Nonlinear Optimization in Machine Learning

Lecture 2 Line search methods

$$\min_{x \in \mathbb{R}^n} f(x)$$

$\{x^k\}$ $\quad f(x^{k+1}) < f(x^k) \quad k = 0, 1, 2, \dots$

"descent direction" $d \in \mathbb{R}^n$

$$f(x + td) < f(x) \text{ for all } \underbrace{t > 0 \text{ sufficiently}}_{\text{small}}$$

Say f is continuously differentiable

$$f(x + td) = f(x) + t \, \nabla f(x + \gamma t d)^T d \quad , \gamma \in (0, 1)$$

If $d^T \nabla f(x) < 0$ then d is a descent direction

"steepest descent"

$$\inf_{\|d\| = 1} d^T \nabla f(x) = - \|\nabla f(x)\|$$

achieved when $d = - \dfrac{\nabla f(x)}{\|\nabla f(x)\|}$

"Steepest descent method"

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k), \quad k = 0, 1, 2, \ldots$$

$\alpha_k > 0$ steplength

"fixed point iteration"

$$x^{k+1} = \Phi(x^k) \qquad k = 0, 1, 2, \ldots$$

$$\Phi(x) = x - \alpha \nabla f(x) \qquad \alpha > 0$$

$$\|\Phi(x) - \Phi(z)\| = \left\| (x-z) - \alpha(\nabla f(x) - \nabla f(z)) \right\|$$

$$= \left\| (x-z) - \alpha \int_0^1 \nabla^2 f(z + \gamma(x-z))(x-z)\, d\gamma \right\|$$

$$= \left\| \left(I - \alpha \int_0^1 \nabla^2 f(z + \gamma(x-z))\, d\gamma \right) (x-z) \right\|$$

$$\lambda\left[I - \alpha \int_0^1 \nabla^2 f(z + \gamma(x-z))\, d\gamma \right] \in [1 - \alpha L, 1 - \alpha m]$$

If x^* is s.t. $\nabla f(x^*) = 0$

and $\|\phi(x) - \phi(z)\| \le \beta \|x - z\|$ $\beta \in [0,1)$

then $\|x^{k+1} - x^*\| = \|\phi(x^k) - \phi(x^*)\|$

$$\le \beta \|x^k - x^*\|$$

$$\le \dots \le \beta^{k+1} \|x_0 - x^*\|$$

"linear convergence rate"

In order $\|x^k - x^*\| \le \varepsilon$

need $T \ge \dfrac{\log\left(\dfrac{\|x^0 - x^*\|}{\varepsilon}\right)}{|\log \beta|}$

We want $-\beta \le 1 - \alpha L \le 1 - \alpha m \le \beta$

So $\alpha \in \left[\dfrac{1-\beta}{m}, \dfrac{1+\beta}{L}\right]$

When $1 - \alpha L = 1 - \alpha m \Rightarrow \beta = \dfrac{L-m}{L+m}$, $\alpha = \dfrac{2}{L+m}$

"steepest descent method"

$$f(x+\alpha d) = f(x) + \alpha \nabla f(x)^T d + \alpha \int_0^1 [\nabla f(x+\alpha d) - \nabla f(x)]d \, d\gamma$$

$$\leq f(x) + \alpha \nabla f(x)^T d + \alpha \int_0^1 \|\nabla f(x+\alpha d) - \nabla f(x)\| \cdot \|d\| \, d\gamma$$

$$\leq f(x) + \alpha \nabla f(x)^T d + \alpha^2 \frac{L}{2} \|d\|^2$$

$x = x^k \qquad d = -\nabla f(x^k)$

take $\alpha = \frac{1}{L}$, $x^{k+1} = x^k - \frac{1}{L} \nabla f(x^k)$, $k = 0, 1, 2, \ldots$

$$f(x^{k+1}) = f\left(x^k - \frac{1}{L} \nabla f(x^k)\right)$$

$$\leq f(x^k) - \frac{1}{2L} \|\nabla f(x^k)\|^2 \quad (*)$$

We derive various convergence rate estimates from $(*)$

General Case

$f(x) \geq \bar{f}$ for all x

$$\sum_{k=0}^{T-1} \|\nabla f(x^k)\|^2 \leq 2L \cdot \left(\sum_{k=0}^{T-1} [f(x^k) - f(x^{k+1})] \right)$$

$$= 2L \cdot [f(x^0) - f(x^T)]$$

$$\leq 2L [f(x^0) - \bar{f}]$$

so $\lim\limits_{T \to \infty} \|\nabla f(x_T)\| = 0$

$$\min_{0 \leq k \leq T-1} \|\nabla f(x^k)\| \leq \sqrt{\frac{2L[f(x^0) - \bar{f}]}{T}}$$

