Lecture 15. Newton's method, Proximal gradient descent.

Definition: $\|Q\|_{a,b} = \max_{x \neq 0} \frac{\|Qx\|_b}{\|x\|_a} = \max_{\|x\|_a = 1} \|Qx\|_b \quad Q \in \mathbb{R}^{m \times n}$.

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

Definition: $\nabla^2 f(x)$ is M-Lipschitz if $\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \leq M \|x - y\|_2$.

Theorem. If $f(x)$ is m-strongly convex. $\nabla^2 f(x)$ is M-Lipschitz. $\{x_k\}$ is the sequence produced by Newton's method (with step size $= 1$). then

$$\|x_{k+1} - x^*\| \leq \frac{M}{2m} \|x_k - x^*\|^2.$$

Remark. Let $y_k = \frac{M}{2m} \|x_k - x^*\|$. Then $y_{k+1} \leq y_k^2 \Rightarrow y_T \leq y_0^{2^T}$.

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

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

$\leq \|\nabla^2 f(x_k)^{-1}\|_2 \|\nabla^2 f(x_k)(x_k - x^*) - (g(1) - g(0))\|_2$.

(where $g(t) = \nabla f(x^* + t(x_k - x^*))$. $Dg(t) = \nabla^2 f(x^* + t(x_k - x^*))(x_k - x^*)$)

$= \|\nabla^2 f(x_k)^{-1}\|_2 \|Dg(1) - \int_0^1 Dg(t)\, dt\|_2 \leq \frac{1}{m} \|\int_0^1 (Dg(1) - Dg(t))\, dt\|_2$

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

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

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

Global convergence for Newton's method: backtracking line search.

while $f(x_k + t_k d_k) > f(x_k) + \alpha t_k \cdot \nabla f(x_k)^T d_k$. $t_k = \beta t_k$.

Now set $d_k = -(\nabla^2 f(x_k))^{-1} \nabla f(x_k)$ for Newton's method.

Summary of descent method: $x_{k+1} = x_k + t_k d_k$. if $\nabla f(x_k)^T d_k < 0$.

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

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

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

Convergence: guaranteed if $f(x_{k+1}) < f(x_k)$ (Lyapunov's stability).

Order and rate: suppose $\lim_{n \to \infty} \frac{\|x_{n+1} - x^*\|}{\|x_n - x^*\|^\rho} = \mu$

then ρ is the order of convergence, and μ is the rate.

Gradient descent: linear convergence (if m-strongly convex).

$$\|x_{k+1} - x^*\| \leq (1 - mt) \|x_k - x^*\|.$$

Newton's method: quadratic convergence.

$$\|x_{k+1} - x^*\| \leq \frac{M}{2m} \|x_k - x^*\|^2.$$

Condition: gradient L-smooth (convergence) m-strongly convex (rate).

Newton. $\nabla^2 f$ M-Lipschitz + m-strongly convex.

For gradient method or Newton's method, assume f differentiable

How to deal with nondifferentiable functions? Proximal gradient

Recall in gradient descent, $x_{k+1} = x_k - t\nabla f(x_k)$, we actually approximate f by

$\hat{f}(x) = f(x_k) + \nabla f(x_k)^T(x - x_k) + \frac{1}{2t}\|x - x_k\|^2$, and let $x_{k+1} = \underset{x}{\operatorname{argmin}}\ \hat{f}(x)$

Now assume $f(x) = g(x) + h(x)$, where $\begin{cases} g(x) \text{ is convex and differentiable} \\ h(x) \text{ is convex} \end{cases}$

Let $\hat{f}(x) = g(x_k) + \nabla g(x_k)^T(x - x_k) + \frac{1}{2t}\|x - x_k\|^2 + h(x) \approx f(x)$.

So $x_{k+1} = \underset{x}{\operatorname{argmin}}\ \hat{f}(x) = \underset{x}{\operatorname{argmin}}\ \frac{1}{2t}\|x - (x_k - t\nabla g(x_k))\|^2 + h(x)$.

since $\frac{1}{2t}\|x - (x_k - t\nabla g(x_k))\|^2 = \frac{1}{2t}\|x - x_k\|^2 + \nabla g(x_k)^T(x - x_k) + C$

$\frac{1}{2t}\|x - (x_k - t\nabla g(x_k))\|^2$: distance to the gradient descent update for g.

If we make $h(x)$ small, $f \approx g$, x_{k+1} is close to the update for g.

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

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

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

A key point: $\text{prox}_{h,t}(y)$ has closed forms for many important $h(x)$.

Lasso: least absolute shrinkage and selection operator.

Reduce overfitting: fitting data set (X, y) with small $\|\beta\|_0$ (# of nonzero entries)

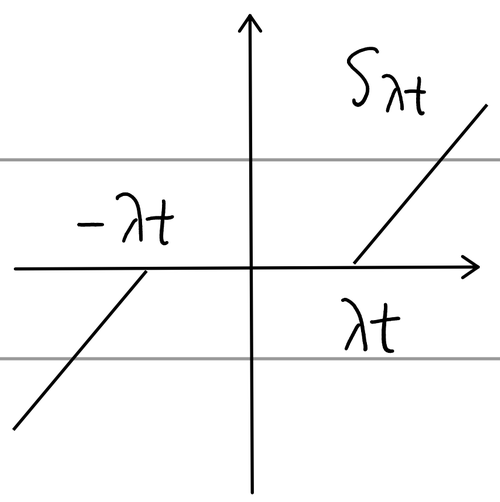
We usually use $\lambda \|\beta\|_p$ $(\lambda > 0,\ p \geq 1)$ to approximate $\|\beta\|_0$.

