Problem Set 1

1. Gradient Descent

1. $$\arg\min_{w} f(w^{(t)}) + \langle w - w^{(t)}, \nabla f(w^{(t)}) \rangle + \frac{\lambda}{2} \|w - w^{(t)}\|^2$$

unconstrainted QP has closed form solution

$$F'(w) = \nabla f(w^{(t)}) + \lambda (w - w^{(t)}) = 0$$

$$w^* = w^{(t)} - \frac{1}{\lambda} \nabla f(w^{(t)}) \qquad \eta = \frac{1}{\lambda}$$

It has similar update rule as gradient descent rule. Therefore, GD can be thought of as find the min of the approximation function plus proximity regularization. When $\lambda \downarrow$, it means less penality on proximity term, and first order approximation is ~~well~~ good enough. As a result, the step along gradient direction can be large ($\eta \uparrow$)

2. Define $D(w^t, w^*) = (w^t - w^*)^T (w^t - w^*)$

$$\begin{aligned}
\Delta D &= D(w^{t+1}, w^*) - D(w^t, w^*) \\
&= D(w^t - \eta V_t, w^*) - D(w^t, w^*) \\
&= (w^t - w^* - \eta V_t)^T (w^t - w^* - \eta V_t) - (w^t - w^*)^T (w^t - w^*) \\
&= \eta^2 \|V_t\|^2 - 2\eta \langle w^t - w^*, V_t \rangle
\end{aligned}$$

$$\begin{aligned}
\sum_{t=1}^{T} \Delta D &= D(w^T, w^*) - D(w^1, w^*) \\
&= \sum_{t=1}^{T} \eta^2 \|V_t\|^2 - 2\eta \sum_{t=1}^{T} \langle w^t - w^*, V_t \rangle
\end{aligned}$$

$$\begin{aligned}
\therefore 2\eta \sum_{t=1}^{T} \langle w^t - w^*, V_t \rangle &= D(w^1 - w^*) - D(w^T - w^*) + \eta^2 \sum_{t=1}^{T} \|V_t\|^2 \\
&\le (0 - w^*)^T (0 - w^*) + \eta^2 \sum_{t=1}^{T} \|V_t\|^2 \qquad (w^{(1)} = 0,\ D \ge 0)
\end{aligned}$$

$$\therefore \sum_{t=1}^{T} \langle w^t - w^*, V_t \rangle \le \frac{\|w^*\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|V_t\|^2$$

3. Assuming f is convex, then $\nabla^2 f$ is positive-semi definite

and $$\begin{aligned}
f(\bar{w}) - f(w^*) &= f\left(\frac{1}{T} \sum_{t=1}^{T} w^{(t)}\right) - f(w^*) \\
&\le \frac{1}{T} \sum_{t=1}^{T} [f(w^t) - f(w^*)]
\end{aligned}$$

$$f(w^*) = f(w^{(t)}) + (w^* - w^{(t)})\nabla f(w^{(t)}) + \underbrace{\text{high order}}_{\geq 0}$$

$$\geq f(w^{(t)}) + (w^* - w^{(t)})\nabla f(w^{(t)})$$

$$\therefore f(\bar{w}) - f(w^*) \leq \frac{1}{T}\sum_{t=1}^{T} \langle w^{(t)} - w^*, \nabla f(w^{(t)}) \rangle$$

$$f(\bar{w}) - f(w^*) \leq \frac{1}{T}\left[\frac{\|w^*\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T} \|\nabla f(w^{(t)})\|\right]$$

$$\leq \frac{1}{T}\left[\frac{B}{2\eta} + \frac{\eta}{2}T\rho\right)$$

$$= \frac{B}{2\eta T} + \frac{\eta}{2}\rho = \frac{B}{2\frac{B}{\rho}\sqrt{T}} + \frac{B}{2\sqrt{T}}$$

$$= \frac{B+\rho}{2}\cdot\frac{1}{\sqrt{T}} \propto \frac{1}{\sqrt{T}}$$

4. No. The SGD choose one term out of objective func randomly.

$$f(w) = \frac{1}{2}(w-2)^2 + \frac{1}{2}(w+1)^2$$

$$= w^2 - w + \frac{5}{2} = (w - \tfrac{1}{2})^2 + \frac{9}{4}$$

when $w_t = 0$ it shout move toward ~~w~~ w^* to improve overall loss function.

