Training/ Testing Errors and Regularization

Dr. Chandranath Adak

Dept. of CSE, Indian Institute of Technology Patna

September 15, 2025

Bias vs Variance Tradeoff

- Bias $(\hat{f}(x)) = E\left[\hat{f}(x)\right] - f(x)$
- Variance $(\hat{f}(x)) = E\left[\left(\hat{f}(x) - E\left[\hat{f}(x)\right]\right)^2\right]$
- Training error $= E\left[(y - \hat{f}(x))^2\right]$
- We observe a tradeoff between bias and variance
  - Simple model: high bias, low variance
  - Complex model: low bias, high variance
- Both bias and variance contribute to the mean square error

$$E\left[(y - \hat{f}(x))^2\right] = \text{Bias}^2 + \text{Variance} + \sigma^2 \text{ (irreducible error)}$$

Please try to prove!

Some background

- $f(x)$ be the true function which we want to estimate
- $y_i = f(x_i) + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$
  Introducing a small noise
- Training sample data set consists of $\{(x_i, y_i)\}_{i=1}^{n}$
- $\hat{f}(x)$ be the approximated function obtained after training the model

What we are interested in?

To compute $\underbrace{E\left[(\hat{f}(x) - f(x))^2\right]}_{\text{True error}}$

Training Error vs Testing Error

Relationship between True Error and Test Error

- We want to compute $\underbrace{E\left[(\hat{f}(x) - f(x))^2\right]}_{\text{True error}}$
- Can we compute?

Relationship between True Error and Test Error

- We want to compute $\underbrace{E\left[(\hat{f}(x) - f(x))^2\right]}_{\text{True error}}$
- Can we compute?
  - No, we do not know what $f(x)$ is
  - We can estimate the error from the test samples

$$E\left[(\hat{f}(x_i) - y_i))^2\right] = E\left[(\hat{f}(x_i) - f(x_i) - \epsilon_i)^2\right] \quad \textbf{since } y_i = f(x_i) + \epsilon_i$$

$$
\begin{aligned}
E\left[(\hat{f}(x_i) - y_i))^2\right] &= E\left[(\hat{f}(x_i) - f(x_i) - \epsilon_i)^2\right] \quad \textbf{since } y_i = f(x_i) + \epsilon_i \\
&= E\left[(\hat{f}(x_i) - f(x_i))^2 - 2\ \epsilon_i\ (\hat{f}(x_i) - f(x_i)) + \epsilon_i^2\right] \\
&= E\left[(\hat{f}(x_i) - f(x_i))^2\right] - E\left[2\ \epsilon_i\ (\hat{f}(x_i) - f(x_i))\right] + E\left[\epsilon_i^2\right]
\end{aligned}
$$

$$
E\left[(\hat{f}(x_i) - f(x_i))^2\right] = E\left[(\hat{f}(x_i) - y_i))^2\right] + E\left[2\ \epsilon_i\ (\hat{f}(x_i) - f(x_i))\right] - E\left[\epsilon_i^2\right]
$$

$$E\left[(\hat{f}(x_i) - y_i))^2\right] =$$

$$E\left[(\hat{f}(x_i) - y_i))^2\right] = \frac{1}{m}\sum_{i=1}^{m}(\hat{f}(x_i) - y_i)^2$$

$$E\left[(\hat{f}(x_i) - f(x_i))^2\right] = E\left[(\hat{f}(x_i) - y_i))^2\right] - E\left[\epsilon_i^2\right] + E\left[2\ \epsilon_i\ (\hat{f}(x_i) - f(x_i))\right]$$

$$= \frac{1}{m}\sum_{i=1}^{m}(\hat{f}(x_i) - y_i)^2 - \frac{1}{m}\sum_{i=1}^{m}\epsilon_i^2 + E\left[2\ \epsilon_i\ (\hat{f}(x_i) - f(x_i))\right]$$