Now choose $p=1$. $f(\beta) = \frac{1}{2}\|y - X\beta\|^2 + \lambda\|\beta\|_1$ $h(\beta) = \lambda\|\beta\|_1$

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

$$\Rightarrow \text{prox}_{h,t}(\gamma) = \underset{\beta}{\text{argmin}} \sum_{i=1}^{n} \left(\frac{1}{2t}(\beta_i - \gamma_i)^2 + \lambda|\beta_i| \right) \quad \beta = (\beta_1 \cdots \beta_n),\ \gamma = (\gamma_1, \cdots, \gamma_n).$$

$$\Rightarrow [\text{prox}_{h,t}(\gamma)]_i = \underset{\beta_i}{\text{argmin}} \frac{1}{2t}(\beta_i - \gamma_i)^2 + \lambda|\beta_i| \qquad \lambda, t > 0.$$



$$= \underset{\beta_i}{\text{argmin}} (\beta_i - \gamma_i)^2 + 2t\lambda \cdot \begin{cases} \beta_i & \beta_i \geq 0 \\ -\beta_i & \beta_i \leq 0 \end{cases}$$

$$= \underset{\beta_i}{\text{argmin}} \beta_i^2 - 2\beta_i \cdot \begin{cases} \gamma_i - \lambda t & \beta_i \geq 0 \\ \gamma_i + \lambda t & \beta_i \leq 0 \end{cases}$$

$$= S_{\lambda t}(\gamma_i) = \begin{cases} \gamma_i - \lambda t & \gamma_i \geq \lambda t \\ 0 & |\gamma_i| \leq \lambda t \\ \gamma_i + \lambda t & \gamma_i \leq -\lambda t \end{cases}.$$

$\text{prox}_{h,t}(\gamma) = S_{\lambda t}(\gamma)$ the soft thresholding operator

proximal update: $\beta_{k+1} = \text{prox}_{h,t}(\beta_k - t\nabla g(\beta_k))$ $\quad g(\beta) = \frac{1}{2}\|y - X\beta\|^2$.

$$= S_{\lambda t}(\beta_k - t(-X^T(y - X\beta_k)))$$

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

Convergence of proximal gradient descent: assume g is L-smooth. $t \leq \frac{1}{L}$.

$$\hat{f}(x) = g(x_k) + \nabla g(x_k)^T(x - x_k) + \frac{1}{2t}\|x - x_k\|^2 + h(x) \geq g(x) + h(x) = f(x).$$

$x_{k+1} = \underset{x}{\text{argmin}}\ \hat{f}(x) \Rightarrow f(x_{k+1}) \leq \hat{f}(x_{k+1}) < \hat{f}(x_k) = f(x_k)$ if $x_{k+1} \neq x_k$.

hope $f(x_{k+1}) < f(x_k)$ if $f(x_k) \neq f^*$. so need $x_{k+1} = x_k$ only if $f(x_k) = f^*$

If $x_{k+1} = x_k$. $x_k = prox_{h,t}(x_k - t\nabla g(x_k))$. we hope $\forall x.\ f(x) \geq f(x_k)$.

Note that $\forall x.\ g(x) \geq g(x_k) + \nabla g(x_k)^T(x - x_k)$. so it suffices to show.

$\forall x.\ h(x) \geq h(x_k) - \nabla g(x_k)^T(x - x_k)$. $-\nabla g(x_k)$ looks like $\nabla h(x_k)$.

Definition (subgradient). Let $f: \mathbb{R}^n \to \mathbb{R}$. $x \in \mathbb{R}^n$. We say $v \in \mathbb{R}^n$ is a subgradient of f at x, denoted by $v \in \partial f(x)$ if $\forall y.\ f(y) \geq f(x) + v^T(y - x)$.

Remark. If f is convex. $\partial f(x) \neq \phi$. (supporting hyperplane of epi f).

$\nabla f(x) \in \partial f(x)$ if differentiable. subgradients may not be unique. ($|x|$).

Lemma. If $w = prox_{h,t}(y)$. then $\frac{1}{t}(y - w) \in \partial h(w)$.

Corollary: $x_k = prox_{h,t}(x_k - t\nabla g(x_k)) \Rightarrow -\nabla g(x_k) \in \partial h(x_k)$.

Proof. $w = prox_{h,t}(y) = \underset{x}{\operatorname{argmin}} \frac{1}{2t}\|x - y\|^2 + h(x)$. So $\forall x \in \mathbb{R}$.

$\frac{1}{2t}\|x - y\|^2 + h(x) \geq \frac{1}{2t}\|w - y\|^2 + h(w)$. Our goal is to show $\forall x$.

$h(x) \geq h(w) + \frac{1}{t}(y - w)^T(x - w)$. If $\exists h(z) < h(w) + \frac{1}{t}(y - w)^T(z - w)$

assume $h(z) = h(w) + \frac{1}{t}(y - w)^T(z - w) - \delta\|z - w\|$ for some $\delta > 0$.

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

however. $\|x - y\|^2 = \|w - y\|^2 + \|x - w\|^2 + 2(w - y)^T(x - w)$. so we have

$\frac{1}{2t}\|x - y\|^2 + h(x) < \frac{1}{2t}\|w - y\|^2 + h(w) + \|x - w\|^2 - \delta\|x - w\|$

contradicts if $\|x-w\|$ sufficiently small. □

Convergence rate for proximal gradient descent: the same as GD.

Suppose $g(x)$ is L-smooth. set $t \le 1/L$. then $f(x_T) - f^* \le \frac{\|x_0 - x^*\|^2}{2tT}$.

If $g(x)$ is further m-strongly convex. $\|x_T - x^*\|^2 \le (1-mt)^T \|x_0 - x^*\|^2$.

If L-smoothness is unknown. then use exact / backtracking line search.