Monday, October 20, 2014 3:11 PM

Today: Course Projects
Proximal Gradient
Fenchel Duality

A2: Pick up (end of class)
A3: Marked version due now
A4: Marked version due next Monday
A5: due wednesday next week

Course Project Suggestions

- apply ML to a new domain
  (supervised problem form)
  - your research? kaggle competition?
  - compare/contrast ML methods from class
- Compare different extensions of a method from class
  on a set of test problems (add your own extensions too!)
- Explore scaling methods up to BIG data sets
  (eg. kernels) (add your own strategies, too)
- Prove a theoretical result
- Do a larger coding project (implement a bunch of ML methods)
- Parallel / distributed stochastic gradient
- apply SAG/SURG to a non-linear, non-convex problem
- convergence rates
- fenchel dual of a non-linear, non-convex problem

Project Proposal: Due Nov 3

- groups 1-3
- max 2 pages (shorter ok)
- audience is Prof (no need for background discussion)
- just a sanity check
  (not marked - no news is good)

Last time (convergence rates)

today: composite, proximal-gradient (back to linear convergence rate)

Gradient: linear convergence

Subgradient: sublinear (rate)

Stochastic Gradient: sublinear convergence (rate)

finite SAG: back to linear convergence rate

tomorrow: 12:30, ESB 4133 (Tues. Oct 21)
"SCAIM Seminar"

Gradient Method

(smooth)

- we want to solve $\min_{x \in \mathbb{R}^d} f(x)$
- minimize quadratic approximation

$$x^{t+1} = \operatorname{argmin}_{y \in \mathbb{R}^d} f(x^t) + \langle \nabla f(x^t), y - x \rangle + \frac{1}{2}\alpha_t \|y-x\|^2$$

(inner product of gradient)

- solution is gradient method:

$$x^{t+1} = x^t - \alpha_t \nabla f(x^t)$$

Guaranteed decrease if $\alpha_t < \frac{2}{L}$

(small step in gradient direction)

Projected Gradient Method

(smooth)

- we want to solve $\min_{x \in C} f(x)$
- minimize quadratic approximation

$$x^{t+1} = \operatorname{argmin}_{y \in C} f(x^t) + \langle \nabla f(x^t), y - x \rangle + \frac{1}{2}\alpha_t \|y-x\|^2$$

(if we base norm on Hessian, we can use more interesting things, but it's hard.)

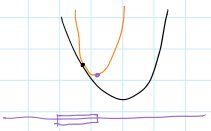
- solution is gradient method:

$$x^{t+1} = P_C\left[x^t - \alpha_t \nabla f(x^t)\right]$$

"projection"

Guaranteed decrease if $\alpha_t < \frac{2}{L}$

- Linear convergence rate

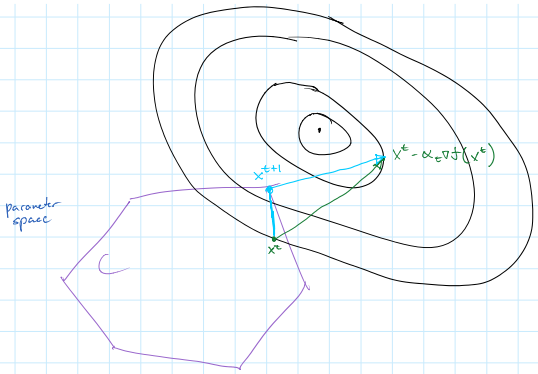


$$P_C[x] = \operatorname{argmin}_{y \in C} \|y - x\|$$

$$C = \{x \mid LB \leq x \leq UB\}$$

$$P_C[x] = \begin{cases} LB & \text{if } x < LB \\ x & \text{if } LB \leq x \leq UB \\ UB & \text{if } x > UB \end{cases}$$

- you can show x^* is optimal (if and only if) iff $x^* = P_C[x^* - \nabla f(x^*)]$
  (similar to gradient of zero)



Proximal Gradient Method

(smooth) (non-smooth)

- we want to solve $\min_{x \in \mathbb{R}^d} f(x) + r(x)$

- minimize quadratic approximation

$$x^{t+1} = \operatorname{argmin}_{y \in \mathbb{R}^d} f(x^t) + \langle \nabla f(x^t), y - x \rangle + \frac{1}{2}\alpha_t \|y - x\|^2 + r(x)$$

- solution is gradient method:

$$x^{t+1} = \operatorname{prox}_{\alpha r(\cdot)}\left[x^t - \alpha_t \nabla f(x^t)\right]$$

"Proximal Operator"

Guaranteed decrease if $\alpha_t < \frac{2}{L}$

- Linear convergence rate

Note: in the case where:

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

it is equivalent to projection as above.