- Convex Case

$$f(x^*) \geq f(x^k) + \nabla f(x^k)^T (x^* - x^k)$$

(Use (*)) $\geq f(x^{k+1}) + \frac{1}{2L}\|\nabla f(x^k)\|^2 + \nabla f(x^k)^T (x^* - x^k)$

so $f(x^{k+1}) \leq f(x^*) + (\nabla f(x^k))^T (x^k - x^*) - \frac{1}{2L}\|\nabla f(x^k)\|$

$$= f(x^*) + \frac{L}{2}\left(\|x^k - x^*\|^2 - \left\| x^k - x^* - \frac{1}{L}\nabla f(x^k) \right\|^2 \right)$$

$$= f(x^*) + \frac{L}{2}\left(\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2 \right)$$

So
$$\sum_{k=0}^{T-1}\left(f(x^{k+1})-f^*\right) \le \frac{L}{2}\sum_{k=0}^{T-1}\left(\|x^k-x^*\|^2-\|x^{k+1}-x^*\|^2\right)$$
$$\le \frac{L}{2}\|x^0-x^*\|^2$$

as $f(x^k)\downarrow$ we get
$$f(x^T)-f(x^*) \le \frac{1}{T}\sum_{k=0}^{T-1}\left(f(x^{k+1})-f^*\right) \le \frac{L}{2T}\|x^0-x^*\|^2$$

- Strongly convex case

$$f(x)+(\nabla f(x))^T(z-x)+\frac{L}{2}\|z-x\|^2 \ge f(z) \ge f(x)+(\nabla f(x))^T(z-x)+\frac{m}{2}\|z-x\|$$

"sandwiching"

$$f_\mu(x) = f(x)+\mu\|x\|^2 \quad \text{convex} \to \text{strong conve[x]}$$

Look at
$$f(z) \ge f(x)+(\nabla f(x))^T(z-x)+\frac{m}{2}\|z-x\|^2$$
$$\min_z f(z) \ge \min_z\left(f(x)+(\nabla f(x))^T(z-x)+\frac{m}{2}\|z-x\|^2\right)$$
$$\left(\text{at } z = x-\frac{\nabla f(x)}{m}\right)$$

$$\Rightarrow f(x^*) \ge f(x)-\frac{1}{2m}\|\nabla f(x)\|^2$$

i.e. $\|\nabla f(x)\|^2 \geq 2m\,[f(x) - f(x^*)]$

"linear convergence rate"

$$f(x^{k+1}) = f\left(x^k - \frac{1}{L}\nabla f(x^k)\right)$$

$$\leq f(x^k) - \frac{1}{2L}\|\nabla f(x^k)\|^2$$

$$\leq f(x^k) - \frac{m}{L}\left(f(x^k) - f^*\right)$$

i.e. $f(x^{k+1}) - f(x^*) \leq \left(1 - \frac{m}{L}\right)\left(f(x^k) - f(x^*)\right)$

$\Rightarrow \quad f(x^T) - f^* \leq \left(1 - \frac{m}{L}\right)\left(f(x^0) - f^*\right)$

Iterate k s.t. $0 \leq k \leq T$ $\quad \|\nabla f(x^k)\| \leq \varepsilon$

General $\qquad T \geq \dfrac{2L(f(x^0) - f^*)}{\varepsilon^2}$ sublinear

Weakly convex $\qquad T \leq \dfrac{f(x^0) - f^*}{\varepsilon}$ sublinear

Strongly convex $\qquad k \geq \dfrac{L}{m}\log\left(\dfrac{f(x^0) - f^*}{\varepsilon}\right)$ linear

— · — · — · — · — · — · —

Descent Methods with variable stepsize & search direction

$$x^{k+1} = x^k + \alpha_k d^k \qquad k = 0, 1, 2, \dots$$

$\alpha_k > 0$ $\qquad d^k$ — search direction

$$-(d^k)^T \nabla f(x^k) \geq \bar{\varepsilon} \|\nabla f(x^k)\| \cdot \|d^k\|$$

$$\gamma_1 \|\nabla f(x^k)\| \leq \|d^k\| \leq \gamma_2 \|\nabla f(x^k)\|$$

where $\bar{\varepsilon}, \gamma_1, \gamma_2 > 0$

If $d^k = -\nabla f(x^k)$ then $\bar{\varepsilon} = \gamma_1 = \gamma_2 = 1$

$$f(x^{k+1}) = f(x^k + \alpha d^k)$$

$$= f(x^k) + \alpha \nabla f(x^k)^T d^k + \alpha \int_0^1 [\nabla f(x^k + \gamma \alpha d^k) - \nabla f(x^k)]^T d^k \, d\gamma$$

