Date ___/___/_____

Problem #1 → Given $Q \in R^{N \times N}$, $P \in R^{M \times M}$, and $B \in R^{M \times N}$

Need to prove $(Q^{-1} + B^T P^{-1} B)^{-1} B^T P^{-1} = Q B^T (B Q B^T + P)^{-1}$

Proof → Lets Multiply both sides by $(Q^{-1} + B^T P^{-1} B)$

$\Rightarrow$ $(Q^{-1} + B^T P^{-1} B) \cdot (Q^{-1} + B^T P^{-1} B)^{-1} B^T P^{-1} = (Q^{-1} + B^T P^{-1} B) \cdot Q B^T (B Q B^T + P)^{-1}$

cancels out

$\Rightarrow$ $B^T P^{-1} = (Q^{-1} + B^T P^{-1} B) \cdot Q B^T (B Q B^T + P)^{-1}$

Now, Lets Multiply both sides by $(B Q B^T + P)$

$\Rightarrow$ $B^T P^{-1} \cdot (B Q B^T + P) = (Q^{-1} + B^T P^{-1} B) \cdot Q B^T \underbrace{(B Q B^T + P)^{-1} \cdot (B Q B^T + P)}$

cancels out

$\Rightarrow$ $B^T P^{-1} \cdot (B Q B^T + P) = (Q^{-1} + B^T P^{-1} B) \cdot Q B^T$

$\Rightarrow$ $B^T P^{-1} B Q B^T + \underbrace{B^T P^{-1} P}_{\text{cancels out}} = \underbrace{Q^{-1} Q}_{\text{cancels out}} B^T + B^T P^{-1} B Q B^T$

$\Rightarrow$ $B^T P^{-1} B Q B^T + B^T = B^T P^{-1} B Q B^T + B^T$

$\Rightarrow$ L.H.S = R.H.S, Hence proved

* This equation is mentioned as a lemma in the paper Kalman filtering by Max Welling.

2nd part → Proof of Woodbury identity –

Need to show $(A + BD^{-1}C)^{-1} = A^{-1} - A^{-1}B(D + CA^{-1}B)^{-1}CA^{-1}$

Proof → Lets Multiply both sides by $(A + BD^{-1}C)$

$\Rightarrow$ $(A + BD^{-1}C) \cdot (A + BD^{-1}C)^{-1} = (A + BD^{-1}C)\left[A^{-1} - A^{-1}B(D + CA^{-1}B)^{-1}CA^{-1}\right]$

cancels out

$\Rightarrow$ $I = \{\underbrace{AA^{-1} - AA^{-1}}_{\text{cancels out}}B(D + CA^{-1}B)^{-1}CA^{-1}\} + \{BD^{-1}CA^{-1} + BD^{-1}CA^{-1}B(D + CA^{-1}B)^{-1}CA^{-1}\}$

$\Rightarrow$ $I = \{I - B(D + CA^{-1}B)^{-1}CA^{-1}\} + \{\ \text{''}\ \}$

Date __/__/____

Saathi

$\Rightarrow I = I - B(D+CA^{-1}B)^{-1}CA^{-1} + BD^{-1}CA^{-1} + BD^{-1}CA^{-1}B(D+CA^{-1}B)^{-1}CA^{-1}$

Now, taking common $(D+CA^{-1}B)^{-1}CA^{-1}$ from both the above terms

$\Rightarrow I = I - BD^{-1}CA^{-1} + [(D+CA^{-1}B)^{-1}CA^{-1}.(B+BD^{-1}CA^{-1}B)]$

Next, taking common BD^{-1} from above 2 terms

$\Rightarrow I = I - BD^{-1}CA^{-1} + [(D+CA^{-1}B)^{-1}CA^{-1} \cdot BD^{-1}(D+CA^{-1}B)]$

Now, cancelling out $(D+CA^{-1}B)^{-1}$ and $(D+CA^{-1}B)$, we get →

$I = I - BD^{-1}CA^{-1} + [CA^{-1}.BD^{-1}]$

$\Rightarrow I = I$ (L.H.S = R.H.S), Hence proved

Problem #2 → Given $x=[x_1 : x_2 : x_3] \in R^3$ and $y=[y_1 : y_2] \in R^2$

$y_1 = x_1^2 - x_2$, $y_2 = x_3^2 + 3x_2$

Solution → $\frac{dy_1}{\partial x_1} = \frac{\partial(x_1^2 - x_2)}{\partial x_1} = 2x_1$

