An introduction to convex methods for life science
Unconstrained minimization for nonsmooth convex problems

Math et Sciences du Vivant – Université Paris-Saclay / Paris-Sud

Autumn semester 2017

http://julien.cremeriefamily.info

References

See Chapter 9 in

Convex Optimization,
Stephen Boyd and Lieve Lieven Vandenberghe
https://web.stanford.edu/~boyd/cvxbook/

All slides stolen (extracted/re-arranged) from Lieve Vandenberghe, Ryan Tibshirani:

- Optimization Methods for Large-Scale Systems
  http://www.seas.ucla.edu/~vandenbe/ee236c/ee236c.html
- Convex Optimization:
  http://www.stat.cmu.edu/~ryantibs/convexopt/

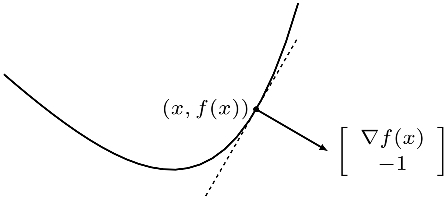
Outline

Outline

Basic inequality

recall the basic inequality for differentiable convex functions:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \quad \forall y \in \operatorname{dom} f$$



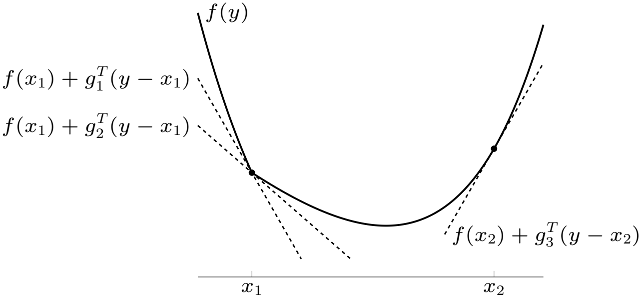
- the first-order approximation of f at x is a global lower bound
- $\nabla f(x)$ defines a non-vertical supporting hyperplane to $\mathbf{epi}\, f$ at $(x, f(x))$:

$$\begin{bmatrix} \nabla f(x) \\ -1 \end{bmatrix}^T \left(\begin{bmatrix} y \\ t \end{bmatrix} - \begin{bmatrix} x \\ f(x) \end{bmatrix} \right) \leq 0 \quad \forall (y, t) \in \mathbf{epi}\, f$$

Subgradient

g is a **subgradient** of a convex function f at $x \in \mathrm{dom}\, f$ if

$$f(y) \geq f(x) + g^T(y - x) \quad \forall y \in \mathrm{dom}\, f$$



g_1, g_2 are subgradients at x_1; g_3 is a subgradient at x_2

Subdifferential

the **subdifferential** $\partial f(x)$ of f at x is the set of all subgradients:

$$\partial f(x) = \{g \mid g^T(y-x) \le f(y) - f(x),\ \forall y \in \mathrm{dom}\, f\}$$

Properties

- $\partial f(x)$ is a closed convex set (possibly empty)

  this follows from the definition: $\partial f(x)$ is an intersection of halfspaces

- if $x \in \mathbf{int}\,\mathrm{dom}\, f$ then $\partial f(x)$ is nonempty and bounded

  proof on next two pages

Proof: we show that $\partial f(x)$ is nonempty when $x \in \mathbf{int}\,\mathrm{dom}\, f$

- $(x, f(x))$ is in the boundary of the convex set $\mathbf{epi}\, f$
- therefore there exists a supporting hyperplane to $\mathbf{epi}\, f$ at $(x, f(x))$:

$$\exists (a, b) \neq 0, \qquad \left[\begin{array}{c} a \\ b \end{array}\right]^T \left(\left[\begin{array}{c} y \\ t \end{array}\right] - \left[\begin{array}{c} x \\ f(x) \end{array}\right] \right) \leq 0 \qquad \forall (y, t) \in \mathbf{epi}\, f$$

- $b > 0$ gives a contradiction as $t \to \infty$
- $b = 0$ gives a contradiction for $y = x + \epsilon a$ with small $\epsilon > 0$
- therefore $b < 0$ and $g = \dfrac{1}{|b|} a$ is a subgradient of f at x

Proof: $\partial f(x)$ is bounded when $x \in \mathbf{int}\,\mathrm{dom}\,f$

- for small $r > 0$, define a set of $2n$ points

$$B = \{x \pm re_k \mid k = 1, \ldots, n\} \subset \mathrm{dom}\,f$$

  and define $M = \max_{y \in B} f(y) < \infty$

- for every nonzero $g \in \partial f(x)$, there is a point $y \in B$ with

$$f(y) \geq f(x) + g^T(y - x) = f(x) + r\|g\|_\infty$$

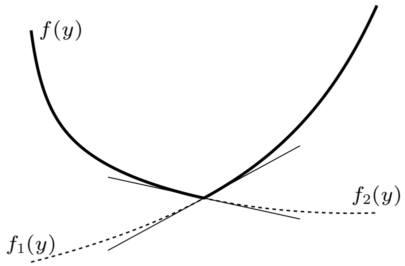
  (choose an index k with $|g_k| = \|g\|_\infty$, and take $y = x + r\,\mathbf{sign}(g_k)e_k$)