$$\underbrace{E\left[(\hat{f}(x_i) - f(x_i))^2\right]}_{\text{True error}} = \underbrace{\frac{1}{m}\sum_{i=1}^{m}(\hat{f}(x_i) - y_i)^2}_{\substack{\text{Estimated error} \\ \text{on test data}}} - \underbrace{\frac{1}{m}\sum_{i=1}^{m}\epsilon_i^2}_{\substack{\text{A small} \\ \text{constant}}} + \underbrace{2\ E\left[\epsilon_i\ (\hat{f}(x_i) - f(x_i))\right]}_{?}$$

$$E\left[\epsilon_i \ (\hat{f}(x_i) - f(x_i))\right] = Cov\left(\epsilon_i, (\hat{f}(x_i) - f(x_i))\right)$$

$$E\left[\epsilon_i \ (\hat{f}(x_i) - f(x_i))\right] = Cov\left(\epsilon_i, (\hat{f}(x_i) - f(x_i))\right)$$

Consider $X = \epsilon_i$; $\quad Y = (\hat{f}(x_i) - f(x_i))$

$$\begin{aligned}
Cov(X, Y) &= E\left[(X - \mu_X)(Y - \mu_Y)\right] \\
&= E\left[X(Y - \mu_Y)\right] \quad (\mu_X = E(\epsilon_i) = 0) \\
&= E\left[XY - X\mu_Y\right] \\
&= E\left[XY\right] - E\left[X\mu_Y\right] \\
&= E\left[XY\right] - \mu_Y \ E\left[X\right] \\
&= E\left[XY\right]
\end{aligned}$$

$$E\left[\epsilon_i \ (\hat{f}(x_i) - f(x_i))\right] = Cov\left(\epsilon_i, (\hat{f}(x_i) - f(x_i))\right)$$

- What is the relationship between ϵ_i and $(\hat{f}(x_i) - f(x_i))$?

$$E\left[\epsilon_i \ (\hat{f}(x_i) - f(x_i))\right] = Cov\left(\epsilon_i, (\hat{f}(x_i) - f(x_i))\right)$$

- What is the relationship between ϵ_i and $(\hat{f}(x_i) - f(x_i))$?
  - $y_i = f(x_i) + \epsilon_i$; where ϵ_i is a noise, independent of $f(x_i)$
  - Is there any dependency between ϵ_i and $\hat{f}(x_i)$?
    - No
    - Here ϵ_i is the noise at testing data
    - $\hat{f}(x_i)$ is estimated on training data

$$\therefore \epsilon_i \perp (\hat{f}(x_i) - f(x_i))$$

$$\begin{aligned} E\left[\epsilon_i \ . \ (\hat{f}(x_i) - f(x_i))\right] &= E\left[\epsilon_i\right] \ . \ E\left[(\hat{f}(x_i) - f(x_i))\right] \\ &= 0 \ . \ E\left[(\hat{f}(x_i) - f(x_i))\right] \\ &= 0 \end{aligned}$$

$$\underbrace{E\left[(\hat{f}(x_i) - f(x_i))^2\right]}_{\text{True error}} = \underbrace{\frac{1}{m}\sum_{i=1}^{m}(\hat{f}(x_i) - y_i)^2}_{\substack{\text{Estimated error} \\ \text{on test data}}} - \underbrace{\frac{1}{m}\sum_{i=1}^{m}\epsilon_i^2}_{\substack{\text{A small} \\ \text{constant}}}$$

The true error can be estimated from empirical test error

Relationship between True Error and Training Error

$$\underbrace{E\left[(\hat{f}(x_i) - f(x_i))^2\right]}_{\text{True error}} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\hat{f}(x_i) - y_i)^2}_{\substack{\text{Estimated error} \\ \text{on training data}}} - \underbrace{\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2}_{\substack{\text{A small} \\ \text{constant}}} + \underbrace{2\ E\left[\epsilon_i\ (\hat{f}(x_i) - f(x_i))\right]}_{=0\ ?}$$

$$E\left[\epsilon_i\ (\hat{f}(x_i) - f(x_i))\right] = Cov\left(\epsilon_i, (\hat{f}(x_i) - f(x_i))\right)$$

- What is the relationship between ϵ_i and $(\hat{f}(x_i) - f(x_i))$?

