Fantastic Question from Last Lecture

> Doing a linesearch over α on $\text{prox}_{\alpha h}(x - \alpha \nabla f(x))$ might not move in a straight line, right?

That is true, but if prox is cheap, we can still compute it at a logarithmic # α guesses

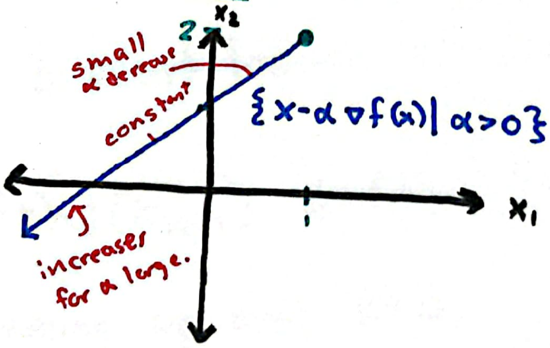
$$\alpha_k = \sup\{\alpha\tau^n \mid n = 0,1,2,3,\ldots \quad f{+}h(x^+) \text{ "}\leq\text{" } f{+}h(x_k)\}$$

$x^+ = \text{prox}_{\alpha\tau^n h}(x_k - \alpha\tau^n \nabla f(x_k))$

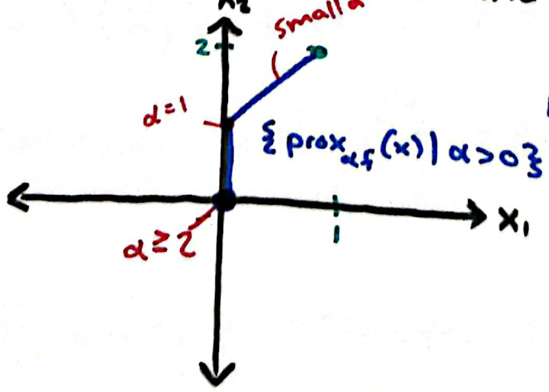
Example $f(x) = |x_1| + |x_2|$

Given $x = (1,2)$, lets look at the GD linesearch:

$\nabla f(1,2) = (1,1)$



Instead, we can look at the prox linesearch



$$[\text{prox}_{\alpha\|\cdot\|_1}(x)]_i = \begin{cases} x_i - \alpha & \text{if } x_i > \alpha \\ x_i + \alpha & \text{if } x_i < -\alpha \\ 0 & \text{if } x_i \in [-\alpha, \end{cases}$$

Picking $\alpha = \frac{1}{L}$ gets descent

$$(f+h)(x^+) \le (f+h)(x) - \frac{1}{2L} \| G_{1/L}(x) \|^2$$

Linesearching (exact, backtracking) work exactly the same.

(Beck Ch 10 repeats these for us).

<u>Theorem</u> For any f with L-Lipschitz gradient and convex h, selecting $\alpha_k = 1/L$, the proximal gradient method has

$$\frac{1}{T} \sum_{k=0}^{T-1} \| G_{\alpha}(x_k) \|_2^2 \le \frac{2L(f+h(x_0) - \min f+h)}{T}.$$

Proof. Our descent lemma at each iteration gives

$$(f+h)(x_{k+1}) \le (f+h)(x_k) - \frac{1}{2L} \| G_{1/L}(x_k) \|_2^2.$$

Summing up over $k=0,\dots,T-1$

$$\min (f+h)(x) \le (f+h)(x_T) \le (f+h)(x_0) - \frac{1}{2L} \sum_{k=0}^{T-1} \| G_{1/L}(x_k) \|_2^2.$$

Simplifying gives $\frac{1}{T} \sum_{k=0}^{T-1} \| G_{1/L}(x_k) \|_2^2 \le \frac{2L((f+h)(x_0) - \min f+h)}{T}$. $\square$

$\Rightarrow$ Approaching necessary condition from last time.

Continuing our parallel analysis to smooth opt.
Look at f being convex.

<u>Theorem</u> For any convex f with L-Lipschitz gradient and h convex, the proximal gradient method has

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

where x^* minimizes $f+h$, and $\alpha_k = 1/L$.

Proof. Lets show the following nice condition:

$$0 \leq \left((f+h)(x_{k+1}) - (f+h)(x^*)\right) \leq \frac{L}{2}\|x_k - x^*\|^2 - \frac{L}{2}\|x_{k+1} - x^*\|^2$$

Proof. Our prox grad method computes x_{k+1} minimizing

$$\varphi(x) = \left[f(x_k) + \nabla f(x_k)^T(x - x_k) + \frac{L}{2}\|x - x_k\|_2^2 . \right] + h(x)$$

$\Rightarrow$ φ is $0 + L = L$-strongly convex.

