Tutorial For A5

Thursday, October 23, 2014   2:54 PM

Q1 "Just make a nicer plot"

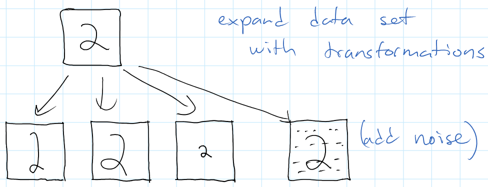
Q2 Should be easy to obtain 0.25 – 0.15

1. more hidden units
2. play with step size, momentum
3. - creating more training examples may help the most

I = reshape (X(i,:)[16, 16])
imagesc(I)

Image Toolbox:

imresize
imrotate
imtranslate



Q3 $\mu I \preccurlyeq \nabla^2 f(x)$ $\nabla^2_{ii} f(x) \le L$

an "ordering" on the matrices

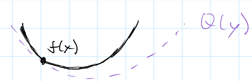
← diagonals less than L

$y^T \nabla^2 f(x) y \geqslant y^T(\mu I) y$

$y^T \nabla^2 f(x) y \geqslant \mu y^T y$

$$f(y) = \underbrace{f(x) + \nabla f(x)^T (y-x) + \tfrac{1}{2}(y-x)^T \nabla^2 f(z)(y-x)}_{Qy}$$

there exists some z
that this holds true



at x, $Q(y) = f(x)$

set y to (y–x)

$(y-x)^T \nabla^2 f(z)(y-x) \geqslant \mu (y-x)^T(y-x)$

$$f(y) \geqslant f(x) + \nabla f(x)^T (y-x) + \overbrace{\tfrac{\mu}{2}(y-x)^T(y-x)}^{\frac{\mu}{2}\|y-x\|^2}$$

$\{a^T y - c + \frac{\mu}{2} y^T y - \mu y^T x + c\}$

$a = b + c$
if $c \geqslant d$
$\therefore a \geqslant b + d$

take derivative, set to zero

minimize wrt y:

$\nabla Q(y) = 0 + \nabla f(x) + \mu(y-x)$

solve: $y^* = x - \left(\frac{1}{\mu}\right)\nabla f(x)$

$$f(y^*) \geqslant f(x) + \underbrace{\nabla f(x)^T ((x - \tfrac{1}{\mu}\nabla f(x)) - x)}_{-1/\mu\, \nabla f(x)^T \nabla f(x) \;\Rightarrow\; -\frac{1}{\mu}\|f(y)\|^2} + \frac{\mu}{2}\|(x - \tfrac{1}{\mu}\nabla f(x)) - x\|^2$$

$f(x^*) \geqslant f(x) - \frac{1}{\mu}\|\nabla f(x)\|^2 + \frac{1}{2\mu}\|\nabla f(x)\|^2$

$f(x^*) \geqslant f(x) - \frac{1}{2\mu}\|f(x)\|^2$

$f(x^*) - f(x) \geqslant -\frac{1}{2\mu}\|\nabla f(x)\|^2$

$-2\mu[f(x) - f(x^*)] \geqslant \|\nabla f(x)\|^2$

$f(y) \geqslant g(y)$

$f(y^*) \geqslant g(y_*)$

minimizes f   minimizes g

$g(y^*) \geqslant g(y_*)$

don't need proof

$y = x^{t+1} = x^t - \frac{1}{L} \nabla_{i_t} f(x^t) e_{i_t}$

pick coordinat i_t to update

$e_i = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \leftarrow i$

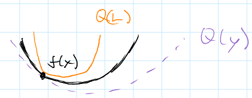
$$= x^t - \begin{array}{c} \\ \\ \\ i \rightarrow \\ \\ \end{array} \begin{bmatrix} 0 \\ 0 \\ 0 \\ \frac{1}{L}\nabla_{i_t} f(x^t) \\ 0 \\ 0 \end{bmatrix}$$