- therefore $\partial f(x)$ is bounded:

$$\sup_{g \in \partial f(x)} \|g\|_\infty \leq \frac{M - f(x)}{r}$$

Example

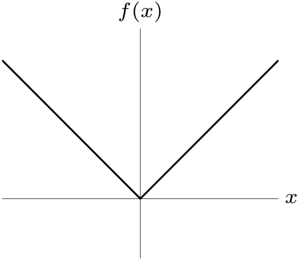
$$f(x) = \max\{f_1(x), f_2(x)\} \qquad \text{with } f_1,\ f_2 \text{ convex and differentiable}$$

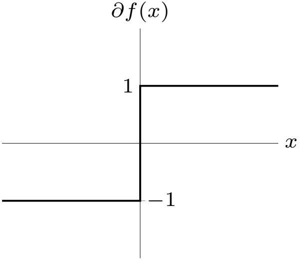


- if $f_1(\hat{x}) = f_2(\hat{x})$, subdifferential at $\hat{x}$ is line segment $[\nabla f_1(\hat{x}), \nabla f_2(\hat{x})]$
- if $f_1(\hat{x}) > f_2(\hat{x})$, subdifferential at $\hat{x}$ is $\{\nabla f_1(\hat{x})\}$
- if $f_1(\hat{x}) < f_2(\hat{x})$, subdifferential at $\hat{x}$ is $\{\nabla f_2(\hat{x})\}$

Examples

Absolute value $f(x) = |x|$

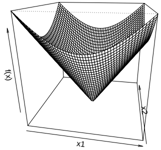




Euclidean norm $f(x) = \|x\|_2$

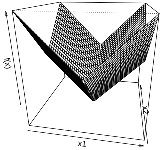
$$\partial f(x) = \{\frac{1}{\|x\|_2}x\} \quad \text{if } x \neq 0, \qquad \partial f(x) = \{g \mid \|g\|_2 \leq 1\} \quad \text{if } x = 0$$

Consider $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = \|x\|_2$



- For $x \neq 0$, unique subgradient $g = x/\|x\|_2$
- For $x = 0$, subgradient g is any element of $\{z : \|z\|_2 \leq 1\}$

Consider $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = \|x\|_1$



- For $x_i \neq 0$, unique ith component $g_i = \mathrm{sign}(x_i)$
- For $x_i = 0$, ith component g_i is any element of $[-1, 1]$

Outline

Connection to convex geometry

Convex set $C \subseteq \mathbb{R}^n$, consider indicator function $I_C : \mathbb{R}^n \to \mathbb{R}$,

$$I_C(x) = I\{x \in C\} = \begin{cases} 0 & \text{if } x \in C \\ \infty & \text{if } x \notin C \end{cases}$$

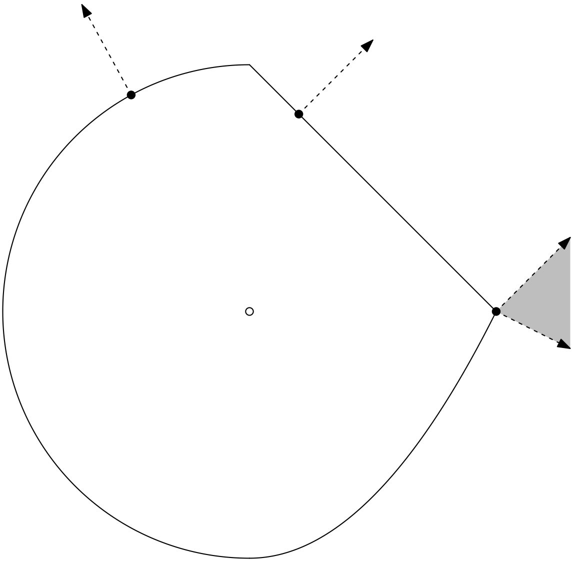
For $x \in C$, $\partial I_C(x) = \mathcal{N}_C(x)$, the normal cone of C at x, recall

$$\mathcal{N}_C(x) = \{g \in \mathbb{R}^n : g^T x \geq g^T y \text{ for any } y \in C\}$$

