Lecture 15. Newton's method cont'd; Proximal gradient descent

Recall the convergence analysis of univariate functions.

$|x_{k+1} - x^*| \le C |x_k - x^*|^2$ by Taylor expansion if $|f'''(x)| \le M$.

also proved by a slightly loose condition $|f''(x) - f''(y)| \le M|x-y|$.

To define M-Lipschitz for $\nabla^2 f(x)$. where $f: \mathbb{R}^n \to \mathbb{R}$. define norm first.

Operator norm: $\|Q\|_{a,b} \triangleq \max_{x \neq 0} \frac{\|Qx\|_b}{\|x\|_a} \Rightarrow \|Qx\|_b \le \|Q\|_{a,b} \|x\|_a$

In particular. if $a = b = 2$. $\|\cdot\|_{a,b}$ is called spectral norm. $\|\cdot\|_2$.

Proposition. $\|Q\|_2 = (\lambda_{max}(Q^T Q))^{1/2}$ since $\|Qx\|^2 = x^T Q^T Q x \le \lambda_{max} \|x\|^2$.

$\nabla^2 f(x)$ is M-Lipschitz if $\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \le M \|x-y\|_2$. $\forall x. y$.

Theorem. If f is m-strongly convex. $\nabla^2 f$ is M-Lipschitz. $\{x_k\}$ produced by Newton's method. then $\|x_{k+1} - x^*\| \le \frac{M}{2m} \|x_k - x^*\|^2$.

Remark: let $y_k = \frac{M}{2m} \|x_k - x^*\|$. then $y_{k+1} \le y_k^2 \Rightarrow y_k \le y_0^{2^k}$.

Proof. $\|x_{k+1} - x^*\|_2 \le \|(\nabla^2 f(x_k))^{-1}\|_2 \|\nabla^2 f(x_k)(x_k - x^*) - (\nabla f(x_k) - \nabla f(x^*))\|_2$.

$= \|(\nabla^2 f(x_k))^{-1}\|_2 \left\| \int_0^1 \left(\nabla^2 f(x_k) - \nabla^2 f(x^* + t(x_k - x^*))\right)(x_k - x^*)\, dt \right\|_2$.

$\le \frac{1}{m} \|x_k - x^*\|_2 \cdot \int_0^1 \|\nabla^2 f(x_k) - \nabla^2 f(x^* + t(x_k - x^*))\|_2 \, dt$.

$\le \frac{M}{m} \|x_k - x^*\|_2 \cdot \int_0^1 (1-t) \|x_k - x^*\|_2 \, dt = \frac{M}{2m} \|x_k - x^*\|_2^2$. □

In summary. descent method $x_{k+1} = x_k - t_k \cdot d_k$. $x_k \to x^*$

direction: gradient descent $x_{k+1} = x_k - t_k \nabla f(x_k)$.

Newton's method $x_{k+1} = x_k - t_k (\nabla^2 f(x_k))^{-1} \nabla f(x_k)$.

step size: fixed step size / exact line search / backtracking

convergence: by Lyapunov's stability theorem. if $f(x_{k+1}) < f(x_k)$.

order and rate: $|f(x_{k+1}) - f(x^*)| \le \alpha |f(x_k) - f(x^*)|$ for some $\alpha < 1$.

$\lim_{n\to\infty} \frac{|x_{n+1} - x^*|}{|x_n - x^*|^\rho} = \mu$.

ρ: order of convergence

μ: rate of convergence.

linear convergence for gradient descent.

$\|x_{k+1} - x^*\| \le C \|x_k - x^*\|^2$ for some $C < \infty$

quadratic convergence for Newton's method.

condition:	gradient descent	Newton's method.	
	L-smooth $\Rightarrow$ convergence.	$\nabla^2 f(x)$ M-Lipschitz	} convergence and rate.
	m-strongly convex. $\Rightarrow$ rate.	m-strongly convex	

How to deal with non differentiable functions: proximal gradient descent.

consider gradient descent: $x_{k+1} = x_k - t \nabla f(x)$.

$$x_{k+1} = \arg\min_y \; f(x_k) + \nabla f(x_k)^T (y - x_k) + \frac{1}{2t} \|y - x_k\|^2.$$

if f is L-smooth. and $t \le 1/L$. RHS gives an upper bound of f.

more sophisticated. $x_{k+1} = \arg\min_y \; f(x_k) + \nabla f(x_k)^T (y - x_k) + \frac{1}{2} (y - x_k)^T \nabla^2 f(x_k) (y - x_k)$.

Now assume f is not differentiable. suppose $f(x) = g(x) + h(x)$.

where $g(x)$ is convex and differentiable. and $h(x)$ is convex.

let $x_{k+1} = \underset{y}{\operatorname{argmin}}\ g(x_k) + \nabla g(x_k)^T (y - x_k) + \frac{1}{2t}\|y - x_k\|^2 + h(y)$.

$= \underset{y}{\operatorname{argmin}}\ \frac{1}{2t}\|y - (x_k - t\nabla g(x_k))\|^2 + h(y)$.