$\frac{\partial y_1}{\partial x_2} = \frac{\partial(x_1^2 - x_2)}{\partial x_2} = -1$

$\frac{\partial y_1}{\partial x_3} = \frac{\partial(x_1^2 - x_2)}{\partial x_3} = 0$

$\frac{\partial y_2}{\partial x_1} = \frac{\partial(x_3^2 + 3x_2)}{\partial x_1} = 0$

$\frac{\partial y_2}{\partial x_2} = \frac{\partial(x_3^2 + 3x_2)}{\partial x_2} = 3$

$\frac{\partial y_2}{\partial x_3} = \frac{\partial(x_3^2 + 3x_2)}{\partial x_3} = 2x_3$

$\Rightarrow \frac{\partial y}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial x_2} \\ \frac{\partial y_1}{\partial x_2} & \frac{\partial y_2}{\partial x_2} \\ \frac{\partial y_1}{\partial x_3} & \frac{\partial y_2}{\partial x_3} \end{bmatrix} = \begin{bmatrix} 2x_1 & 0 \\ -1 & 3 \\ 0 & 2x_3 \end{bmatrix}$

2nd part $x = r\sin\theta\cos\phi$, $y = r\sin\theta\sin\phi$, $z = r\cos\theta$

here $r>0$, $0<\theta<\pi$ and $0 \le \phi < 2\pi$. $x=[x;y;z]$

and $y=[r;\theta;\phi]$

Saathi

Date ___/___/_____

Solution → $\frac{\partial x}{\partial r} = \frac{\partial(r\sin\theta\cos\phi)}{\partial r} = \sin\theta\cos\phi$

$\frac{\partial x}{\partial \theta} = \frac{\partial(r\sin\theta\cos\phi)}{\partial \theta} = r\cos\theta\cos\phi$

$\frac{\partial x}{\partial \phi} = \frac{\partial(r\sin\theta\cos\phi)}{\partial \phi} = -r\sin\theta\sin\phi$

$\frac{\partial y}{\partial r} = \frac{\partial(r\sin\theta\sin\phi)}{\partial r} = \sin\theta\sin\phi$

$\frac{\partial y}{\partial \theta} = \frac{\partial(r\sin\theta\sin\phi)}{\partial \theta} = r\cos\theta\sin\phi$

$\frac{\partial y}{\partial \phi} = \frac{\partial(r\sin\theta\sin\phi)}{\partial \phi} = r\sin\theta\cos\phi$

$\frac{\partial z}{\partial r} = \frac{\partial(r\cos\theta)}{\partial r} = \cos\theta$

$\frac{\partial z}{\partial \theta} = \frac{\partial(r\cos\theta)}{\partial \theta} = -r\sin\theta$

$\frac{\partial z}{\partial \phi} = \frac{\partial(r\cos\theta)}{\partial \phi} = 0$

$$\Rightarrow \frac{\partial x}{\partial y} = \begin{bmatrix} \frac{\partial x}{\partial r} & \frac{\partial y}{\partial r} & \frac{\partial z}{\partial r} \\ \frac{\partial x}{\partial \theta} & \frac{\partial y}{\partial \theta} & \frac{\partial z}{\partial \theta} \\ \frac{\partial x}{\partial \phi} & \frac{\partial y}{\partial \phi} & \frac{\partial z}{\partial \phi} \end{bmatrix} = \begin{bmatrix} \sin\theta\cos\phi & \sin\theta\sin\phi & \cos\theta \\ r\cos\theta\cos\phi & r\cos\theta\sin\phi & -r\sin\theta \\ -r\sin\theta\sin\phi & r\sin\theta\cos\phi & 0 \end{bmatrix}$$

Problem#3- Newton's Method to solve least squares loss in Linear regression

Solution → 1. The Hessian of the least squares loss function

$$L(w) = \frac{1}{2}\sum_{i=1}^{n}(x_i^T w - y_i)^2 \quad \text{w.r.t parameter vector } w$$