Why? By definition of subgradient g,

$$I_C(y) \geq I_C(x) + g^T(y - x) \quad \text{for all } y$$

- For $y \notin C$, $I_C(y) = \infty$
- For $y \in C$, this means $0 \geq g^T(y - x)$



Subgradient calculus

Basic rules for convex functions:

- Scaling: $\partial(af) = a \cdot \partial f$ provided $a > 0$
- Addition: $\partial(f_1 + f_2) = \partial f_1 + \partial f_2$
- Affine composition: if $g(x) = f(Ax + b)$, then

$$\partial g(x) = A^T \partial f(Ax + b)$$

- Finite pointwise maximum: if $f(x) = \max_{i=1,\ldots m} f_i(x)$, then

$$\partial f(x) = \operatorname{conv}\left(\bigcup_{i: f_i(x) = f(x)} \partial f_i(x) \right)$$

  convex hull of union of subdifferentials of all active functions at x

- General pointwise maximum: if $f(x) = \max_{s \in S} f_s(x)$, then

$$\partial f(x) \supseteq \mathrm{cl}\left\{\mathrm{conv}\left(\bigcup_{s: f_s(x) = f(x)} \partial f_s(x)\right)\right\}$$

and under some regularity conditions (on S, f_s), we get an equality above

- Norms: important special case, $f(x) = \|x\|_p$. Let q be such that $1/p + 1/q = 1$, then

$$\|x\|_p = \max_{\|z\|_q \leq 1} \; z^T x$$

Hence

$$\partial f(x) = \underset{\|z\|_q \leq 1}{\mathrm{argmax}} \; z^T x$$

Why subgradients?

Subgradients are important for two reasons:

- Convex analysis: optimality characterization via subgradients, monotonicity, relationship to duality
- Convex optimization: if you can compute subgradients, then you can minimize (almost) any convex function

Optimality condition

For any f (convex or not),

$$f(x^\star) = \min_x \ f(x) \quad \Longleftrightarrow \quad 0 \in \partial f(x^\star)$$

I.e., $x^\star$ is a minimizer if and only if 0 is a subgradient of f at $x^\star$. This is called the subgradient optimality condition

Why? Easy: $g = 0$ being a subgradient means that for all y

$$f(y) \geq f(x^\star) + 0^T(y - x^\star) = f(x^\star)$$

Note the implication for a convex and differentiable function f, with $\partial f(x) = \{\nabla f(x)\}$

Derivation of first-order optimality

Example of the power of subgradients: we can use what we have learned so far to derive the first-order optimality condition. Recall that for f convex and differentiable, the problem

$$\min_x \; f(x) \;\; \text{subject to} \;\; x \in C$$

is solved at x if and only if

$$\nabla f(x)^T (y - x) \geq 0 \quad \text{for all } \; y \in C$$

Intuitively says that gradient increases as we move away from x. How to see this? First recast problem as

$$\min_x \; f(x) + I_C(x)$$

Now apply subgradient optimality: $0 \in \partial(f(x) + I_C(x))$

But

$$
\begin{aligned}
0 \in \partial\big(f(x) + I_C(x)\big)& \\
&\iff 0 \in \{\nabla f(x)\} + \mathcal{N}_C(x) \\
&\iff -\nabla f(x) \in \mathcal{N}_C(x) \\
&\iff -\nabla f(x)^T x \geq -\nabla f(x)^T y \text{ for all } \in C \\
&\iff \nabla f(x)^T (y - x) \geq 0 \text{ for all } y \in C
\end{aligned}
$$

as desired

Note: the condition $0 \in \partial f(x) + \mathcal{N}_C(x)$ is a fully general condition for optimality in a convex problem. But this is not always easy to work with (KKT conditions, later, are easier)

Outline

Subgradients and subdifferentials

- Definitions
- Important Properties
- Example: ℓ_1-regularization aka Lasso

Subgradients methods

Proximal methods

Example: lasso optimality conditions

Given $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, lasso problem can be parametrized as:

$$\min_{\beta} \; \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

where $\lambda \geq 0$. Subgradient optimality:

$$\begin{aligned} 0 \in \partial\Big(\frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1\Big) \\ &\iff \; 0 \in -X^T(y - X\beta) + \lambda\partial\|\beta\|_1 \\ &\iff \; X^T(y - X\beta) = \lambda v \end{aligned}$$

for some $v \in \partial\|\beta\|_1$, i.e.,

$$v_i \in \begin{cases} \{1\} & \text{if } \beta_i > 0 \\ \{-1\} & \text{if } \beta_i < 0 \\ [-1, 1] & \text{if } \beta_i = 0 \end{cases}, \quad i = 1, \ldots p$$