$$\leq f(x^k) + \alpha (\nabla f(x^k))^T d^k + \alpha \int_0^1 \|\nabla f(x^k + \gamma \alpha d^k) - \nabla f(x^k)\| \cdot \|d^k\| \, d\gamma$$

$$\leq f(x^k) - \alpha \bar{\varepsilon} \|\nabla f(x^k)\| \cdot \|d^k\| + \frac{\alpha^2}{2} L \|d^k\|^2$$

$$\leq f(x^k) - \alpha\left(\bar{\varepsilon} - \alpha \frac{L}{2} \gamma_2\right) \|\nabla f(x^k)\| \cdot \|d^k\|$$

If $\alpha \in \left(0, \frac{2\bar{\varepsilon}}{L\gamma_2}\right)$ then $f(x^{k+1}) < f(x^k)$

① $d^k = -S^k \nabla f(x^k)$ S^k symmetric positive definite

$\lambda(S^k) \in [\gamma_1, \gamma_2]$

② Gauss-Southwell $d^k = -[\nabla f(x^k)]_{i_k}$

$$i_k = \arg\min_{i=1,2,\dots n} \left|[\nabla f(x^k)]_i\right|$$

③ Stochastic Coordinate Descent

$d^k = -[\nabla f(x^k)]_{i_k}$ $i_k \in \{1, 2, \dots n\}$ chosen uniformly

④ Stochastic Gradient methods

⑤ Exact line search $\min_{\alpha > 0} f(x^k + \alpha d^k)$

⑥ Approximate Line search

"Weak Wolfe Conditions"

$f(x^k + \alpha d^k) \leq f(x^k) + c_1 \alpha \nabla f(x^k)^T d^k$ "sufficient decrease"

$\nabla f(x^k + \alpha d^k)^T d^k \geq c_2 \nabla f(x^k)^T d^k$

⑦ Backtracking line search

$$\bar{\alpha},\ \beta\bar{\alpha},\ \beta^2\bar{\alpha},\ \beta^3\bar{\alpha},\ \dots$$

How to obtain convergence?

$$f(x^{k+1}) \le f(x^k) - C\|\nabla f(x^k)\|^2$$

for some $C>0$

Approximate Second-Order Necessary Points

$$\|\nabla f(x)\| \le \varepsilon_g, \quad \lambda_{min}(\nabla^2 f(x)) \ge -\varepsilon_H \qquad (**)$$

(i) If $\|\nabla f(x^k)\| > \varepsilon_g$ take the steepest descent

$$x^{k+1} = x^k - \frac{1}{L}\nabla f(x^k)$$

(ii). Otherwise let λ_k be the minimum eigenvalue of $\nabla^2 f(x^k)$. If $\lambda_k < -\varepsilon_H$ choose p^k to be the eigenvector of the most negative eigenvalue of $\nabla^2 f(x^k)$

$\|p^k\| = 1$, $(p^k)^T \nabla f(x^k) \le 0$

$$\text{Set} \quad x^{k+1} = x^k + \alpha_k p^k, \quad \alpha_k = \frac{2|\lambda_k|}{M}$$

(iii) If both (i) & (ii) are false then (**) is true

For (i)

$$f(x^{k+1}) \le f(x^k) - \frac{1}{2L}\|\nabla f(x^k)\|^2$$
$$\le f(x^k) - \frac{\varepsilon_g^2}{2L}$$

For (ii)

$$f(x^{k+1}) \le f(x^k) + \alpha_k \nabla f(x^k)^T p^k + \frac{1}{2}\alpha_k^2 (p^k)^T \nabla^2 f(x^k) p^k$$
$$+ \frac{1}{6} M \alpha_k^3 \|p^k\|^3$$
$$\le f(x^k) - \frac{1}{2}\left(\frac{2|\lambda_k|}{M}\right)^2 |\lambda_k| + \frac{1}{6} M \left(\frac{2|\lambda_k|}{M}\right)^3$$
$$= f(x^k) - \frac{2}{3}\frac{|\lambda_k|^3}{M^2} = f(x^k) - \frac{2}{3}\frac{\varepsilon_H^3}{M^2}$$

Here $\|\nabla^2 f(x) - \nabla^2 f(y)| \le M \|x-y\|$.

$$f(x+p) \le f(x) + \nabla f(x)^T p + \frac{1}{2} p^T \nabla^2 f(x) p + \frac{1}{6} M \|p\|$$