Relationship between True Error and Training Error

$$\underbrace{E\left[(\hat{f}(x_i) - f(x_i))^2\right]}_{\text{True error}} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\hat{f}(x_i) - y_i)^2}_{\substack{\text{Estimated error} \\ \text{on training data}}} - \underbrace{\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2}_{\substack{\text{A small} \\ \text{constant}}} + \underbrace{2\ E\left[\epsilon_i\ (\hat{f}(x_i) - f(x_i))\right]}_{=0\ ?}$$

$$E\left[\epsilon_i\ (\hat{f}(x_i) - f(x_i))\right] = Cov\left(\epsilon_i, (\hat{f}(x_i) - f(x_i))\right)$$

- What is the relationship between ϵ_i and $(\hat{f}(x_i) - f(x_i))$?
  - Is there any dependency between ϵ_i and $\hat{f}(x_i)$?
    - Yes, of course
    - Here ϵ_i is the noise in training data
    - $\hat{f}(x_i)$ is estimated on training data
    - ϵ_i and $\hat{f}(x_i)$ are highly dependent

$$E\left[\epsilon_i\ .\ (\hat{f}(x_i) - f(x_i))\right] \neq E\left[\epsilon_i\right]\ .\ E\left[(\hat{f}(x_i) - f(x_i))\right] \neq 0$$

$$E\left[(\hat{f}(x_i) - f(x_i))^2\right] = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\hat{f}(x_i) - y_i)^2}_{\mathcal{L}(\theta)} - \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 + \underbrace{2\ E\left[\epsilon_i\ (\hat{f}(x_i) - f(x_i))\right]}_{\mathcal{R}(\theta)}$$

$$E\left[(\hat{f}(x_i) - f(x_i))^2\right] = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\hat{f}(x_i) - y_i)^2}_{\mathcal{L}(\theta)} - \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 + \underbrace{2\ E\left[\epsilon_i\ (\hat{f}(x_i) - f(x_i))\right]}_{\mathcal{R}(\theta)}$$

Claim: $\mathcal{R}(\theta)$ is the component responsible for capturing model complexity.

$$E\left[(\hat{f}(x_i) - f(x_i))^2\right] = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\hat{f}(x_i) - y_i)^2}_{\mathcal{L}(\theta)} - \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 + \underbrace{2\ E\left[\epsilon_i\ (\hat{f}(x_i) - f(x_i))\right]}_{\mathcal{R}(\theta)}$$

Claim: $\mathcal{R}(\theta)$ is the component responsible for capturing model complexity.

$$\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\ (\hat{f}(x_i) - f(x_i)) = \frac{\sigma^2}{n}\sum_{i=1}^{n}\frac{\partial \hat{f}(x_i)}{\partial y_i} \quad \text{(Using Stein's Lemma we can show)}$$

$$E\left[(\hat{f}(x_i) - f(x_i))^2\right] = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\hat{f}(x_i) - y_i)^2}_{\mathcal{L}(\theta)} - \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 + \underbrace{2\ E\left[\epsilon_i\ (\hat{f}(x_i) - f(x_i))\right]}_{\mathcal{R}(\theta)}$$

Claim: $\mathcal{R}(\theta)$ is the component responsible for capturing model complexity.

$$\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\ (\hat{f}(x_i) - f(x_i)) = \frac{\sigma^2}{n}\sum_{i=1}^{n}\frac{\partial \hat{f}(x_i)}{\partial y_i} \quad \text{(Using Stein's Lemma we can show)}$$

- $\frac{\partial \hat{f}(x_i)}{\partial y_i}$ captures the rate of change of $\hat{f}(x_i)$ w.r.t. y_i
- When is $\frac{\partial \hat{f}(x_i)}{\partial y_i}$ high?

$$E\left[(\hat{f}(x_i) - f(x_i))^2\right] = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\hat{f}(x_i) - y_i)^2}_{\mathcal{L}(\theta)} - \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 + \underbrace{2\ E\left[\epsilon_i\ (\hat{f}(x_i) - f(x_i))\right]}_{\mathcal{R}(\theta)}$$

Claim: $\mathcal{R}(\theta)$ is the component responsible for capturing model complexity.

