex but doable in $O(n)$, similarly ℓ_1-norm ball

Subgradient method – remarks

- Why care?
  - simple
  - low-memory
  - large-scale versions possible

Subgradient method – remarks

- Why care?
  - simple
  - low-memory
  - large-scale versions possible
- Another perspective

$$x^{k+1} = \min_{x \in \mathcal{X}} \langle x, g^k \rangle + \frac{1}{2\alpha_k} \|x - x_k\|^2$$

Mirror Descent

Subgradient method – remarks

- Why care?
  - simple
  - low-memory
  - large-scale versions possible
- Another perspective

$$x^{k+1} = \min_{x \in \mathcal{X}} \langle x, g^k \rangle + \frac{1}{2\alpha_k} \|x - x_k\|^2$$

Mirror Descent

- Improvements using more information (heavy-ball, filtered subgradient, ...)

Subgradient method – remarks

- Why care?
  - simple
  - low-memory
  - large-scale versions possible
- Another perspective

$$x^{k+1} = \min_{x \in \mathcal{X}} \langle x, g^k \rangle + \frac{1}{2\alpha_k} \|x - x_k\|^2$$

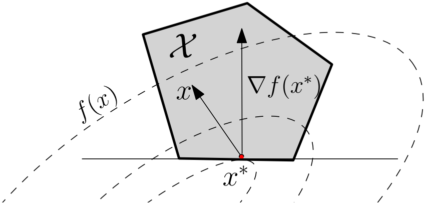
  Mirror Descent
- Improvements using more information (heavy-ball, filtered subgradient, ...)
- Don't forget the dual
  - may be more amenable to optimization
  - duality gap

What we did not cover

- Adaptive stepsize tricks
- Space dilation methods, quasi-Newton style subgrads
- Barrier subgradient method
- Sparse subgradient method
- Ellipsoid method, center of gravity, etc. as subgradient methods
- ...

Feasible descent

$$\min \quad f(x) \quad \text{s.t. } x \in \mathcal{X}$$
$$\langle \nabla f(x^*), x - x^* \rangle \geq 0, \qquad \forall x \in \mathcal{X}.$$



Feasible descent

$$\boxed{x^{k+1} = x^k + \alpha_k d^k}$$

Feasible descent

$$\boxed{x^{k+1} = x^k + \alpha_k d^k}$$

- d^k – **feasible direction**, i.e., $x^k + \alpha_k d^k \in \mathcal{X}$

Feasible descent

$$\boxed{x^{k+1} = x^k + \alpha_k d^k}$$

- d^k – **feasible direction**, i.e., $x^k + \alpha_k d^k \in \mathcal{X}$
- d^k must also be **descent direction**, i.e., $\langle \nabla f(x^k), d^k \rangle < 0$
- Stepsize α_k chosen to ensure **feasibility and descent**.

Feasible descent

$$\boxed{x^{k+1} = x^k + \alpha_k d^k}$$

- d^k – **feasible direction**, i.e., $x^k + \alpha_k d^k \in \mathcal{X}$
- d^k must also be **descent direction**, i.e., $\langle \nabla f(x^k), d^k \rangle < 0$
- Stepsize α_k chosen to ensure **feasibility and descent**.

Since $\mathcal{X}$ is convex, all feasible directions are of the form

$$d^k = \gamma(z - x^k), \quad \gamma > 0,$$

where $z \in \mathcal{X}$ is any feasible vector.

Feasible descent

$$\boxed{x^{k+1} = x^k + \alpha_k d^k}$$

- d^k – **feasible direction**, i.e., $x^k + \alpha_k d^k \in \mathcal{X}$
- d^k must also be **descent direction**, i.e., $\langle \nabla f(x^k), d^k \rangle < 0$
- Stepsize α_k chosen to ensure **feasibility and descent**.