To find the Hessian, we will have to compute 2nd derivative

$\nabla^2 L(w)$ w.r.t w

Date ___/___/_____ Saathi

So firstly, the 1st derivative (gradient) of $L(W)$ =

$$\nabla L(w) = \frac{\partial L(W)}{\partial W} = \frac{\partial \left(\frac{1}{2} * \sum_{i=1}^{n} x_i^T w - y_i\right)^2}{\partial W} = \sum_{i=1}^{n} (x_i^T w - y_i) * x_i$$

Now, Hessian (2nd derivative) of $L(W)$ w.r.t W =

$$\text{Hessian } (H) = \nabla^2 L(W) = \frac{\partial^2 L(W)}{\partial W^2} = \frac{\partial^2 \left(\sum_{i=1}^{n} (x_i^T w - y_i) x_i\right)}{\partial W^2}$$

$$\boxed{= \sum_{i=1}^{n} x_i^T x_i}$$

So, the Hessian is the summation of outer products of the input data points x_i.

2. Show that first iteration of Newton's method gives us:

$$w^* = (X X^T)^{-1} X y$$

Solution → We know that $L(w) = \frac{1}{2} \sum_{i=1}^{n} (x_i^T w - y_i)^2$

$$\text{Gradient of } L(w) = \sum_{i=1}^{n} (x_i^T w - y_i) x_i = X^T X w - X^T y$$

$$\text{Hessian of } L(w) = \sum_{i=1}^{n} x_i^T x_i = X^T X$$

Now, Newton's method sets the parameters to an initial guess w_0, then iteratively them. Let w_t be the current parameters on iteration t. Then, the updated parameter w_{t+1} as Newton's method gives:

$$w_{t+1} = w_t - H_L(w_t)^{-1} \cdot \nabla L(w_t) \quad \left(\begin{matrix}\text{Product of Hessian inverse} \\ \text{and gradient}\end{matrix}\right)$$

$$= w_t - (X^T X)^{-1} \cdot (X^T X w_t - X^T y)$$

$$= w_t - (X^T X)^{-1} \cdot (X^T X) w_t + (X^T X)^{-1} X^T y$$

$$\Rightarrow \boxed{w_{t+1} = (X^T X)^{-1} X^T y}$$, hence, Newton's method converges in first iteration.

5. Need to prove the following theorem:-

Suppose the function $f: R^n \to R$ is convex & differentiable & that is gradient ∇f is Lipschitz continous with constant $L > 0$, then if we run gradient descent for k iterations with a fixed Learning rate $0 < \alpha \leq \frac{1}{L}$, it will satisfies :-

$$f(x^k) - f(x^*) \leq \frac{1}{2\alpha k} \|x^0 - x^*\|_2^2$$ (where $f(x)$ is the optimal value)

We will prove the given hints step by step in order to prove this theorm finally.

1) Hints 1: $f(x^{k+1}) \leq f(x^k) - \left(1 - \frac{1}{2} L\alpha\right) \cdot \alpha \|\nabla f(x^k)\|_2^2$

Proof :- Given Lipschitz continuity of the gradient (∇f), we know that:- $\|f(y) - f(x)\|_2 \leq L \|y - x\|_2$

Now, applying this gradient update equation of gradient descent:-

~~$f(x^{k+1}) \leq f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{L}{2}$~~

$$f(x^{k+1}) \leq f(x^k) + \nabla f(x^k)^T (x^{k+1} - x^k) + \frac{L}{2} \|x^{k+1} - x^k\|_2^2$$

Using gradient descent update rule:-

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

putting this above value in our equation :-

$$f(x^{k+1}) \leq f(x^k) - \alpha \nabla f(x^k)^T \nabla f(x^k) + \frac{L\alpha^2}{2} \|\nabla f(x^k)\|_2^2$$

$$\Rightarrow f(x^{k+1}) \leq f(x^k) - \alpha \|\nabla f(x^k)\|_2^2 + \frac{L\alpha^2}{2} \|\nabla f(x^k)\|_2^2$$

$$\Rightarrow \boxed{f(x^{k+1}) \leq f(x^k) - \left(1 - \frac{1}{2} L\alpha\right) \alpha \|\nabla f(x^k)\|_2^2}$$