But if $\frac{1}{2}(w+1)^2$ is selected, then $w_t - \eta \cdot c w_t = 0$

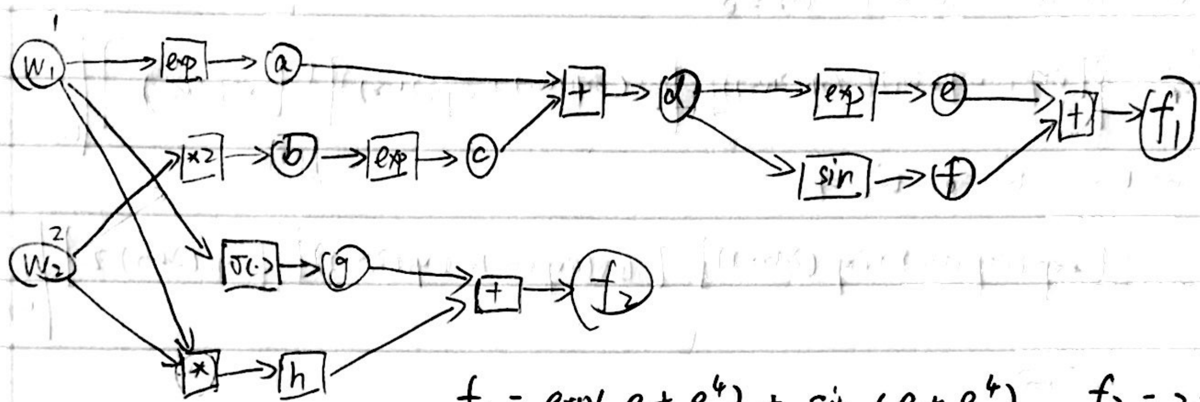
it will deviate from the course toward $w^* = \frac{1}{2}$

2. Auto Differentiation

$$f_1(w_1, w_2) = e^{e^{w_1} + e^{2w_2}} + \sin(e^{w_1} + e^{2w_2})$$

$$f_2(w_1, w_2) = w_1 w_2 + \sigma(w_1)$$

(a)



$f_1 = \exp(e + e^4) + \sin(e + e^4)$ $\quad f_2 = 2 + \frac{1}{1+e^{-1}}$

$= 7.8021e24$ $\quad = 2.7311$

(b)

$$\text{Jacobian} = \begin{bmatrix} \frac{\partial f_1}{\partial w_1} & \frac{\partial f_1}{\partial w_2} \\ \frac{\partial f_2}{\partial w_1} & \frac{\partial f_2}{\partial w_2} \end{bmatrix}$$

$$\frac{\partial f_1}{\partial w_1} = \frac{f_1(w_1+\Delta w_1, w_2)}{\Delta w_1} = \frac{\exp(e^{1.01}+e^4)+\sin(e^{1.01}+e^4)}{0.01} = 8.0182e26$$

$$\frac{\partial f_1}{\partial w_2} = \frac{f_1(w_1, w_2+\Delta w_2)}{\Delta w_2} = \frac{\exp(e+e^{4.02})+\sin(e+e^{4.02})}{0.01} = 2.3508e27$$

$$\frac{\partial f_2}{\partial w_1} = \frac{f_2(w_1+\Delta w_1, w_2)}{\Delta w_1} = \frac{1.01\times 2 + \sigma(1.01)}{0.01} = 275.3020$$

$$\frac{\partial f_2}{\partial w_2} = \frac{f_2(w_1, w_2+\Delta w_2)}{\Delta w_2} = \frac{1\times 2.01 + \sigma(1)}{0.01} = 274.1059$$

(c) $\frac{\partial f_1}{\partial a_0}$ $\quad a_0 = w_1$ or w_2

$$\dot{w}_1 = \frac{\partial w_1}{\partial a_0} \qquad \dot{w}_2 = \frac{\partial w_2}{\partial a_0}$$

$$\frac{\partial a}{\partial a_0} = \exp(w_1)\,\dot{w}_1 \qquad \frac{\partial b}{\partial a_0} = 2\dot{w}_2 \qquad \frac{\partial c}{\partial a_0} = \exp(2w_2)\cdot 2\dot{w}_2$$

$$\frac{\partial d}{\partial a_0} = \exp(w_1)\dot{w}_1 + \exp(2w_2)\cdot 2\dot{w}_2$$

$$\frac{\partial e}{\partial a_0} = \exp(\exp(w_1)+\exp(2w_2))\left[\exp(w_1)\dot{w}_1 + \exp(2w_2)\,2\dot{w}_2\right]$$