Write $X_1, \ldots X_p$ for columns of X. Then subgradient optimality reads:

$$\begin{cases} X_i^T(y - X\beta) = \lambda \cdot \mathrm{sign}(\beta_i) & \text{if } \beta_i \neq 0 \\ |X_i^T(y - X\beta)| \leq \lambda & \text{if } \beta_i = 0 \end{cases}$$

Note: the subgradient optimality conditions do not directly lead to an expression for a lasso solution ... however they do provide a way to check lasso optimality

They are also helpful in understanding the lasso estimator; e.g., if $|X_i^T(y - X\beta)| < \lambda$, then $\beta_i = 0$

Example: soft-thresholding

Simplfied lasso problem with $X = I$:

$$\min_{\beta} \ \frac{1}{2}\|y - \beta\|_2^2 + \lambda\|\beta\|_1$$

This we can solve directly using subgradient optimality. Solution is $\beta = S_\lambda(y)$, where S_λ is the soft-thresholding operator:

$$[S_\lambda(y)]_i = \begin{cases} y_i - \lambda & \text{if } y_i > \lambda \\ 0 & \text{if } -\lambda \le y_i \le \lambda \\ y_i + \lambda & \text{if } y_i < -\lambda \end{cases}, \quad i = 1, \dots n$$

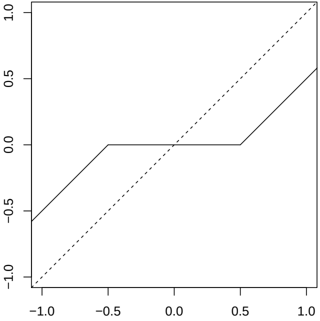
Check: from last slide, subgradient optimality conditions are

$$\begin{cases} y_i - \beta_i = \lambda \cdot \mathrm{sign}(\beta_i) & \text{if } \beta_i \neq 0 \\ |y_i - \beta_i| \le \lambda & \text{if } \beta_i = 0 \end{cases}$$

Now plug in $\beta = S_\lambda(y)$ and check these are satisfied:

- When $y_i > \lambda$, $\beta_i = y_i - \lambda > 0$, so $y_i - \beta_i = \lambda = \lambda \cdot 1$
- When $y_i < -\lambda$, argument is similar
- When $|y_i| \leq \lambda$, $\beta_i = 0$, and $|y_i - \beta_i| = |y_i| \leq \lambda$

Soft-thresholding in one variable:



Outline

Outline

Subgradient method

to minimize a nondifferentiable convex function f: choose $x^{(0)}$ and repeat

$$x^{(k)} = x^{(k-1)} - t_k g^{(k-1)}, \quad k = 1, 2, \ldots$$

$g^{(k-1)}$ is any subgradient of f at $x^{(k-1)}$

Step size rules

- fixed step: t_k constant
- fixed length: $t_k \|g^{(k-1)}\|_2 = \|x^{(k)} - x^{(k-1)}\|_2$ is constant
- diminishing: $t_k \to 0$, $\sum\limits_{k=1}^{\infty} t_k = \infty$

Assumptions

- f has finite optimal value $f^\star$, minimizer $x^\star$
- f is convex, $\mathrm{dom}\, f = \mathbf{R}^n$
- f is Lipschitz continuous with constant $G > 0$:

$$|f(x) - f(y)| \le G\|x - y\|_2 \qquad \forall x, y$$

this is equivalent to $\|g\|_2 \le G$ for all x and $g \in \partial f(x)$ (see next page)

Proof.

- assume $\|g\|_2 \leq G$ for all subgradients; choose $g_y \in \partial f(y)$, $g_x \in \partial f(x)$:

$$g_x^T(x-y) \geq f(x) - f(y) \geq g_y^T(x-y)$$

  by the Cauchy-Schwarz inequality

$$G\|x-y\|_2 \geq f(x) - f(y) \geq -G\|x-y\|_2$$

- assume $\|g\|_2 > G$ for some $g \in \partial f(x)$; take $y = x + g/\|g\|_2$:

$$\begin{aligned} f(y) &\geq f(x) + g^T(y-x) \\ &= f(x) + \|g\|_2 \\ &> f(x) + G \end{aligned}$$