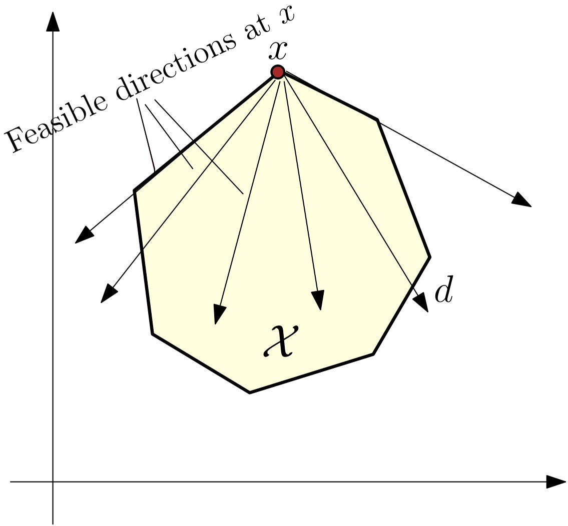
Since $\mathcal{X}$ is convex, all feasible directions are of the form

$$d^k = \gamma(z - x^k), \quad \gamma > 0,$$

where $z \in \mathcal{X}$ is any feasible vector.

$$\boxed{x^{k+1} = x^k + \alpha_k(z^k - x^k), \quad \alpha_k \in (0, 1]}$$

Cone of feasible directions



Conditional gradient method

Optimality: $\langle \nabla f(x^k), z^k - x^k \rangle \geq 0$ for all $z^k \in \mathcal{X}$

Conditional gradient method

Optimality: $\langle \nabla f(x^k), z^k - x^k \rangle \geq 0$ for all $z^k \in \mathcal{X}$
Aim: If not optimal, then generate feasible direction
$d^k = z^k - x^k$ that obeys **descent condition** $\langle \nabla f(x^k), d^k \rangle < 0$.

Conditional gradient method

Optimality: $\langle \nabla f(x^k), z^k - x^k \rangle \geq 0$ for all $z^k \in \mathcal{X}$
Aim: If not optimal, then generate feasible direction $d^k = z^k - x^k$ that obeys **descent condition** $\langle \nabla f(x^k), d^k \rangle < 0$.

Frank-Wolfe (Conditional gradient) method

- Let $z^k \in \operatorname{argmin}_{x \in \mathcal{X}} \langle \nabla f(x^k), x - x^k \rangle$
- Use different methods to select α_k
- $x^{k+1} = x^k + \alpha_k (z^k - x^k)$

Conditional gradient method

Optimality: $\langle \nabla f(x^k), z^k - x^k \rangle \geq 0$ for all $z^k \in \mathcal{X}$
Aim: If not optimal, then generate feasible direction $d^k = z^k - x^k$ that obeys **descent condition** $\langle \nabla f(x^k), d^k \rangle < 0$.

Frank-Wolfe (Conditional gradient) method

- ▲ Let $z^k \in \operatorname{argmin}_{x \in \mathcal{X}} \langle \nabla f(x^k), x - x^k \rangle$
- ▲ Use different methods to select α_k
- ▲ $x^{k+1} = x^k + \alpha_k (z^k - x^k)$

- ♠ Practical when solving *linear* problem over $\mathcal{X}$ easy
- ♠ Became popular in machine learning in recent years
- ♠ Refinements, several variants

Composite objectives

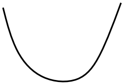
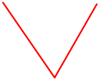
Frequently nonsmooth problems take the form

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

Composite objectives

Frequently nonsmooth problems take the form

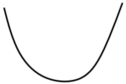
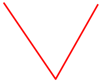
$$\boxed{\text{minimize } f(x) := \ell(x) + r(x)}$$

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

Composite objectives

Frequently nonsmooth problems take the form

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

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

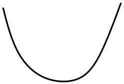
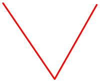
Example: $\ell(x) = \frac{1}{2}\|Ax - b\|^2$ and $r(x) = \lambda\|x\|_1$