$$\frac{\partial f}{\partial a_0} = \cos(\exp(w_1)+\exp(2w_2))\left[\exp(w_1)\dot{w}_1 + \exp(2w_2)\,2\dot{w}_2\right]$$

$$\frac{\partial f_1}{\partial a_0} = \left\{\left[\exp(\exp(w_1)+\exp(2w_2))\right] + \left[\cos(\exp(w_1)+\exp(2w_2))\right]\right\}\left[\exp(w_1)\dot{w}_1 + \exp(2w_2)\cdot 2\dot{w}_2\right]$$

when $a_0 = w_1$ $\dot{w}_1 = 1$ $\dot{w}_2 = 0$

$$\frac{\partial f_1}{\partial w_1} = \left\{ [\exp(\exp(w_1) + \exp(2w_2))] + [\cos(\exp(w_1) + \exp(2w_2))] \right\} [\exp(w_1)] \Big|_{w_1 = 1, w_2 = 2}$$

$a_0 = w_2$ $\dot{w}_1 = 0$ $\dot{w}_2 = 1$

$$\frac{\partial f_1}{\partial w_2} = \left\{ [\exp(\exp(w_1) + \exp(2w_2))] + [\cos(\exp(w_1) + \exp(2w_2))] \right\} [\exp(2w_2) \cdot 2] \Big|_{w_1 = 1, w_2 = 2}$$

~~when $a_0 = w_1$~~

$$\frac{\partial g}{\partial a_0} = \sigma(\cdot)(1 - \sigma(\cdot))\, \dot{w}_1 \qquad \partial f_2 = \sigma(\cdot)(1 - \sigma(\cdot))\, \dot{w}_1 + w_2 \dot{w}_1 + w_1 \dot{w}_2$$

$$\frac{\partial h}{\partial a_0} = w_2 \dot{w}_1 + w_1 \dot{w}_2$$

$a_0 = w_1$ $\dot{w}_1 = 1$ $\dot{w}_2 = 0$

$$\frac{\partial f_2}{\partial w_1} = \sigma(w_1)(1 - \sigma(w_1)) + w_2$$

$a_0 = w_2$ $\dot{w}_1 = 0$ $\dot{w}_2 = 1$

$$\frac{\partial f_2}{\partial w_2} = w_1$$

d. $\frac{\partial f_1}{\partial f_1} = 1$ $\quad \frac{\partial f_1}{\partial e} = 1$ $\quad \frac{\partial f_1}{\partial f} = 1$ $\quad \frac{\partial f_1}{\partial d} = \frac{\partial f_1}{\partial e} \exp(d) + \frac{\partial f_1}{\partial f} \cos(d)$

$$= \exp(d) + \cos(d)$$

$$\frac{\partial f_1}{\partial a} = \frac{\partial f_1}{\partial d} \qquad \frac{\partial f_1}{\partial c} = \frac{\partial f_1}{\partial d} \qquad \frac{\partial f_1}{\partial b} = \frac{\partial f_1}{\partial d} \exp(b) = \exp(\exp(w_1) + \exp(2w_2)) + \cos(\exp(w_1) + \exp(2w_2))$$

$$\frac{\partial f_1}{\partial w_1} = \frac{\partial f_1}{\partial d} \exp(w_1) = [\exp(\exp(w_1) + \exp(2w_2)) + \cos(\exp(w_1) + \exp(2w_2))] \exp(w_1)$$

$$\cancel{\frac{\partial f_1}{\partial w_2} = 2 \frac{\partial f_1}{\partial f}} \quad \frac{\partial f_1}{\partial w_2} = \frac{\partial f_1}{\partial b} \cdot 2 = [\exp(\exp(w_1) + \exp(2w_2)) + \cos(\exp(w_1) + \exp(2w_2))]\, 2 \cdot \exp(2w_2)$$

$$\frac{\partial f_2}{\partial f_2} = 1 \quad \frac{\partial f_2}{\partial g} = 1 \quad \frac{\partial f_2}{\partial h} = 1 \quad \frac{\partial f_2}{\partial w_2} = \frac{\partial f_2}{\partial h} \cdot w_1 = w_1$$

$$\frac{\partial f_2}{\partial w_1} = \frac{\partial f_2}{\partial g} \sigma \cdot (1 - \sigma) + \frac{\partial f_2}{\partial h} w_2 = \sigma(w_1)(1 - \sigma(w_1)) + w_2$$

e. Yes.

Scanned by CamScanner