HW 2, Q3(a), we then know

$$\varphi(x^*) - \varphi(x_{k+1}) \geq \frac{L}{2}\|x^* - x_{k+1}\|_2^2 .$$

$\Rightarrow$ Characterization of smooth convex func

$[\;\,] \geq f(x)$

$\Rightarrow (f+h)(x_{k+1}) \leq \varphi(x_{k+1})$

Convexity of f ——— $\leq f(x)$

$\Rightarrow (f+h)(x^*) + \frac{L}{2}\|x^* - x_k\|_2^2 \geq \varphi(x^*)$

$$\Longrightarrow (f+h)(x^*) - (f+h)(x_{k+1}) + \frac{L}{2}\|x^* - x_k\|_2^2 \geq \frac{L}{2}\|x^* - x_{k+1}\|_2^2 .$$

$\square$

Summing this up, gives (and divide by T)

$$\frac{1}{T}\sum_{k=0}^{T-1}\left((f+h)(x_{k+1}) - (f+h)(x^*)\right) \le \frac{1}{T}\frac{L}{2}\|x_0-x^*\|_2^2 - \underbrace{\frac{1}{T}\frac{L}{2}\|x_T-x^*\|_2^2}_{\geq 0}$$

$$\le \frac{L}{T2}\|x_0-x^*\|_2^2$$

Our descent lemma ensures $(f+h)(x_{k+1}) - (f+h)(x^*)$ is decreasing / nonincreasing.

$$\Rightarrow f^{+h}(x_T) - (f+h)(x^*) \le \frac{1}{T}\sum_{k=0}^{T-1}\left(f+h(x_{k+1}) - (f+h)(x^*)\right). \quad \square$$

Things should speedup to linear / geometric convergence under strong convexity. (Lets steal the restarting arguement. Easier since "no memory").

<u>Theorem</u> In addition to our previous assumptions, suppose $f+h$ is μ-strongly convex. Then prox grad method converges linearly.

Proof. By previous theorem

HW2, Q3, $(f+h)(x_0) - (f+h)(x^*) \geq \frac{\mu}{2}\|x_0-x^*\|_2^2$

$$f^{+h}(x_{k+1}) - (f+h)(x^*) \le \frac{L\|x_0-x^*\|_2^2}{2k}$$

$$\le \frac{L}{\mu k}\cdot\left((f+h)(x_0) - (f+h)(x^*)\right).$$

$\Rightarrow$ Pick $k \geq \lceil 2 L/\mu \rceil$, gives us half the gap.

Reach accuracy ε after $\log_2\left(\frac{(f+h)(x_0) - (f+h)(x^*)}{\varepsilon}\right)$. $\square$

4. Acceleration

Define $y_0 = x_0$, $\lambda_0 = 0$

$$y_{k+1} = \mathrm{prox}_{\alpha h}(x_k - \alpha \nabla f(x_k))$$

$$x_{k+1} = y_{k+1} + \frac{\lambda_k - 1}{\lambda_{k+1}}(y_{k+1} - y_k)$$

where $\lambda_{k+1} = \dfrac{1 + \sqrt{1 + 4\lambda_k^2}}{2}$

"Accelerated / Fast Proximal/Projected Gradient Method"

"FISTA" ← good name for LASSO.

Theorem (10.34 of Beck)

For any convex f with L-Lipschitz gradient and convex h, the accelerated method with $\alpha = 1/L$ has

$$(f+h)(y_k) - (f+h)(x^*) \leq \frac{2L\,\|x_0 - x^*\|_2^2}{(k+1)^2}$$

Proof. See Beck, nearly identical to our smooth opt.

4. More Proximal Methods

Alternating Projections

Given convex sets S_i, find $x \in \bigcap_{i=1}^{n} S_i$

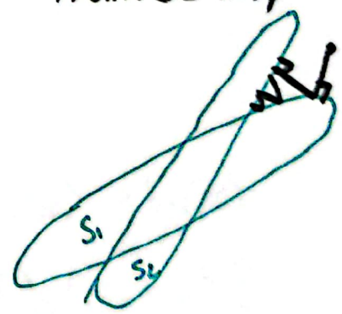
$\text{proj}_{\cap S_i}(x)$ is often hard

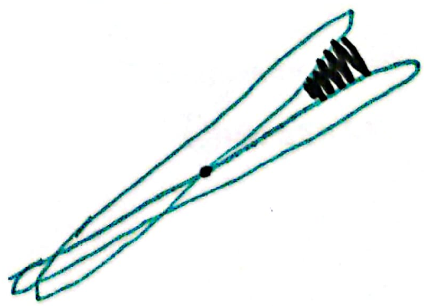
$\text{proj}_{S_i}(x)$ may be easy.

$$x_{k+1} = \text{proj}_{S_n}(\text{proj}_{S_{n-1}}(\cdots \text{proj}_{S_1}(x_k)) \cdots)$$

<u>Thm</u> If $\bigcap_{i=1}^{n} S_i \neq \emptyset$, then $x_k \to x^* \in \cap S_i$.

Rate is controlled by "tranversality"





$$\min \ f(Ax) + h(Bx)$$

both simple, proxable

$$\min \ f(y) + h(z) \quad \text{s.t.} \begin{cases} Ax = y \\ Bx = z \\ x \in \mathbb{R}^d \end{cases} \iff \min \ f(y) + h(z) + \delta(y, z)$$

proxable

Lagrange multipliers

ADMM

(Alternating Direction Method of Multipliers).

[Next Semester w/ Duality]

Mirror Descent / Bregman Methods

Improve on 2-norm

$$x_{k+1} = \operatorname{argmin} \left\{ f(x_k) + \nabla f(x_k)^T (x - x_k) + h(x) + D_\phi(x, x_k) \right\}$$

Bregman Divergence

KL-Divergence

$x, y \in \Delta_n$ $\quad D_\phi(x, y) = \sum x_i \log\left(\frac{x_i}{y_i}\right)$

Beautiful Duality Theory

Open Questions about how accelerate this?

Next Week

Tuesday Subgradient Methods

Thursday Stochastic Methods

Friday Midterm posted