Lasso, L1-LS, compressed sensing

Composite objectives

Frequently nonsmooth problems take the form

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

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

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

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

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

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

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

Projected gradient

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

Proximal Gradient Method

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

Projected gradient

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

$$\boxed{\min \quad f(x) + h(x)}$$

Proximal gradient

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

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

Proximal Gradient Method

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

Projected gradient

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

$$\boxed{\min \quad f(x) + h(x)}$$

Proximal gradient

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

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

NOTE: non-Euclidean versions (mirror-descent) also exist

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

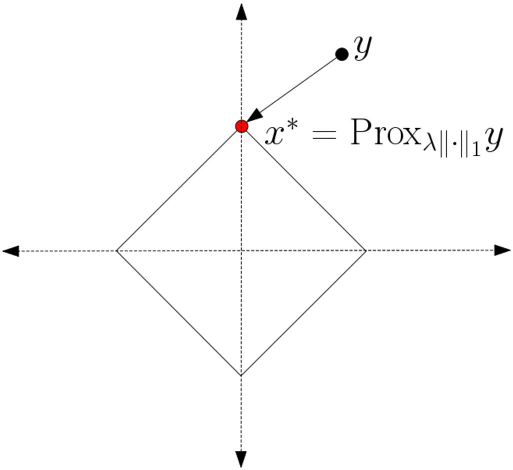
Projection

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

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

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

Proximity operator



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

Proximity operators

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

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

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

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

Hint 2: Consider the two cases: 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 $x^* = \text{prox}_{\alpha h}(x^* - \alpha \nabla f(x^*)), \forall \alpha > 0$

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

Where does it come from?

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

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

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?

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

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?

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

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,

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

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,

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

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

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

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

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

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

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

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

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

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

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

Exercise: Why this inequality suffices to show convergence.
Use $z = x^*$ in corollary to obtain progress in terms of iterates:

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

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

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

Convergence rate

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

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

Convergence rate

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

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

Convergence rate

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

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

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

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

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 $O(1/k^2)$

Faster methods

Optimal gradient methods

♠ We saw following efficiency estimates for the gradient method

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

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

Optimal gradient methods

- We saw following efficiency estimates for the gradient method

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

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

- We also saw **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.

Gradient with "momentum"

Polyak's method (aka heavy-ball) for $f \in S^1_{L,\mu}$

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

Gradient with "momentum"

Polyak's method (aka heavy-ball) for $f \in S^1_{L,\mu}$

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

- **Converges** (locally, i.e., for $\|x^0 - x^*\|_2 \leq \epsilon$) as

$$\|x^k - x^*\|_2^2 \leq \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}\right)^{2k} \|x^0 - x^*\|_2^2,$$

for $\alpha_k = \frac{4}{(\sqrt{L}+\sqrt{\mu})^2}$ and $\beta_k = \left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^2$

Nesterov's optimal gradient method

$$\boxed{\min_x f(x), \text{ where } S^1_{L,\mu} \text{ with } \mu \geq 0}$$

Nesterov's optimal gradient method

$\min_x f(x)$, where $S^1_{L,\mu}$ with $\mu \geq 0$

1. Choose $x^0 \in \mathbb{R}^n$, $\alpha_0 \in (0, 1)$
2. Let $y^0 \leftarrow x^0$; set $q = \mu/L$

Nesterov's optimal gradient method

$$\min_x f(x), \text{ where } S^1_{L,\mu} \text{ with } \mu \geq 0$$

1. Choose $x^0 \in \mathbb{R}^n$, $\alpha_0 \in (0,1)$
2. Let $y^0 \leftarrow x^0$; set $q = \mu/L$
3. k-th iteration ($k \geq 0$):
   a). Compute intermediate update

$$x^{k+1} = y^k - \tfrac{1}{L}\nabla f(y^k)$$

Nesterov's optimal gradient method

$\min_x f(x)$, where $S^1_{L,\mu}$ with $\mu \geq 0$

1. Choose $x^0 \in \mathbb{R}^n$, $\alpha_0 \in (0, 1)$
2. Let $y^0 \leftarrow x^0$; set $q = \mu/L$
3. k-th iteration ($k \geq 0$):
   a). Compute intermediate update

$$x^{k+1} = y^k - \tfrac{1}{L}\nabla f(y^k)$$

   b). Compute stepsize α_{k+1} by solving

$$\alpha_{k+1}^2 = (1 - \alpha_{k+1})\alpha_k^2 + q\alpha_{k+1}$$

Nesterov's optimal gradient method

$\min_x f(x)$, where $S^1_{L,\mu}$ with $\mu \geq 0$

1. Choose $x^0 \in \mathbb{R}^n$, $\alpha_0 \in (0, 1)$
2. Let $y^0 \leftarrow x^0$; set $q = \mu/L$
3. k-th iteration ($k \geq 0$):
   a). Compute intermediate update
   $$x^{k+1} = y^k - \tfrac{1}{L}\nabla f(y^k)$$
   b). Compute stepsize α_{k+1} by solving
   $$\alpha_{k+1}^2 = (1 - \alpha_{k+1})\alpha_k^2 + q\alpha_{k+1}$$
   c). Set $\beta_k = \alpha_k(1 - \alpha_k)/(\alpha_k^2 + \alpha_{k+1})$
   d). Update solution estimate
   $$y^{k+1} = x^{k+1} + \beta_k(x^{k+1} - x^k)$$

