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

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