$\frac{1}{2t}\|y - (x - t\nabla g(x))\|^2$: closed to update for g. $h(y)$: make h small.

Proximal mapping: $\text{prox}_{h,t}(x) = \underset{y}{\operatorname{argmin}}\ \frac{1}{2t}\|x - y\|^2 + h(y)$

Proximal gradient descent: $x_{k+1} = \text{prox}_{h,t}(x_k - t \cdot \nabla g(x_k))$.

$x_{k+1} = x_k - t_k G_{t_k}(x_k)$. where $G_t(x) = \frac{1}{t}(x - \text{prox}_{h,t}(x - t \cdot \nabla g(x)))$

Key point is that $\text{prox}_{h,t}(x)$ has closed form for many important h.

Lasso: least absolute shrinkage and selection operator. (to reduce overfitting)

$f(\beta) = \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \qquad h(\beta) = \lambda\|\beta\|_1 \quad (\lambda > 0)$.

proximal mapping $\text{prox}_{h,t}(\beta) = \underset{y}{\operatorname{argmin}}\ \frac{1}{2t}\|\beta - y\|_2^2 + \lambda\|y\|_1$

$\text{prox}_{h,t}(\beta) = \underset{y}{\operatorname{argmin}}\ \sum\left(\frac{1}{2t}(\beta_i - y_i)^2 + \lambda|y_i|\right)$.

$\Rightarrow [\text{prox}_{h,t}(\beta)]_i = \underset{y_i}{\operatorname{argmin}}\ \frac{1}{2t}(\beta_i - y_i)^2 + \lambda|y_i| = \begin{cases} \beta_i - \lambda t & \text{if } \beta_i > \lambda t \\ 0 & \text{if } |\beta_i| \le \lambda t \\ \beta_i + \lambda t & \text{if } \beta_i < -\lambda t \end{cases}$

$\Rightarrow \text{prox}_{h,t}(\beta) = S_{\lambda t}(\beta)$. the soft thresholding operator ↗

proximal update $\beta_{k+1} = \text{prox}_{h,t}(\beta_k - t\nabla g(\beta_k)) = S_{\lambda t}(\beta_k + tX^T(y - X\beta_k))$.

$\nabla g(\beta_k) = -X^T(y - X\beta_k)$

Remark: it is called the iterative soft-thresholding algorithm (ISTA).

Subgradient: note $f(y) \geq f(x) + \nabla f(x)^T(y-x)$. how about non differentiable f?

d_x is a subgradient. denoted by $d_x \in \partial f(x)$, if $f(y) \geq f(x) + d_x^T(y-x)$.

$x_{k+1} = x_k - t \cdot G_t(x_k)$ $G_t(x_k)$ is gradient or subgradient of $f = g+h$? X

$w = prox_{h,t} = \underset{y}{\operatorname{argmin}} \frac{1}{2t}\|y-x\|^2 + h(y) \triangleq \tilde{h}(y)$. $0 \in \partial \tilde{h}(w)$.

$\frac{1}{t}(w-x) \in \partial(\frac{1}{2t}\|y-x\|^2)(w)$. conj. $\frac{1}{t}(x-w) \in \partial h(w)$.

note $\tilde{h}(y) \geq \tilde{h}(w) \iff h(y) \geq h(w) + \frac{1}{2t}\underbrace{(2x-w-y)^T(y-w)}_{\leq 2(x-w)^T(y-w)}$. >0

assume $h(z) = h(w) + \frac{1}{t}(x-w)^T(z-w) - \delta\|z-w\|$.

by convexity. $h(\gamma) \leq h(w) + \frac{1}{t}(x-w)^T(\gamma-w) - \delta\|\gamma-w\|$. $\gamma = \theta z + \bar{\theta} w$.

however. $\frac{1}{2t}\|\gamma-x\|^2 = \frac{1}{2t}\|w-x\|^2 + \frac{1}{t}(w-x)(\gamma-w) + \frac{1}{2t}\|\gamma-w\|^2$.

$\Rightarrow h(\gamma) + \frac{1}{2t}\|\gamma-x\|^2 < h(w) + \frac{1}{2t}\|w-x\|^2$ for sufficiently small $\|\gamma-w\|$.

contradiction!

In general. we have $\partial(f_1+f_2) = \partial f_1 + \partial f_2$.

Firmly nonexpansive: $(prox_h(x) - prox_h(y))^T(x-y) \geq \|prox_h(x) - prox_h(y)\|^2$

let $u = prox_h(x)$. $v = prox_h(y)$. then $x-u \in \partial h(u)$. $y-v \in \partial h(v)$.

$\Rightarrow h(v) \geq h(u) + (x-u)^T(v-u)$. and $h(u) \geq h(v) + (y-v)^T(u-v)$.

$\Rightarrow (x-u)^T(v-u) + (y-v)^T(u-v) \leq 0 \Rightarrow (x-u-y+v)^T(u-v) \geq 0$.