Optimal gradient method – rate

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

Strongly convex case – simplification

If $\mu > 0$, select $\alpha_0 = \sqrt{\mu/L}$. The two main steps get simplified:

1. Set $\beta_k = \alpha_k(1 - \alpha_k)/(\alpha_k^2 + \alpha_{k+1})$
2. $y^{k+1} = x^{k+1} + \beta_k(x^{k+1} - x^k)$

$$\alpha_k = \sqrt{\tfrac{\mu}{L}} \qquad \beta_k = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}, \qquad k \geq 0.$$

Strongly convex case – simplification

If $\mu > 0$, select $\alpha_0 = \sqrt{\mu/L}$. The two main steps get simplified:

1. Set $\beta_k = \alpha_k(1-\alpha_k)/(\alpha_k^2 + \alpha_{k+1})$
2. $y^{k+1} = x^{k+1} + \beta_k(x^{k+1} - x^k)$

$$\alpha_k = \sqrt{\frac{\mu}{L}} \qquad \beta_k = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}, \qquad k \geq 0.$$

Optimal method simplifies to

1. Choose $y^0 = x^0 \in \mathbb{R}^n$
2. k-th iteration ($k \geq 0$):

Strongly convex case – simplification

If $\mu > 0$, select $\alpha_0 = \sqrt{\mu/L}$. The two main steps get simplified:

1. Set $\beta_k = \alpha_k(1-\alpha_k)/(\alpha_k^2 + \alpha_{k+1})$
2. $y^{k+1} = x^{k+1} + \beta_k(x^{k+1} - x^k)$

$$\alpha_k = \sqrt{\frac{\mu}{L}} \qquad \beta_k = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}, \qquad k \geq 0.$$

Optimal method simplifies to

1. Choose $y^0 = x^0 \in \mathbb{R}^n$
2. k-th iteration ($k \geq 0$):
   a). $x^{k+1} = y^k - \frac{1}{L}\nabla f(y^k)$
   b). $y^{k+1} = x^{k+1} + \beta(x^{k+1} - x^k)$

Strongly convex case – simplification

If $\mu > 0$, select $\alpha_0 = \sqrt{\mu/L}$. The two main steps get simplified:

1. Set $\beta_k = \alpha_k(1-\alpha_k)/(\alpha_k^2 + \alpha_{k+1})$
2. $y^{k+1} = x^{k+1} + \beta_k(x^{k+1} - x^k)$

$$\alpha_k = \sqrt{\tfrac{\mu}{L}} \qquad \beta_k = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}, \qquad k \geq 0.$$

Optimal method simplifies to

1. Choose $y^0 = x^0 \in \mathbb{R}^n$
2. k-th iteration ($k \geq 0$):
   a). $x^{k+1} = y^k - \frac{1}{L}\nabla f(y^k)$
   b). $y^{k+1} = x^{k+1} + \beta(x^{k+1} - x^k)$

$$\boxed{\text{Notice similarity to Polyak's method!}}$$

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).

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

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

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

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

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 of monotone operator is monotone.

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

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

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

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

- 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

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

- 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

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

- 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

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

More proximal splitting

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

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

More proximal splitting

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

- 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:

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

More proximal splitting

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

- 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!

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

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!

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

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

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!

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

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

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

Douglas-Rachford splitting

Reflection operator

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

Douglas-Rachford splitting

Reflection operator

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

Douglas-Rachford method

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

Douglas-Rachford splitting

Reflection operator

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

Douglas-Rachford method

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

Douglas-Rachford splitting

Reflection operator

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

Douglas-Rachford method

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

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

Douglas-Rachford splitting

Reflection operator

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

Douglas-Rachford method

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

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

Douglas-Rachford splitting

Reflection operator

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

Douglas-Rachford method

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

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

Douglas-Rachford splitting

Reflection operator

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

Douglas-Rachford method

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

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

Douglas-Rachford splitting

Reflection operator

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

Douglas-Rachford method

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

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

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

Douglas-Rachford method

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

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

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

Douglas-Rachford method

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

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

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

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

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

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

Exercise: Prove this claim.

Other methods

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