$y = x$, except for coordinate i — assume only differ in one coordinate

$$f(y) = f(x) + \nabla f(x)^T (y-x) + \tfrac{1}{2}(y-x)^T \nabla f(z)(y-x)$$

$$f(y) = f(x) + \underbrace{\nabla_i f(x)(y-x)_i}_{\text{all multiplied by 0 except at coordinate } i} + \tfrac{1}{2}(y-x)_i^2 \nabla_{ii} f(z) \qquad \nabla_{ii}^2 f(x) \le L$$

$$f(y) \le \underbrace{f(x) + \nabla_i f(y)(y-x)_i + \tfrac{L}{2}(y-x)_i^2}_{Q_L}$$



set $x = x^t$
$y = x^{t+1}$

$$f(x^{t+1}) \le f(x^t) + \nabla_i f(x^t)(x^{t+1} - x^t)_i + \tfrac{L}{2}(x^{t+1} - x^t)_i^2$$

$$y = x^{t+1} = x^t - \tfrac{1}{L}\nabla_{i_t} f(x^t) e_{i_t}$$

$$(x^{t+1} + x^t) = -\tfrac{1}{L}\nabla_{i_t} f(x^t) e_{i_t}$$

$$f(x^{t+1}) \le f(x^t) + \nabla_i f(x^t)\left(-\tfrac{1}{L}\nabla_{i_t} f(x^t) e_{i_t}\right) + \tfrac{L}{2}\left\|\tfrac{1}{L}\nabla_{i_t} f(x^t) e_{i_t}\right\|^2$$

$$\Rightarrow f(x^t) - \tfrac{1}{L}(\nabla_i f(x^t))^2 + \tfrac{1}{2L}(\nabla_i f(x^t))^2$$

$$\Rightarrow f(x^t) - \tfrac{1}{2L}(\nabla_i f(x^t))^2$$

$\Rightarrow$ this is for all methods (all 3 parts of question)

Coordinate Descent:

upper bound: $f(x^{t+1}) \le f(x^t) - \frac{1}{2L}(\nabla_i f(x^t))^2$

lower bound: $f(x^*) \ge f(x^t) - \frac{1}{2\mu}\|\nabla f(x^t)\|^2$

thing inside is infinity norm

$i_t = \underset{i}{\operatorname{argmax}} |\nabla_i f(x^t)|$ — biggest improvement by changing

Greedy:

$$(\nabla_{i_t} f(x^t))^2 = \|\nabla f(x^t)\|_\infty^2$$

$$\ge d\|\nabla f(x)\|^2$$

Randomize i:

$$f(x^{t+1}) \le f(x^t) - \tfrac{1}{2L}(\nabla_{i_k} f(x^t))^2$$

$$E_{i_t}[f(x^{t+1})] \le E_{i_t}[f(x^t) - \tfrac{1}{2L}(\nabla_{i_t} f(x^t))^2]$$

$$\Rightarrow E_{i_t}[f(x^t)] - E_{i_t}[\tfrac{1}{2L}(\nabla_{i_t} f(x^t))^2]$$

$$\Rightarrow f(x^t) - \tfrac{1}{2L}\sum_{i=1}^{d} p(i_t = i)(\nabla_{i_t} f(x^t))^2$$

$$\Rightarrow f(x^t) - \tfrac{1}{2L}\sum_{i=1}^{d}\left(\tfrac{1}{d}\right)(\nabla_{i_t} f(x^t))$$

$$\Rightarrow f(x^t) - \tfrac{1}{2Ld}\underbrace{\sum_{i=1}^{d}(\nabla_i f(x_t))^2}_{\|\nabla f(x)\|^2}$$

$E(a+b) = E(a) + E(b)$

$E(c(a+b)) = cE(a) + cE(b)$

"linearity of expectation"