$$\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\ (\hat{f}(x_i) - f(x_i)) = \frac{\sigma^2}{n}\sum_{i=1}^{n}\frac{\partial \hat{f}(x_i)}{\partial y_i} \quad \text{(Using Stein's Lemma we can show)}$$

- $\frac{\partial \hat{f}(x_i)}{\partial y_i}$ captures the rate of change of $\hat{f}(x_i)$ w.r.t. y_i
- When is $\frac{\partial \hat{f}(x_i)}{\partial y_i}$ high?
  - With a small change in y_i, the change in $\hat{f}(x_i)$ considerably large

$$E\left[(\hat{f}(x_i) - f(x_i))^2\right] = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\hat{f}(x_i) - y_i)^2}_{\mathcal{L}(\theta)} - \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 + \underbrace{2\ E\left[\epsilon_i\ (\hat{f}(x_i) - f(x_i))\right]}_{\mathcal{R}(\theta)}$$

Claim: $\mathcal{R}(\theta)$ is the component responsible for capturing model complexity.

$$\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\ (\hat{f}(x_i) - f(x_i)) = \frac{\sigma^2}{n}\sum_{i=1}^{n}\frac{\partial \hat{f}(x_i)}{\partial y_i} \quad \text{(Using Stein's Lemma, can be shown)}$$

- $\frac{\partial \hat{f}(x_i)}{\partial y_i}$ captures the rate of change of $\hat{f}(x_i)$ w.r.t. y_i
- When is $\frac{\partial \hat{f}(x_i)}{\partial y_i}$ high?
  - When with a small change in y_i, the change in $\hat{f}(x_i)$ considerably large
  - When the model is sensitive to the training sample (such as complex model)

Modification of the Objective Function

$$E\left[(\hat{f}(x_i) - f(x_i))^2\right] = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\hat{f}(x_i) - y_i)^2}_{\mathcal{L}(\theta)} - \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 + \underbrace{2\ E\left[\epsilon_i\ (\hat{f}(x_i) - f(x_i))\right]}_{\mathcal{R}(\theta)}$$

To capture *true error* during training, a modification in the objective function is required.
Modified objective function

$$\min_{\theta} \mathcal{L}(\theta) + \mathcal{R}(\theta) = \underbrace{\mathcal{L}_R(\theta)}_{\text{Regularized loss}}$$

Where $\mathcal{R}(\theta)$ is the regularized term, which is to be high for complex models and small for simple models

Different Forms of Regularization

- l_2 regularization
- Early stopping
- Dropout
- Adding Noise to the inputs
- Adding Noise to the outputs
- Ensemble methods
- Dataset augmentation
- Parameter Sharing and tying

l_2 Regularization

Objective Function for l_2 Regularization

$$\mathcal{L}_R(\theta) = \mathcal{L}(\theta) + \frac{\gamma}{2} ||\theta||^2$$

- γ is a regularized hyper-parameter
- What if γ is close to 0

Objective Function for l_2 Regularization

$$\mathcal{L}_R(\theta) = \mathcal{L}(\theta) + \frac{\gamma}{2} ||\theta||^2$$

- γ is a regularized hyper-parameter
- What if γ is close to 0
  - $\mathcal{L}_R(\theta) = \mathcal{L}(\theta)$

Objective Function for l_2 Regularization

$$\mathcal{L}_R(\theta) = \mathcal{L}(\theta) + \frac{\gamma}{2} ||\theta||^2$$

- γ is a regularized hyper-parameter
- What if γ is close to 0
  - $\mathcal{L}_R(\theta) = \mathcal{L}(\theta)$
  - Overfitting problem remains

Objective Function for l_2 Regularization

$$\mathcal{L}_R(\theta) = \mathcal{L}(\theta) + \frac{\gamma}{2} ||\theta||^2$$

- γ is a regularized hyper-parameter
- What if γ is close to 0
  - $\mathcal{L}_R(\theta) = \mathcal{L}(\theta)$
  - Overfitting problem remains
- What if γ is high