Analysis

- the subgradient method is not a descent method
- the key quantity in the analysis is the distance to the optimal set

with $x^+ = x^{(i)}$, $x = x^{(i-1)}$, $g = g^{(i-1)}$, $t = t_i$:

$$\begin{array}{rcl}\|x^+ - x^\star\|_2^2 & = & \|x - tg - x^\star\|_2^2 \\ & = & \|x - x^\star\|_2^2 - 2tg^T(x - x^\star) + t^2\|g\|_2^2 \\ & \leq & \|x - x^\star\|_2^2 - 2t\left(f(x) - f^\star\right) + t^2\|g\|_2^2\end{array}$$

combine inequalities for $i = 1, \ldots, k$, and define $f_{\mathrm{best}}^{(k)} = \min_{0 \leq i < k} f(x^{(i)})$:

$$\begin{array}{rcl}2(\sum_{i=1}^k t_i)(f_{\mathrm{best}}^{(k)} - f^\star) & \leq & \|x^{(0)} - x^\star\|_2^2 - \|x^{(k)} - x^\star\|_2^2 + \sum_{i=1}^k t_i^2\|g^{(i-1)}\|_2^2 \\ & \leq & \|x^{(0)} - x^\star\|_2^2 + \sum_{i=1}^k t_i^2\|g^{(i-1)}\|_2^2\end{array}$$

Fixed step size: $t_i = t$

$$f_{\text{best}}^{(k)} - f^\star \leq \frac{\|x^{(0)} - x^\star\|_2^2}{2kt} + \frac{G^2 t}{2}$$

- does not guarantee convergence of $f_{\text{best}}^{(k)}$
- for large k, $f_{\text{best}}^{(k)}$ is approximately $G^2 t/2$-suboptimal

Fixed step length: $t_i = s/\|g^{(i-1)}\|_2$

$$f_{\text{best}}^{(k)} - f^\star \leq \frac{G\|x^{(0)} - x^\star\|_2^2}{2ks} + \frac{Gs}{2}$$

- does not guarantee convergence of $f_{\text{best}}^{(k)}$
- for large k, $f_{\text{best}}^{(k)}$ is approximately $Gs/2$-suboptimal

Diminishing step size: $t_i \to 0$, $\sum_{i=1}^{\infty} t_i = \infty$

$$f_{\text{best}}^{(k)} - f^\star \le \frac{\|x^{(0)} - x^\star\|_2^2 + G^2 \sum_{i=1}^{k} t_i^2}{2 \sum_{i=1}^{k} t_i}$$

can show that $(\sum_{i=1}^{k} t_i^2)/(\sum_{i=1}^{k} t_i) \to 0$; hence, $f_{\text{best}}^{(k)}$ converges to $f^\star$

Outline

Example: regularized logistic regression

Given $(x_i, y_i) \in \mathbb{R}^p \times \{0,1\}$ for $i = 1, \ldots n$, consider the logistic regression loss:

$$f(\beta) = \sum_{i=1}^n \Big(-y_i x_i^T \beta + \log(1 + \exp(x_i^T \beta)\Big)$$

This is a smooth and convex, with

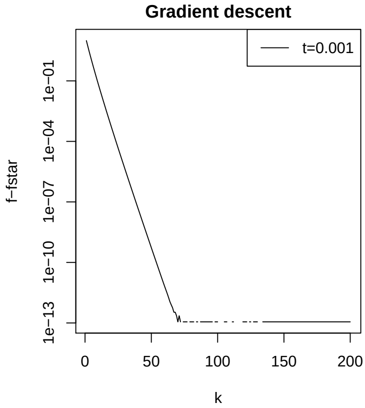
$$\nabla f(\beta) = \sum_{i=1}^n \big(y_i - p_i(\beta)\big) x_i$$

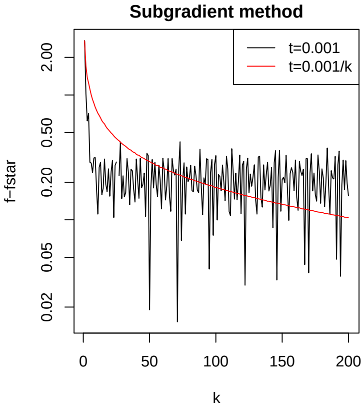
where $p_i(\beta) = \exp(x_i^T \beta)/(1 + \exp(x_i^T \beta))$, $i = 1, \ldots n$. We will consider the regularized problem:

$$\min_\beta \; f(\beta) + \lambda \cdot P(\beta)$$

where $P(\beta) = \|\beta\|_2^2$ (ridge penalty) or $P(\beta) = \|\beta\|_1$ (lasso penalty)

Ridge problem: use gradients; lasso problem: use subgradients. Data example with $n = 1000$, $p = 20$:





Step sizes hand-tuned to be favorable for each method (of course comparison is imperfect, but it reveals the convergence behaviors)

Outline

Outline

Proximal mapping

if h is convex and closed (has a closed epigraph), then

$$\mathrm{prox}_h(x) = \operatorname*{argmin}_u \left(h(u) + \frac{1}{2}\|u - x\|_2^2 \right)$$

exists and is unique for all x

- will be studied in more detail in lecture 8
- from optimality conditions of minimization in the definition:

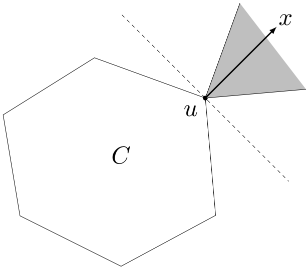
$$\begin{aligned} u = \mathrm{prox}_h(x) \quad &\Longleftrightarrow \quad x - u \in \partial h(u) \\ &\Longleftrightarrow \quad h(z) \geq h(u) + (x-u)^T(z-u) \quad \forall z \end{aligned}$$

Projection on closed convex set

proximal mapping of indicator function δ_C is Euclidean projection on C

$$\mathrm{prox}_{\delta_C}(x) = \operatorname*{argmin}_{u \in C} \|u - x\|_2^2 = P_C(x)$$

$$u = P_C(x)$$
$$\Updownarrow$$
$$(x-u)^T(z-u) \le 0 \quad \forall z \in C$$



we will see that proximal mappings have many properties of projections

Proximal gradient method

unconstrained optimization with objective split in two components

$$\text{minimize} \quad f(x) = g(x) + h(x)$$

- g convex, differentiable, $\operatorname{dom} g = \mathbf{R}^n$
- h convex with inexpensive prox-operator (many examples in lecture 8)

Proximal gradient algorithm

$$x^{(k)} = \operatorname{prox}_{t_k h}\left(x^{(k-1)} - t_k \nabla g(x^{(k-1)})\right)$$

- $t_k > 0$ is step size, constant or determined by line search
- can start at infeasible $x^{(0)}$ (however $x^{(k)} \in \operatorname{dom} f = \operatorname{dom} h$ for $k \geq 1$)

Interpretation

$$x^+ = \mathrm{prox}_{th}\left(x - t\nabla g(x)\right)$$

from definition of proximal mapping:

$$\begin{aligned} x^+ &= \underset{u}{\mathrm{argmin}}\left(h(u) + \frac{1}{2t}\left\|u - x + t\nabla g(x)\right\|_2^2\right) \\ &= \underset{u}{\mathrm{argmin}}\left(h(u) + g(x) + \nabla g(x)^T(u - x) + \frac{1}{2t}\|u - x\|_2^2\right) \end{aligned}$$

x^+ minimizes $h(u)$ plus a simple quadratic local model of $g(u)$ around x

Example: ISTA

Given $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, recall lasso criterion:

$$f(\beta) = \underbrace{\frac{1}{2}\|y - X\beta\|_2^2}_{g(\beta)} + \underbrace{\lambda\|\beta\|_1}_{h(\beta)}$$

Prox mapping is now

$$\begin{aligned} \mathrm{prox}_t(\beta) &= \underset{z}{\mathrm{argmin}} \ \frac{1}{2t}\|\beta - z\|_2^2 + \lambda\|z\|_1 \\ &= S_{\lambda t}(\beta) \end{aligned}$$

where $S_\lambda(\beta)$ is the soft-thresholding operator,

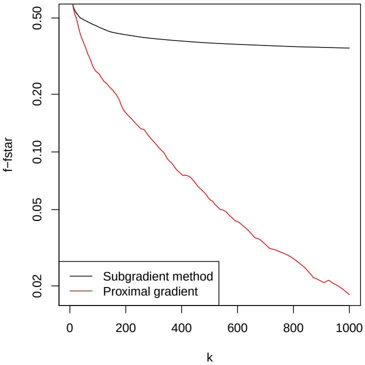
$$[S_\lambda(\beta)]_i = \begin{cases} \beta_i - \lambda & \text{if } \beta_i > \lambda \\ 0 & \text{if } -\lambda \le \beta_i \le \lambda\,, \\ \beta_i + \lambda & \text{if } \beta_i < -\lambda \end{cases} \quad i = 1, \dots n$$

Recall $\nabla g(\beta) = -X^T(y - X\beta)$, hence proximal gradient update is:

$$\beta^+ = S_{\lambda t}\big(\beta + tX^T(y - X\beta)\big)$$

Often called the iterative soft-thresholding algorithm (ISTA).[1] Very simple algorithm

Example of proximal gradient (ISTA) vs. subgradient method convergence rates



[1] Beck and Teboulle (2008), "A fast iterative shrinkage-thresholding algorithm for linear inverse problems"

Outline

Assumptions

$$\text{minimize} \quad f(x) = g(x) + h(x)$$

- h is closed and convex (so that prox_{th} is well defined)
- g is differentiable with $\text{dom}\, g = \mathbf{R}^n$
- there exist constants $m \geq 0$ and $L > 0$ such that the functions

  $$g(x) - \frac{m}{2}x^T x, \qquad \frac{L}{2}x^T x - g(x)$$

  are convex
- the optimal value $f^\star$ is finite and attained at $x^\star$ (not necessarily unique)

