Starting 10/5

Note on homeworks:

Please only submit standard file formats

.pdf, .zip, images (jpeg, png, etc).

HW1 posted grades (shortly)

Questions about grading details, see corresponding TA:

Q1 ← Jinke
Q2 ← Thabo
Q3 ← Ning
Q4 ← Salma

(~16/20)

3. Projected/Proximal Gradient Descent

Two examples currently motivating us...

LASSO $\min_{x \in \mathbb{R}^d} \|Ax-b\|_2^2 + \gamma \|x\|_1$

Grading Program $\min_{x \in \mathbb{R}^d} f(x) + \delta_S(x)$

Lets consider the more general model

$$\min_{x \in \mathbb{R}^d} f(x) + h(x)$$

f: smooth, nonconvex(?)

h: convex, "structured" (i.e. we can compute $\mathrm{prox}_{\alpha h}(\cdot)$)

Lemma (Sum Rule) $\partial(f+h)(x) = \partial f(x) + \partial h(x)$

if f, h convex.

Proof. "$\supseteq$" Let $g_1 \in \partial f(x)$, $g_2 \in \partial h(x)$

$$
\begin{aligned}
& f(y) \geq f(x) + g_1^T(y-x) \quad \forall y \\
+ \; & h(y) \geq h(x) + g_2^T(y-x) \quad \forall y \\
\hline
& (f+h)(y) \geq (f+h)(x) + (g_1+g_2)^T(y-x) \; \forall y. \checkmark
\end{aligned}
$$

"$\subseteq$" Harder to prove, Need Seperating Hyperplane Thm.

("Intro to Convexity") $\square$

~~Lemma~~ For particular problem, we consider

$$\nabla f(x) + \partial h(x).$$

prox is our angle of attack.

Lemma (Proximal First-Order Condition). For any f with L-Lipschitz gradient (only need C^1) and h convex,

if x^* is a local min of $f+h$, then

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

Proof. Suppose $-\nabla f(x^*) \notin \partial h(x^*)$.

$\Rightarrow$ Some $y \in \mathbb{R}^d$ s.t. $\underline{h(y) < h(x^*) + -\nabla f(x^*)^T(y-x^*)}$

Consider $z = x^* + \lambda(y - x^*)$ for $0 \le \lambda \le 1$.

$$\begin{aligned} h(z) &\le (1-\lambda)h(x^*) + \lambda h(y) \\ &= h(x^*) + \lambda(\underline{h(y) - h(x^*)}) \\ &< h(x^*) - \lambda \nabla f(x^*)^T(y - x^*) \end{aligned}$$

As $\lambda \to 0$, $z \to x^*$, $\nabla f(x^*)^T(y-x^*) \leftrightarrows \dfrac{f(z) - f(x^*)}{\lambda}$

$\Rightarrow h(z) < h(x^*) - \lambda \cdot \dfrac{f(z) - f(x^*)}{\lambda}$

$\Rightarrow (f+h)(z) < (f+h)(x^*)$.

$\Rightarrow$ z is better than x^*

$\Rightarrow$ x^* is not a local min. $\square$

Lets develop an iterative algorithm to min $f+h$:

At step k, we can linearize f

$$(f+h)(x) \approx \underbrace{f(x_k) + \nabla f(x_k)^T(x - x_k)}_{\substack{1^{st}\text{-order model} \\ \text{of } f \text{ at } x_k.}} + h(x)$$

Lets iterate

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

$$= \underset{x \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ f(x_k) + \nabla f(x_k)^T (x - x_k) + h(x) + \frac{1}{2\alpha_k} \|x - x_k\|_2^2 \right\}$$

$$= \operatorname{argmin} \left\{ \underbrace{f(x_k) - \frac{1}{2\alpha_k} \|\nabla f(x_k)\|^2}_{\text{constant}} + h(x) + \frac{1}{2\alpha_k} \|x - (x_k - \alpha_k \nabla f(x_k))\|^2 \right\}$$

$$= \operatorname{argmin} \left\{ h(x) + \frac{1}{2\alpha_k} \|x - (x_k - \alpha_k \nabla f(x_k))\|^2 \right\}$$

$$= \text{prox}_{\alpha_k h}(x_k - \alpha_k \nabla f(x_k)).$$

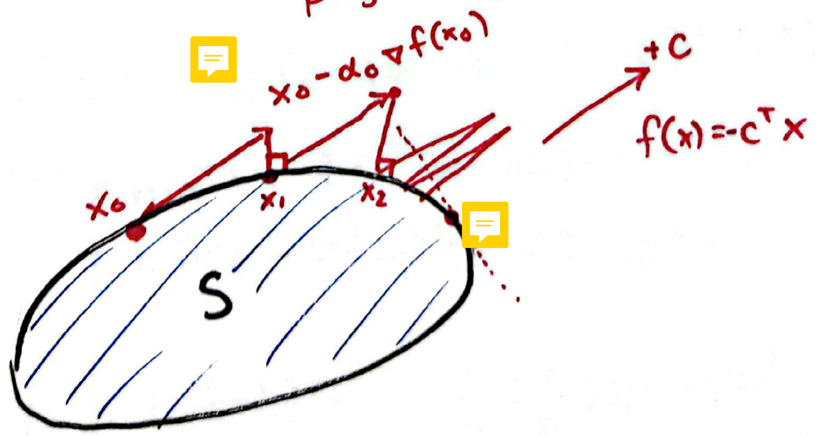
This alternates grad descent on f, prox on h, repeat....

Called "projected/proximal gradient descent"

"ISTA"

"Forward-Backward Method"

projected gradient descent.



Define $G_\alpha(x) = \frac{1}{\alpha}\left(x - \underbrace{\text{prox}_{\alpha h}(x - \alpha \nabla f(x))}_{\text{"}x^+}\right)$

as the gradient mapping.

Check "gradient-like". HW 2, Q2.

$$\frac{1}{\alpha}\left(x - \alpha\nabla f(x) - \text{prox}_{\alpha h}(x - \alpha\nabla f(x))\right) \in \partial h(x^+)$$

$$\Leftrightarrow \quad G_\alpha(x) \in \nabla f(x) + \partial h(x^+)$$

Small $G_\alpha(x) \Rightarrow x, x^+$ are essentially the same.

Lemma (Descent) For any x, let $x^+ = \text{prox}_{\alpha h}(x - \alpha\nabla f(x))$.

$$(f+h)(x^+) \le (f+h)(x) - \left(\alpha - \frac{L\alpha^2}{2}\right)\|G_\alpha(x)\|^2$$

whenever ∇f is L-Lipschitz.

Proof. By our Taylor Approximation Thms,

$$f(x^+) \le f(x) + \nabla f(x)^T(x^+ - x) + \frac{L}{2}\|x^+ - x\|_2^2 . \quad (1)$$

By HW2, Q2, $\frac{1}{\alpha}(x - \alpha\nabla f(x) - x^+) \in \partial h(x^+)$

$$\Rightarrow h(x) \ge h(x^+) + \frac{1}{\alpha}(x - \alpha\nabla f(x) - x^+)^T(x - x^+)$$

$$= h(x^+) - \nabla f(x)^T(x - x^+) + \frac{1}{\alpha}\|x - x^+\|^2 . \quad (2)$$

(1) - (2)

$$(f+h)(x^+) \le (f+h)(x) - \left(\frac{1}{\alpha} - \frac{L}{2}\right)\|x - x^+\|^2$$

$$= (f+h)(x) - \left(\alpha - \frac{L\alpha^2}{2}\right)\|G_\alpha(x)\|^2 .$$

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

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

Proof. Our descent lemma at each iteration gives

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