Objective Function for l_2 Regularization

$$\mathcal{L}_R(\theta) = \mathcal{L}(\theta) + \frac{\gamma}{2} ||\theta||^2$$

- γ is a regularized hyper-parameter
- What if γ is close to 0
  - $\mathcal{L}_R(\theta) = \mathcal{L}(\theta)$
  - Overfitting problem remains
- What if γ is high
  - $\frac{\gamma}{2} ||\theta||^2$ will be dominating term in $\mathcal{L}_R(\theta)$

Objective Function for l_2 Regularization

$$\mathcal{L}_R(\theta) = \mathcal{L}(\theta) + \frac{\gamma}{2} ||\theta||^2$$

- γ is a regularized hyper-parameter
- What if γ is close to 0
  - $\mathcal{L}_R(\theta) = \mathcal{L}(\theta)$
  - Overfitting problem remains
- What if γ is high
  - $\frac{\gamma}{2} ||\theta||^2$ will be dominating term in $\mathcal{L}_R(\theta)$
  - Optimizer tries to optimize $\frac{\gamma}{2} ||\theta||^2$

Objective Function for l_2 Regularization

$$\mathcal{L}_R(\theta) = \mathcal{L}(\theta) + \frac{\gamma}{2} ||\theta||^2$$

- γ is a regularized hyper-parameter
- What if γ is close to 0
  - $\mathcal{L}_R(\theta) = \mathcal{L}(\theta)$
  - Overfitting problem remains
- What if γ is high
  - $\frac{\gamma}{2} ||\theta||^2$ will be dominating term in $\mathcal{L}_R(\theta)$
  - Optimizer tries to optimize $\frac{\gamma}{2} ||\theta||^2$
  - Causes underfitting problem

We have now two loss functions:

- $\mathcal{L}(\theta)$: Unregularized training loss
- $\mathcal{L}_R(\theta) = \mathcal{L}(\theta) + \frac{\gamma}{2} ||\theta||^2$: Regularized training loss

Next step is to establish the relationship between the optimal solutions to unregularized and regularized losses, respectively

- We assume, θ^* and θ_R^* are the optimal solutions to unregularized and regularized losses, respectively
- This implies, $\nabla\mathcal{L}(\theta^*) = 0$ and $\nabla\mathcal{L}_R(\theta_R^*) = 0$
- For Stochastic GD or its variants, we have:

$$\nabla\mathcal{L}_R(\theta) = \nabla\mathcal{L}(\theta) + \gamma\ \theta$$

Algorithm 1 Pseudocode for Feedforward Network with Backpropagation with Regularized Loss

1: $t \leftarrow 0$ {Iteration count}
2: $\theta_0 = (W_1, W_2, \ldots, W_L, B_1, B_2, \ldots, B_L)$; {Initialize learning parameters}
3: **repeat**
4: $M \leftarrow$ ForwardPropagation(θ_t); {M is the model $(z_i, a_i, \hat{y})$}
5: $\nabla_\theta^t \leftarrow$ Backpropagation (M);
6: $\theta_{t+1} \leftarrow \theta_t - \eta \, \nabla_\theta^t \; - \eta \, \gamma \, \theta_{\mathbf{t}}$;
7: $t+ = 1$;
8: **until** Converge

Consider any solution $\theta = \theta^* + \psi$;
Using Taylor series approximation (upto 2^{nd} order)

$$\mathcal{L}(\theta^* + \psi) = \mathcal{L}(\theta^*) + \psi^T \left(\nabla \mathcal{L}(\theta^*)\right) + \frac{1}{2}\psi^T \left(\nabla^2 \mathcal{L}(\theta^*)\right)\psi$$

$$\mathcal{L}(\theta) = \mathcal{L}(\theta^*) + \psi^T .0 + \frac{1}{2}\psi^T \left(\nabla^2 \mathcal{L}(\theta^*)\right)\psi$$

$$\nabla_\theta \mathcal{L}(\theta) = \nabla_\theta \left(\mathcal{L}(\theta^*) + \frac{1}{2}(\theta - \theta^*)^T \left(\nabla^2 \mathcal{L}(\theta^*)\right)(\theta - \theta^*)\right)$$

$$\nabla \mathcal{L}(\theta) = \left(\nabla^2 \mathcal{L}(\theta^*)\right)(\theta - \theta^*)$$