$$\operatorname{prox}_{\alpha r(\cdot)}[x] \quad \operatorname*{argmin}_{y \in \mathbb{R}^d} \frac{1}{2}\|y - x\|^2 + \alpha r(y)$$

in the case where $r(x) = \lambda \|x\|_1$

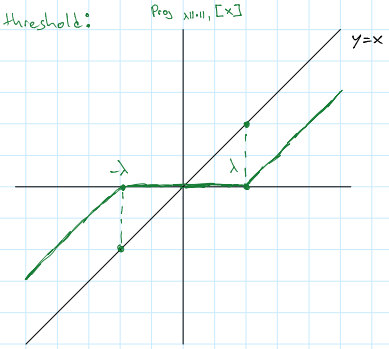
$$x^{t+1/2} = x^t - \alpha_t \nabla f(x^t)$$

$$x^{t+1} = \operatorname{sign}(x^{t+1/2}) \max\{0, |x^{t+1/2}| - \alpha_t \lambda\}$$

"soft threshold"

(sparse iterations as you go)

Soft threshold: $\operatorname{Prox}_{\lambda\|\cdot\|_1}[x]$



⊛ Other constrained / non-smooth optimizers:

- penalty methods (turn constrained to unconstrained with penalty)
- log barrier
- augmented Lagrangian
- ADMM
- Interior Point

Fenchel Duality

Why?
- generic smoothing of non-smooth convex functions
- dual may have fewer variables
- generic Kernel trick
- formulate non-convex problems as convex problems

⊛ $\min_{x_1, x_2} f(x_1) + g(x_2)$, s.t. $x_1 = Ax_2$ (take lagrange dual)

Primal Problem: $\min_{x \in \mathbb{R}^d} \; p(x) = f(Ax) + g(x)$ (no restriction on g)

Fenchel Dual: $\max_{y \in \mathbb{R}^n} \; D(y) = -f^*(y) - g^*(A^T y)$

f^* is the "convex conjugate" of f

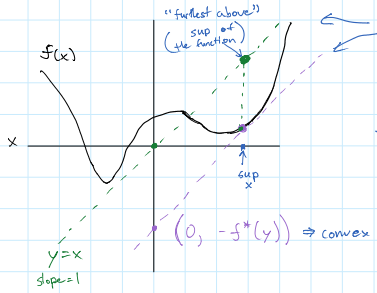
Notes: (properties of the dual)
1. $y \in \mathbb{R}^n$ $d \Rightarrow n$ (can be smaller)
2. $D(y) \leq p(x)$ (weak duality)
3. $D(y^*) = p(x^*)$ (strong duality) under conditions (f, g are convex)
4. $-D(y)$ is always convex
5. If $p(x)$ strongly convex $\overset{\text{then}}{\Rightarrow}$ $D(y)$ is smooth
6. $f^{**} = f$, if f is convex and 'closed' ⊛

$f(x) + \frac{1}{2}\|x\|^2 \Rightarrow$ dual is smooth

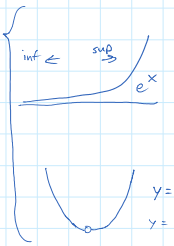
Convex Conjugate

$$f^*(y) = \sup_x \{ y^T x - f(x) \}$$

$$= -\inf_x \{ -y^T x + f(x) \}$$



What is 'sup'? (similar to max)



$\inf_x \{e^x\} = 0$ $\quad \min_x \{e^x\} \to 0$ (DNE)
no "actual" value of x (it's the limit)
so; min DNE

$y = x^2$, $x \neq 0$ • min DNE
$y = \infty$, $x = 0$ • no arg inf
• inf = 0

Interesting note: If $f(x)$ is convex and differentiable, the maximum gap occurs at point x_m where $\nabla f(x_m) = y$.

Examples:

1. $f(x) = \frac{1}{2}\|x\|^2$

$$f^*(y) = \sup_x \{ y^T x - \tfrac{1}{2}\|x\|^2 \}$$

take derivative, set to 0

$0 = y - x \qquad x = y$

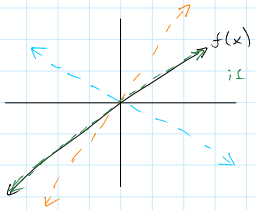
$$f^*(y) = y^T y - \tfrac{1}{2}\|y\|^2$$

$$f^*(y) = \|y\|^2$$

(same as original! this is the only function where this will happen)

2. $f(x) = a^T x$

$$f^*(y) = \sup_x \{ y^T x - a^T x \}$$



if $a = y$, sup = 0 (it's on the line)

$$f^*(y) = \begin{cases} 0 & \text{if } y = a \\ \infty & \text{elsewise} \end{cases}$$

Fenchel Dual with L_2 Regularization

$$P(x) = f(Ax) + \frac{\lambda}{2}\|x\|^2$$

$$-D(x) = f^*(y) + \underbrace{\frac{1}{2\lambda}\|A^T y\|^2}_{\frac{1}{2\lambda} y^T A A^T y} \longrightarrow \text{gram matrix}$$

replace this with kernel matrix

* kernel trick
  for any L_2-reg linear model

take fenchel and do coordinate descent - state of the art!