Nonexpansive: $\|\text{prox}_h(x)-\text{prox}_h(y)\| \le \|x-y\|$ by Cauchy-Schwarz.

Convergence analysis: assume g is L-smooth and m-strongly convex.

Recall $\frac{m}{2}\|x-y\|^2 \le g(y)-g(x)-\nabla g(x)^T(y-x) \le \frac{L}{2}\|x-y\|^2$. $y \leftarrow x_{k+1}$

$$\frac{m}{2}t^2\|G_t(x_k)\|^2 \le g(x_{k+1})-g(x_k)+t\nabla g(x_k)^T G_t(x_k) \le \frac{L}{2}t^2\|G_t(x_k)\|^2$$

$$\Rightarrow g(x_{k+1}) \le g(x_k) - t\nabla g(x_k)^T G_t(x_k) + \frac{t}{2}\|G_t(x_k)\|^2 \quad (\text{select } t \le 1/L).$$

$$\overset{\forall y}{\Rightarrow} f(x_{k+1}) \le g(y) - \nabla g(x_k)^T(y-x_k) - \frac{m}{2}\|x_k-y\|^2 + h(x_{k+1})$$

(note $x_{k+1} = x_k - tG_t(x_k) = \text{prox}_{h,t}(x_k - t\nabla g(x_k)) \Rightarrow G_t(x_k)-\nabla g(x_k) \in \partial h(x_{k+1})$)

$$\Rightarrow f(x_{k+1}) \le g(y)+h(y) - G_t(x_k)^T(y-x_{k+1}) - \frac{m}{2}\|x_k-y\|^2 + \nabla g(x_k)^T(x_k-x_{k+1})$$

$$f(y) \le g(y)+h(y) - G_t(x_k)^T(y-x_k) - \frac{m}{2}\|x_k-y\|^2 - \frac{t}{2}\|G_t(x_k)\|^2 \quad (= tG_t(x_k)).$$

let $y = x_k \Rightarrow f(x_{k+1}) \le f(x_k) - \frac{t}{2}\|G_t(x_k)\|^2$. (descent method).

let $y = x^* \Rightarrow f(x_{k+1}) - f(x^*) \le G_t(x_k)^T(x_k-x^*) - \frac{m}{2}\|x_k-x^*\|^2 - \frac{t}{2}\|G_t\|^2$

$$= G_t(x_k)^T\left(x_k - x^* - \frac{t}{2}G_t(x_k)\right) - \frac{m}{2}\|x_k-x^*\|^2.$$

$$= \frac{1}{2t}(x_k-x_{k+1})^T\left((x_k-x^*)+(x_{k+1}-x^*)\right) - \frac{m}{2}\|x_k-x^*\|^2.$$

$$\le \frac{1}{2t}\left(\|x_k-x^*\|^2 - \|x_{k+1}-x^*\|^2\right) - \frac{m}{2}\|x_k-x^*\|^2.$$

$$\Rightarrow \sum_{k=0}^{T}\left(f(x_{k+1})-f(x^*)\right) \le \frac{1}{2t}\|x_0-x^*\|^2 \Rightarrow f(x_T)-f(x^*) \le \frac{\|x_0-x^*\|^2}{2tT}.$$

Furthermore, $m>0 \Rightarrow \|x_{k+1}-x^*\|^2 \le (1-mt)\|x_k-x^*\|^2$. fix $t=1/L$

Require L-smooth. if L is unknown.? Backtracking line search.

choose $0<\alpha,\beta<1$. and initial $\hat{t}>0$. often choose $\alpha=1/2$.

Recall in analysis we need $g(y)-g(x)-\nabla g(x)^T(y-x)\le \frac{L}{2}\|x-y\|^2$

while $g(x_k-t_k G_{t_k}(x_k)) > g(x_k)-t\nabla g(x_k)^T G_{t_k}(x_k)+\frac{t_k}{2}\|G_{t_k}(x_k)\|^2$ $t_k=\beta t_k$

Convergence analysis for backtracking line search. note $t_k\ge \min\{\hat{t},\beta/L\}$

$f(x_{k+1})-f(x^*)\le \frac{1}{2t_k}\left((1-mt_k)\|x_k-x^*\|^2-\|x_{k+1}-x^*\|^2\right)$. $\overset{\triangle}{=} t_{min}$

$$\Rightarrow \sum_{k=0}^{T-1} t_k(f(x_{k+1})-f(x^*)) \le \frac{1}{2}\left(\|x_0-x^*\|^2-\|x_T-x^*\|^2\right)$$

$$\Rightarrow (\Sigma t_k)(f(x_T)-f(x^*)) \le \frac{1}{2}\|x_0-x^*\|^2 \Rightarrow f(x_T)-f(x^*)\le \frac{\|x_0-x^*\|^2}{2t_{min}T}$$

Again. if m-strongly convex for $m>0$. $\|x_{k+1}-x^*\|^2\le(1-mt_{min})\|x_k-x^*\|^2$.