Now, considering regularized loss

$$\mathcal{L}_R(\theta) = \mathcal{L}(\theta) + \frac{\gamma}{2} ||\theta||^2$$

$$\nabla \mathcal{L}_R(\theta) = \nabla \mathcal{L}(\theta) + \gamma\ \theta$$

$$= \left(\nabla^2 \mathcal{L}(\theta^*)\right)(\theta - \theta^*) + \gamma\ \theta$$

$$\nabla \mathcal{L}_R(\theta) = \left(\nabla^2 \mathcal{L}(\theta^*)\right)(\theta - \theta^*) + \gamma\ \theta = \mathcal{H}(\theta - \theta^*) + \gamma\ \theta$$

Recollect, θ_R^* is the optimal solution for $\mathcal{L}_R(\theta)$, then

$$\nabla \mathcal{L}_R(\theta_R^*) = \mathcal{H}(\theta_R^* - \theta^*) + \gamma\ \theta_R^* = 0$$

$$\begin{aligned} \mathcal{H}(\theta_R^* - \theta^*) + \gamma\ \theta_R^* &= 0 \\ \implies (\mathcal{H} + \gamma \mathbb{I})\ \theta_R^* &= \mathcal{H}\ \theta^* \\ \implies \theta_R^* &= (\mathcal{H} + \gamma \mathbb{I})^{-1} \mathcal{H}\ \theta^* \end{aligned}$$

Observation

If $\gamma \to 0$, $\theta_R^* = \theta^*$, No regularization, Overfitting problem persists

Assumption

$\mathcal{H}$ is a positive semi definite matrix

$$\theta_R^* = (\mathcal{H} + \gamma \mathbb{I})^{-1} \mathcal{H}\ \theta^*$$

$$\mathcal{H} = Q\ \Lambda\ Q^T \quad (Q \text{ is a orthogonal matrix})$$

$$\begin{aligned}
\theta_R^* &= (Q\ \Lambda\ Q^T + \gamma \mathbb{I})^{-1} Q\ \Lambda\ Q^T\ \theta^* \\
&= (Q\ \Lambda\ Q^T + \gamma Q\ \mathbb{I}\ Q^T)^{-1} Q\ \Lambda\ Q^T\ \theta^* \\
&= \left[Q(\Lambda + \gamma \mathbb{I}) Q^T\right]^{-1} Q\ \Lambda\ Q^T\ \theta^* \\
&= (Q^T)^{-1} (\Lambda + \gamma \mathbb{I})^{-1} Q^{-1} Q\ \Lambda\ Q^T\ \theta^* \\
&= Q(\Lambda + \gamma \mathbb{I})^{-1}\ \Lambda\ Q^T\ \theta^* \\
&= Q\ D\ Q^T\ \theta^* \quad (D = (\Lambda + \gamma \mathbb{I})^{-1}\ \Lambda)
\end{aligned}$$

Assumption

$$\theta_R^* = Q\ D\ Q^T\ \theta^* \quad (D = (\Lambda + \gamma \mathbb{I})^{-1}\ \Lambda)$$

Assumption

$$\theta_R^* = Q\ D\ Q^T\ \theta^* \quad (D = (\Lambda + \gamma\mathbb{I})^{-1}\ \Lambda)$$

- Q^T rotates θ^*
- D scales $Q^T\ \theta^*$
  - i^{th} element of $Q^T\ \theta^*$ scaled by $\frac{\lambda_i}{\lambda_i+\gamma}$
- Q rotates back $D\ Q^T\ \theta^*$

Analysis

- If $\lambda_i >> \gamma$, $\frac{\lambda_i}{\lambda_i+\gamma} = 1$
- If $\gamma >> \lambda_i$, $\frac{\lambda_i}{\lambda_i+\gamma} = 0$
- Therefore, $\sum_{i=1}^{n} \frac{\lambda_i}{\lambda_i+\gamma} \leq n$

Thank You!