Jan 30, 2020

$$\min_x f(x) \qquad f: \mathbb{R}^n \to \mathbb{R}, \ f \text{ smooth}$$

Iterative: $x_{k+1} = G(x_k)$

Idea: $x_{k+1} \approx x_k + \alpha_k p_k$ (step size α_k, search direction p_k)

Taylor's theorem: $f_k = f(x_k)$, $g_k = \nabla f(x_k)$

$$f(x_k + \varepsilon p_k) = f(x_k) + \varepsilon p_k^T \nabla f(x_k) + O(\varepsilon^2)$$

if $p_k^T g_k < 0$, then $f(x_k + \varepsilon p_k) < f(x_k)$ for small ε

$p_k = -g_k \quad p_k^T g_k = -g_k^T g_k = -\|g_k\|_2^2 < 0$

Gradient descent: $x_{k+1} = x_k - \alpha_k g_k$

$f(x) = \frac{1}{2} x^T A x + b^T x + c$, A SPD $\quad g(x) = Ax + b$

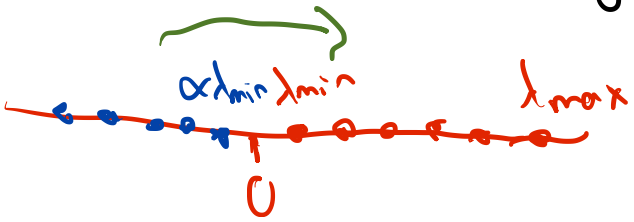
$x_{k+1} = x_k - \alpha_k g_k \qquad x_{k+1} = x_k - \alpha(Ax_k + b)$

at opt. $\qquad x_* = x_* - \alpha(Ax_* + b)$

$$\overbrace{x_{k+1} - x^*}^{e_{k+1}} = \overbrace{x_k - x_*}^{e_k} - \alpha A \overbrace{(x_k - x_*)}^{e_k}$$

$e_{k+1} = (I - \alpha A) e_k \qquad \|e_{k+1}\|_2 \le \|I - \alpha A\|_2 \|e_k\|_2$

$\|I - \alpha A\|_2 = \max_j |1 - \alpha \lambda_j| = \max(|1 - \alpha\lambda_{min}|, |1 - \alpha\lambda_{max}|)$



$\alpha\lambda_{max} - 1 = 1 - \alpha\lambda_{min}$

$\alpha = \frac{2}{\lambda_{min} + \lambda_{max}}$

$\|e_{k+1}\|_2 \le \|I - \alpha A\|_2 \|e_k\|_2 \qquad 1 - \frac{2\lambda_{min}}{\lambda_{min} + \lambda_{max}}$

$\frac{\lambda_{min}}{\lambda_{max}}$ small $\Rightarrow$ slow convergence

What about a different search direction?

$$p_k = -M^{-1} g_k \qquad M \text{ SPD} \qquad x_{k+1} = x_k - \alpha_k M^{-1} g_k$$

$p_k^T g_k < 0 \qquad p_k^T g_k = -\underbrace{g_k M^{-1} g_k}_{>0} < 0$

$$e_{k+1} = (I - \alpha M^{-1} A) e_k \qquad \alpha = 1 \qquad M = A$$

converge in one step: $x_1 = x_0 - A^{-1} g_0$

$g_0 = A x_0 + b \Rightarrow x_1 = x_0 - A^{-1}(A x_0 + b) = A^{-1} b$

Problem: only a model

Solution: model true locally

Taylor expansion:

$$f(x+p) = f(x) + p^T g(x) + \frac{1}{2} p^T H(x) p + O(\|p\|^3)$$

$$\min_p f(x+p) \approx c + p^T b + \frac{1}{2} p^T A p$$

Taylor expansion:

$$f(x+p) \approx f(x) + p^T g(x) + \frac{1}{2} p^T H(x) p + O(\|p\|^3)$$

$$\min_p f(x+p) \approx c + p^T b + \frac{1}{2} p^T A p$$

solution: $Ap = -b$ $\quad H(x)p = -g(x)$

$$x_{k+1} = x_k - H(x_k)^{-1} g_k \quad \text{Newton iteration}$$

Possible issue: $H_k = H(x_k)$ not SPD Possible solution: $H(x_k) + \lambda I$

Possible problem: unit step length might be too far

H_k SPD $\Rightarrow$ $H_k^{-1} g_k$ is a descent direction

$$x_{k+1} = x_k - \alpha_k H_k^{-1} g_k \quad \text{(line search)}$$

Problems with Newton:

- $O(n^2)$ to store
- $O(n^3)$ to solve $H_k p_k = g_k$ (in general)

Idea: approx H_k

- just compute diagonal (diagonal scaling)
- Quasi-Newton $H_k \approx B_k$

Taylor expansion:

$$g_{k+1} = g(x_k + (x_{k+1} - x_k)) \approx g_k + H_k(x_{k+1} - x_k)$$

$$\overbrace{H_k(x_{k+1} - x_k)}^{s_k} \approx \overbrace{g_{k+1} - g_k}^{y_k}$$

$$B_k(x_{k+1} - x_k) = g_{k+1} - g_k$$

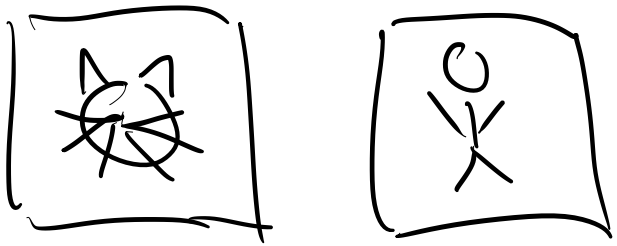
BFGS: $$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k}$$

Stochastic gradient descent

Idea: replace $g(x)$ with $\tilde{g}(x)$ where $\mathbb{E}(\tilde{g}(x)) = g(x)$

Example: $f(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x)$ $\quad g(x) = \frac{1}{N}\sum_{i=1}^{N} g_i(x)$

$\frac{1}{N}\|Ax-b\|_2^2 = \frac{1}{N}\sum_{i=1}^{N}(a_i^T x - b)^2$ $\quad \frac{1}{N}\|A - BC\|_F = \frac{1}{N}\sum_{ij}(b_{ij} - b_i^T c_j)^2$



$\frac{1}{N}\sum_{i=1}^{N} \ell(\text{score}_i, \text{image}_i)$

- sample $J \sim \{1, 2, \dots, N\}$ i.i.d (UAR)

$$\mathbb{E}(g_J(x)) = \sum_{i=1}^{N} g_i(x)\Pr(J=i) = \frac{1}{N}\sum_{i=1}^{N} g_i(x) = g(x)$$

unbiased estimate

SGD: $x_{k+1} = x_k - \alpha_k g_{J_k}$ (random var; iterate)

$J \sim \{1, \ldots, N\}$ UAR $\quad x_{k+1} = x_k - \alpha_k g_{Jk}$

$\mathbb{E}(g_{Jk}) = g_k$

Possible problem: variance might be large

Partial solution: "mini-batching"

Sample $J \subseteq \{1, 2, \ldots, N\}$ UAR $\quad |J| = 10, 50, 100$

$g_{Jk} = \frac{1}{|J|} \sum_{j \in J} g_j(x_k) \quad \mathbb{E}(g_{Jk}) = g(x_k)$ still unbiased!

Possible problem: random sub-samples might be hard to get

Partial solution: approx by "pass" over data

for $i = 1, 2, \ldots N$

$J = i$

$x_{k+1} = x_k - \alpha_k g_{Jk}$

} epoch