Now, since $\alpha \leq \frac{1}{L}$, we have $1 - \frac{1}{2} L\alpha \geqslant \frac{1}{2}$

Hence, inequality will be -

$$\boxed{f(x^{k+1}) \leq f(x^k) - \frac{\alpha}{2} \|\nabla f(x^k)\|_2^2}$$

Date

~~2) Hints 2 : $f(x^{k+1}) \leq f(x^*) - \frac{\alpha}{2}$~~

Hints 2 : $f(x^{k+1}) - f(x^*) \leq \frac{1}{2\alpha}\left(\|x^k - x^*\|_2^2 - \|x^k - \alpha \nabla f(x^k) - x^*\|_2^2\right)$

Now, using Hints 1 inequality we got ;-

$$f(x^{k+1}) \leq f(x^k) - \frac{\alpha}{2}\|\nabla f(x^k)\|_2^2$$

$$\Rightarrow f(x^{k+1}) - f(x^*) \leq f(x^k) - f(x^*) - \frac{\alpha}{2}\|\nabla f(x^k)\|_2^2$$

Putting Lipschitz property i.e. $\boxed{f(x^k) - f(x^*) \leq \nabla f(x^k)^T (x^k - x^*)}$

$$\Rightarrow f(x^{k+1}) - f(x^*) \leq \nabla f(x^k)^T (x^k - x^*) - \frac{\alpha}{2}\|\nabla f(x^k)\|_2^2$$

$$\Rightarrow f(x^{k+1}) - f(x^*) \leq \frac{1}{2\alpha}\|x^k - x^*\|_2^2 - \left[\left(\|\nabla f(x^k)\|_2^2\right) - \nabla f(x^k)^T (x^k - x^*)\right]$$

By applying gradient descent update rule, we get -

$$\Rightarrow \boxed{f(x^{k+1}) - f(x^*) \leq \frac{1}{2\alpha}\|x^k - x^*\|_2^2 - \|\nabla f(x^k) - (x^k - x^*)\|_2^2}$$

Hints 3 :- $\sum_{k=1}^{K}\left[f(x^k) - f(x^*)\right] \leq \frac{1}{2\alpha}\|x_0 - x^*\|^2$

↳ Summation of Hints 2 formulae/inequality from $k=1$ to $k \to$

$$\sum_{k=1}^{K}\left(f(x^{k+1}) - f(x^*)\right) \leq \sum_{k=1}^{K}\left(\frac{1}{2\alpha}\|x^k - x^*\|_2^2 - \|\nabla f(x^k) - (x^k - x^*)\|_2^2\right)$$

$$\Rightarrow \sum_{k=1}^{K} f(x^{k+1}) =$$

$$\leq \frac{1}{2\alpha}\sum_{k=1}^{K}\|x^k - x^*\|_2^2 - \sum_{k=1}^{K}\|\nabla f(x^k) - (x^k - x^*)\|_2^2$$

Since, $\sum_{k=1}^{K}\left(f(x^{k+1}) - f(x^*)\right) = f(x^{k+1}) - f(x^*)$

On applying property, f is decreasing on every iteration

$f(x^{k+1}) \leq f(x^k) \leq \ldots\ldots \leq f(x^2) \leq f(x^1)$

Page No.

Date ___ / ___ / _____

Re-evaluating L·H·S as $K(f(x') - f(x^*))$

$$\Rightarrow K(f(x') - f(x^*)) \leq \frac{1}{2\alpha} \sum_{K=1}^{K} \|x^K - x^*\|_2^2 - \sum_{K=1}^{K} \|\nabla f(x^K - (x^K - x^*)\|_2^2$$

Dividing both sides by K

$$\Rightarrow f(x') - f(x^*) \leq \frac{1}{2\alpha K} \sum_{K=1}^{K} \|x^K - x^*\|_2^2 \sum_{K=1}^{K} \|\nabla f(x^K) - (x^K - x^*)\|_2^2$$

Taking Limit as $K \to \infty$, $\frac{1}{K} \sum_{K=1}^{K} \|x^K - x^*\|_2^2$

approaches $\frac{1}{K} \|x^0 - x^*\|_2^2$

and $\frac{1}{K} \sum_{K=1}^{K} \|\nabla f(x^K - x^*)\|_2^2$ approaches 0.