Implications of assumptions on g

Lower bound

- convexity of the the function $g(x) - (m/2)x^T x$ implies (page 1-18):

$$g(y) \geq g(x) + \nabla g(x)^T (y - x) + \frac{m}{2} \|y - x\|_2^2 \qquad \forall x, y \tag{1}$$

- if $m = 0$, this means g is convex; if $m > 0$, strongly convex (lecture 1)

Upper bound

- convexity of the function $(L/2)x^T x - g(x)$ implies (page 1-12):

$$g(y) \leq g(x) + \nabla g(x)^T (y - x) + \frac{L}{2} \|y - x\|_2^2 \qquad \forall x, y \tag{2}$$

- this is equivalent to Lipschitz continuity and co-coercivity of gradient (lecture 1)

Gradient map

$$G_t(x) = \frac{1}{t}\left(x - \mathrm{prox}_{th}(x - t\nabla g(x))\right)$$

$G_t(x)$ is the negative 'step' in the proximal gradient update

$$\begin{aligned} x^+ &= \mathrm{prox}_{th}\left(x - t\nabla g(x)\right) \\ &= x - tG_t(x) \end{aligned}$$

- $G_t(x)$ is not a gradient or subgradient of $f = g + h$
- from subgradient definition of prox-operator (page 6-7),

$$G_t(x) \in \nabla g(x) + \partial h\left(x - tG_t(x)\right)$$

- $G_t(x) = 0$ if and only if x minimizes $f(x) = g(x) + h(x)$

Consequences of quadratic bounds on g

substitute $y = x - tG_t(x)$ in the bounds (1) and (2): for all t,

$$\frac{mt^2}{2}\|G_t(x)\|_2^2 \leq g\left(x - tG_t(x)\right) - g(x) + t\nabla g(x)^T G_t(x) \leq \frac{Lt^2}{2}\|G_t(x)\|_2^2$$

- if $0 < t \leq 1/L$, then the upper bound implies

$$g(x - tG_t(x)) \leq g(x) - t\nabla g(x)^T G_t(x) + \frac{t}{2}\|G_t(x)\|_2^2 \qquad (3)$$

- if the inequality (3) is satisfied and $tG_t(x) \neq 0$, then $mt \leq 1$
- if the inequality (3) is satisfied, then for all z,

$$f(x - tG_t(x)) \leq f(z) + G_t(x)^T(x - z) - \frac{t}{2}\|G_t(x)\|_2^2 - \frac{m}{2}\|x - z\|_2^2 \qquad (4)$$

(proof on next page)

Proof of (4):

$$
\begin{aligned}
&f(x - tG_t(x)) \\
&\quad \leq \quad g(x) - t\nabla g(x)^T G_t(x) + \frac{t}{2}\|G_t(x)\|_2^2 + h(x - tG_t(x)) \\
&\quad \leq \quad g(z) - \nabla g(x)^T (z - x) - \frac{m}{2}\|z - x\|_2^2 - t\nabla g(x)^T G_t(x) + \frac{t}{2}\|G_t(x)\|_2^2 \\
&\qquad\qquad + h(z) - (G_t(x) - \nabla g(x))^T (z - x + tG_t(x)) \\
&\quad = \quad g(z) + h(z) + G_t(x)^T (x - z) - \frac{t}{2}\|G_t(x)\|_2^2 - \frac{m}{2}\|x - z\|_2^2
\end{aligned}
$$

- in the first step we add $h(x - tG_t(x))$ to both sides of the inequality (3)
- in the next step we use the lower bound on $g(z)$ from (2) and

$$G_t(x) - \nabla g(x) \in \partial h(x - tG_t(x))$$

(see page 6-12)

Progress in one iteration

for a step size t that satisfies the inequality (3), define

$$x^+ = x - tG_t(x)$$

- inequality (4) with $z = x$ shows the algorithm is a descent method:

$$f(x^+) \leq f(x) - \frac{t}{2}\|G_t(x)\|_2^2$$

- inequality (4) with $z = x^\star$ shows that

$$\begin{aligned} f(x^+) - f^\star &\leq G_t(x)^T(x - x^\star) - \frac{t}{2}\|G_t(x)\|_2^2 - \frac{m}{2}\|x - x^\star\|_2^2 \\ &= \frac{1}{2t}\left(\|x - x^\star\|_2^2 - \|x - x^\star - tG_t(x)\|_2^2\right) - \frac{m}{2}\|x - x^\star\|_2^2 \\ &= \frac{1}{2t}\left((1 - mt)\|x - x^\star\|_2^2 - \|x^+ - x^\star\|_2^2\right) && (5) \\ &\leq \frac{1}{2t}\left(\|x - x^\star\|_2^2 - \|x^+ - x^\star\|_2^2\right) && (6) \end{aligned}$$