Review from class: Gradient Descent:

upper bound $f(x^{t+1}) \le f(x^t) - \frac{1}{2L}\|\nabla f(x^t)\|^2$

lower bound: $f(x^*) \ge f(x^t) - \frac{1}{2\mu}\|\nabla f(x^t)\|^2$

(lower bound is the same)

$$-\tfrac{1}{2\mu}\|\nabla f(x^t)\|^2 \le f(x^*) - f(x^t)$$

$$-\|\nabla f(x^t)\|^2 \le -2\mu[f(x^t) - f(x^*)]$$

$$f(x^{t+1}) \le f(x^t) - \tfrac{2\mu}{2L}[f(x^t) - f(x^*)]$$

subtract $f(x^*)$ from both sides $\left(\begin{array}{c} a \ge b \\ a - c \ge b - c \end{array}\right)$

$$f(x^{t+1}) - f(x^*) \le f(x^t) - f(x^*) - \tfrac{\mu}{L}(f(x^t) - f(x^*))$$

$$\Rightarrow \left(1 - \tfrac{\mu}{L}\right)[f(x^t) - f(x^*)]$$

$p(i_t = i) = \frac{1}{d}$

$$E_{i_t}[g(i)] = \sum_{i=1}^{d} p(i_t = i)\, g(i_t)$$

$$E_{i_t}[f(x^t)] \le f(x^t) - \frac{1}{2Ld} \|\nabla f(x)\|^2$$

combine with lower bound as before, then we're done

Lipschitz $\nabla_{ii}^2 f(x) \le L_i$

$$p(i_t = i) = \frac{L_i}{\sum_j L_j}$$

$x^{t+1} = x^t - \frac{1}{L_i} \nabla f(x^t) e_{i_t}$ $\quad$ L_{i_t} in upper bound $\quad$ $f(x^{t+1}) \le f(x^t) - \frac{1}{2L_{i_t}} (\nabla_{i_t} f(x^t))^2$

$1 - \frac{\mu}{dL} \Rightarrow 1 - \frac{\mu}{\sum L_i}$ $\quad$ $\sum L_i \le d \max\{L_i\}$, $\sum L_i \le dL$

Q4

final result:

$$\|x^{t+1} - x^*\| \le \rho \|x^t - x^*\|$$

at x^t, $\rho < 1$, we are moving closer to solution

$$\|x^{t+1} - x^*\|^2 = \|x^t - x^*\|^2 - 2\alpha \nabla f(x^t)^T \underbrace{(x^t - x^*)}_{\substack{y = x^* \\ x = x^t}} + \alpha^2 \|\nabla f(x^t)\|^2$$

$$-(\nabla f(x) - \nabla f(y))^T (x - y) \le -\frac{\mu L}{\mu + L} \|x - y\|^2 - \frac{1}{\mu + L} \|\nabla f(x) - \nabla f(y)\|^2$$

$(\nabla f(x^t) - \nabla f(x^*))$

$\nabla f(x^*) = 0$ at optimal

$$= \|x^t - x^*\|^2 - 2a\left[\frac{\mu L}{\mu + L}\|x^t - x^*\|^2 - \frac{1}{\mu + L}\|\nabla f(x^t) - \underbrace{\nabla f(x^*)}_{\text{zero}}\|^2\right] + \alpha^2 \|\nabla f(x_t)\|^2$$

$$= \left(1 - \frac{2\alpha\mu L}{\mu + L}\right)\|x^t - x^*\|^2 + \cdots \qquad - \alpha \|\nabla f(x)\|^2 \left(\frac{2}{\mu + L} - \alpha\right)$$

goes to zero

$\Downarrow$ 0 $\qquad \alpha = \frac{2}{\mu + L}$

4.2

step 1. $x^{t+1} = \text{prox}\left[x^t - \alpha_t \nabla f(x^t)\right]$

$x^* = \text{prox}\left[x^* - \alpha_t \nabla f(x^*)\right]$

step 2. $\|\text{prox}(x) - \text{prox}(y)\| \le \|x - y\|$

continue as in 4.1,
will give exact same as 4.1 in the end.

4.3

$$\log(\epsilon) = -\log(1/\epsilon)$$