As K gets larger, $\|\nabla f(x^K) - (x^K - x^*)\|$ become smaller, since gradient descent is converging.

Hence, in limit $\boxed{f(x') - f(x^*) \leq \frac{1}{2\alpha} \|x^0 - x^*\|_2^2}$

Therefore, we can say here -

$$\sum_{K=1}^{K} f(x^{K+1}) - f(x^*) \leq \frac{1}{2\alpha} \|x^0 - x^*\|_2^2$$

$$\Rightarrow \boxed{f(x^K) - f(x^*) \leq \frac{1}{2\alpha K} \cdot \|x^0 - x^*\|_2^2} \quad \text{Theorem proved.}$$

Thus, this completes the proof and depicts that gradient descent with the given conditions converges at a specified rate.

Date ___/___/_____

4. Given, ① Minimizing ordinal least square loss subject to Lp norm- Minimize $L(w) = \sum f(x_n, w) - t_n)^2$; $\|w\|_p^p \leq \gamma$

② Minimizing regularized least square loss with Lp regularization
Minimize $L(w) = \sum (f(x_n, w) - t_n)^2 + \lambda \|w\|_p^p$

Starting with OLS equation, applying Langrang multiplier α:

$L_1(w, \alpha) = \sum (f(x_n, w) - t_n)^2 + \alpha (\|w\|_p^p - \gamma)$

on taking partial derivative of L_1 w.r.t. w and α

$\partial L1/\partial w = 2 \sum f(x_n, w) - t_n) \, df/\partial w - \alpha p \|w\|_p^{p-1} w/\|w\|p = 0$

$\partial L1/\partial \alpha = \|w\|_p - \gamma = 0$

$\Rightarrow \|w\|_p = \gamma$, putting this in above equation -

$2 \sum f(x_n, w) - t_n) \frac{df}{\partial w} - \alpha p \gamma^{(p-1)} w/\gamma = 0$

or $2 \sum f(x_n, w) - t_n) \frac{df}{\partial w} - \alpha p \|w\|_p^{p-1} w/\|w\|_p = 0$

Next, applying Lagrangian in Regularized Least square with Langrange Multiplier β -

$L_2(w, \beta) = \sum (f(x_n, w) - t_n)^2 + \beta \|w\|_p^{p-1}$

on Taking partial derivative of L_2 w.r.t w and β

$\partial L2/\partial w = 2 \sum (f(x_n, w) - t_n) \partial f/\partial w + \beta (p-1) \|w\|_p^{p-2} w = 0$

$\partial L2/\partial \beta = \|w\|_p^{p-1} = 0$

↳ we can write $\|w\|_p^p = 0$, as p is constant

This means, we get $\|w\|_p^p = \gamma$ and also $\|w\|_p^p = 0$

$\Rightarrow \gamma = 0$ (under given conditions)

Hence → $\min_w L(w) = \sum_{n=1}^{N} (f(x_n; w) - t_n)^2$, s.t. $\|w\|_p^p \leq \gamma$

$\Leftrightarrow \min_w L(w) \sum_{n=1}^{N} (f(x_n; w) - t_n)^2 + \lambda \|w\|_p^p$

Date __/__/____

Saathi

In summary, minimizing OLS with Lp norm is equivalent to minimizing Lp regularised least square loss when $\gamma = 0$, where λ controls the strength of regularization.

Part 2 Relationship between λ and γ.

In the regularised least square loss problem, the hyperparameter λ controls the strength of regularization. The larger the value of λ, the stronger the regularization and more will be penalizing of weight vector w.

On the other hand, Lp norm constraint, γ sets an upper bound on the Lp norm of the weight vector w. If γ is small, it means Lp norm constraint is strict, and w must be close to the origin. If γ is large, it allows w to have a larger Lp norm.

The relationship between λ and γ depends on the particular optimization problem and value of other parameters, generally λ controls trade-off between fitting the data and having smaller Lp norm w.

The accurate relationship would require other details of dataset, model, optimization problem etc. In general, these hyperparameters are tuned by cross-validation to find best balance between fitting and regularization for a given task.