Analysis for fixed step size

add inequalities (6) for $x = x^{(i-1)}$, $x^+ = x^{(i)}$, $t = t_i = 1/L$

$$\begin{aligned}\sum_{i=1}^{k}(f(x^{(i)}) - f^\star) &\leq \frac{1}{2t}\sum_{i=1}^{k}\left(\|x^{(i-1)} - x^\star\|_2^2 - \|x^{(i)} - x^\star\|_2^2\right) \\ &= \frac{1}{2t}\left(\|x^{(0)} - x^\star\|_2^2 - \|x^{(k)} - x^\star\|_2^2\right) \\ &\leq \frac{1}{2t}\|x^{(0)} - x^\star\|_2^2\end{aligned}$$

since $f(x^{(i)})$ is nonincreasing,

$$f(x^{(k)}) - f^* \leq \frac{1}{k}\sum_{i=1}^{k}(f(x^{(i)}) - f^\star) \leq \frac{1}{2kt}\|x^{(0)} - x^\star\|_2^2$$

Distance to optimal set

- from (5) and $f(x^+) \geq f^\star$, the distance to the optimal set does not increase:

$$\begin{aligned}\|x^+ - x^\star\|_2^2 &\leq (1 - mt)\|x - x^\star\|_2^2 \\ &\leq \|x - x^\star\|_2^2\end{aligned}$$

- for fixed step size $t_k = 1/L$

$$\|x^{(k)} - x^\star\|_2^2 \leq c^k \|x^{(0)} - x^\star\|_2^2, \qquad c = 1 - \frac{m}{L}$$

i.e., linear convergence if g is strongly convex ($m > 0$)

Outline

Accelerated proximal gradient method

Our problem, as before:

$$\min_x \; g(x) + h(x)$$

where g convex, differentiable, and h convex. Accelerated proximal gradient method: choose initial point $x^{(0)} = x^{(-1)} \in \mathbb{R}^n$, repeat:

$$\begin{aligned} v &= x^{(k-1)} + \frac{k-2}{k+1}(x^{(k-1)} - x^{(k-2)}) \\ x^{(k)} &= \mathrm{prox}_{t_k}\big(v - t_k \nabla g(v)\big) \end{aligned}$$

for $k = 1, 2, 3, \ldots$

- First step $k = 1$ is just usual proximal gradient update
- After that, $v = x^{(k-1)} + \frac{k-2}{k+1}(x^{(k-1)} - x^{(k-2)})$ carries some "momentum" from previous iterations
- $h = 0$ gives accelerated gradient method

FISTA

Recall lasso problem,

$$\min_{\beta} \; \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

and ISTA (Iterative Soft-thresholding Algorithm):

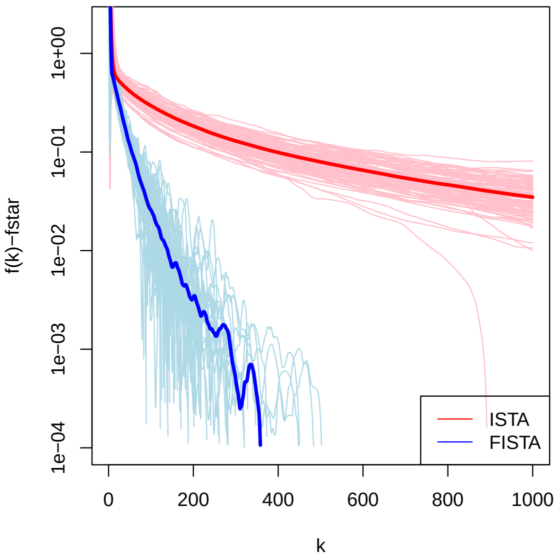
$$\beta^{(k)} = S_{\lambda t_k}\big(\beta^{(k-1)} + t_k X^T (y - X\beta^{(k-1)})\big), \quad k = 1, 2, 3, \ldots$$

$S_\lambda(\cdot)$ being vector soft-thresholding. Applying acceleration gives us FISTA (F is for Fast):[6] for $k = 1, 2, 3, \ldots,$

$$\begin{aligned} v &= \beta^{(k-1)} + \frac{k-2}{k+1}(\beta^{(k-1)} - \beta^{(k-2)}) \\ \beta^{(k)} &= S_{\lambda t_k}\big(v + t_k X^T (y - Xv)\big), \end{aligned}$$

[6] Beck and Teboulle (2008) actually call their general acceleration technique (for general g, h) FISTA, which may be somewhat confusing

Lasso regression: 100 instances (with $n = 100$, $p = 500$):



Lasso logistic regression: 100 instances $(n = 100,\ p = 500)$: