

Probability and Random Variables

- Probability
- Random Variables
- Expectation
- Conditional Probability
- Conditional Expectation

Probability

- A probability space is a triple (Ω, \mathcal{B}, P) s.t.
 - Ω be a sample space
 - \mathcal{B} is the set of “well behaved” subsets of Ω
 - Note: If Ω is discrete, then $\mathcal{B} = 2^\Omega$, i.e., the power set of Ω
 - $A \in \mathcal{B}$ is an event
 - $P(A)$ is the probability of the event A

- Axioms of probability

$$P(\Omega) = 1$$

$$\forall A \in \mathcal{B}, P(A) \geq 0$$

$$\text{If } \forall i \neq j A_i \cap A_j = \emptyset, \text{ then } P(\cup_i A_i) = \sum_i P(A_i)$$

Random Variables

- A random variable is a “well behaved” function
 - $X: \Omega \rightarrow \mathfrak{R}$
 - Can be thought of as the outcome of an experiment
- Probability that $X = 2$ is given by
 - Event $A = \{\omega \in \Omega: X(\omega) = 2\}$
 - Then $P\{X = 2\} = P(A) = P(\{\omega \in \Omega: X(\omega) = 2\})$

Discrete Probability Density

- A discrete random variable, X , with density

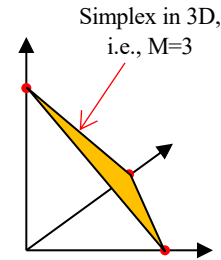
$$P\{X = i\} = p_i$$

for $i = 0, \dots, M - 1$, with $p_i \geq 0$ and $1 = \sum_{i=0}^{M-1} p_i$.

Notice that $p \in \mathcal{S}^M \Leftarrow$ is a simplex set.

- Then we have

$$P\{X \in A\} = \sum_{i \in A} p_i$$



Expectation, Mean, and Variance

- Let $P\{X = i\} = p_i$,
 - Expected value of X

$$\mu = E[X] = \sum_{n=0}^{M-1} n p_n$$

- Expected value of $f(X)$

$$E[f(X)] = \sum_{n=0}^{M-1} f(n) p_n$$

- Variance X

$$\sigma^2 = E[(X - \mu)^2] = \sum_{n=0}^{M-1} (n - \mu)^2 p_n$$

Marginal and Conditional Probability

- Consider the two random variables, X and Y

$$P\{X = i, Y = j\} = p(i, j)$$

- Marginal densities

$$P\{X = i\} = \sum_{j=0}^{M-1} P\{X = i, Y = j\} = \sum_{j=0}^{M-1} p(i, j) = p_x(i)$$

$$P\{Y = j\} = \sum_{i=0}^{M-1} P\{X = i, Y = j\} = \sum_{i=0}^{M-1} p(i, j) = p_y(j)$$

- Conditional densities

$$p_{x|y}(i|j) = P\{X = i|Y = j\} = \frac{P\{X = i, Y = j\}}{P\{Y = j\}} = \frac{p(i, j)}{p_y(j)} = \frac{p(i, j)}{\sum_{i=0}^{M-1} p(i, j)}$$

$$p_{y|x}(j|i) = P\{Y = j|X = i\} = \frac{P\{X = i, Y = j\}}{P\{X = i\}} = \frac{p(i, j)}{p_x(i)} = \frac{p(i, j)}{\sum_{j=0}^{M-1} p(i, j)}$$

Conditional Expectation

- Consider the two random variables, X and Y

$$P\{X = i, Y = j\} = p(i, j)$$

- Conditional densities


$$p_{x|y}(i|j) = \frac{p(i, j)}{\sum_{i=0}^{M-1} p(i, j)}$$

$$p_{y|x}(j|i) = \frac{p(i, j)}{\sum_{j=0}^{M-1} p(i, j)}$$

- Conditional expectation


$$E[f(X)|Y] = \sum_{n=0}^{M-1} f(n) p_{x|y}(n|Y)$$

*This is a function of
 Y*



$$E[f(Y)|X] = \sum_{n=0}^{M-1} f(n) p_{y|x}(n|X)$$

*This is a function of
 X*



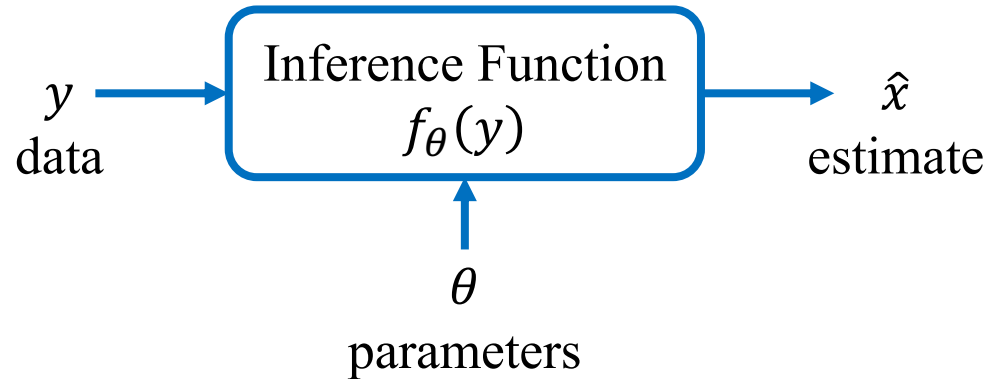
Random Variables versus Numbers

- Remember:
 - An function of a random variable is a random variable.
 - The expectation is a number.
 - But the conditional expectation is a random variable.
- Examples: Let X be a random variable, then consider
 - X^2
 - $E[X^2]$
 - $E[Y|X]$
 - $E[X|X]$
 - $E[YX|X]$

Estimation

- Frequentist Viewpoint
- The ML Estimate
- Bayesian Viewpoint
- The MMSE and MAP Estimates
- The Bias Variance Tradeoff
- Regularization

The Big Picture



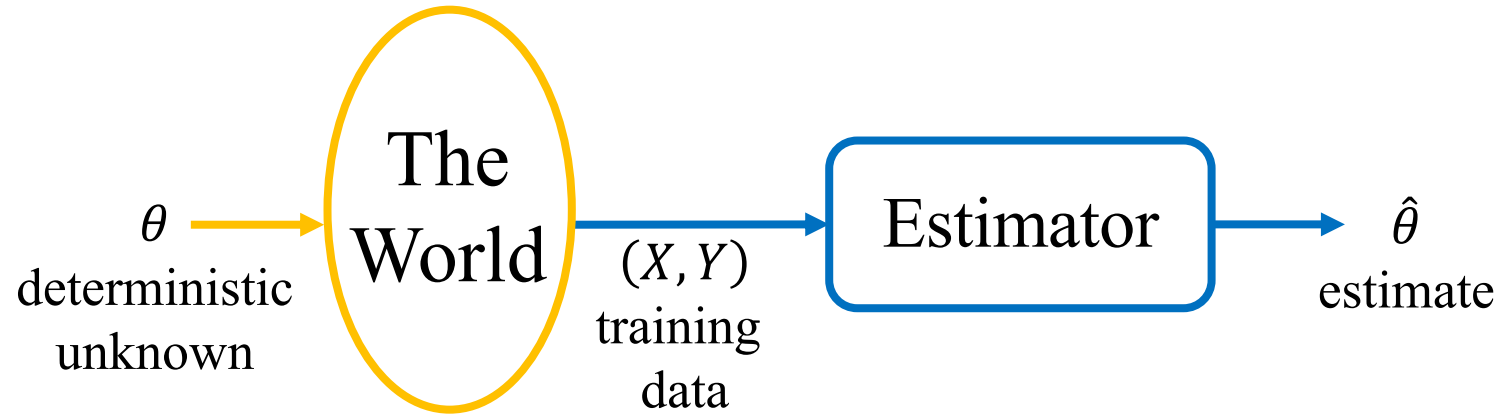
■ Inference

- Requires estimation of \hat{x} from y
- Typically Bayesian approach
- Always biased

■ Training

- Requires estimation of $\hat{\theta}$ from training data $(x_k, y_k) |_{k=0}^{K-1}$
- Typically Frequentist approach
- Typically unbiased (unless regularized)

Frequentist View of the World



- Model of world

$$(X, Y) \sim p_{\theta}(x, y) = P_{\theta}\{X = x, Y = y\}$$

- Estimate of unknown

$$\hat{\theta} = T(X, Y)$$

- Unbiased estimate

$$E[\hat{\theta}|\theta] = \theta$$

- Bias and variance are given by

$$\text{Bias} = E[\hat{\theta}|\theta] - \theta$$

$$\text{Variance} = E[\|\hat{\theta} - E[\hat{\theta}|\theta]\|^2|\theta]$$

Maximum Likelihood Estimate (MLE)

- Model of world

$$(X, Y) \sim p_{\theta}(x, y)$$

- The MLE estimate is defined as

$$\hat{\theta} = \arg \max_{\theta} \{p_{\theta}(X, Y)\}$$

$$= \arg \min_{\theta} \{-\log p_{\theta}(X, Y)\}$$

usually can be computed as the solution to

$$0 = \nabla_{\theta} \log p_{\theta}(X, Y) \Big|_{\theta=\hat{\theta}}$$

MLE for Regression

- Example: MLE training for regression

Assume that we have K i.i.d. training pairs, $(X_0, Y_0), \dots, (X_{K-1}, Y_{K-1})$, such that

$$X_k = f_\theta(Y_k) + W_k$$

where $W_k \sim N(0, \sigma^2 I)$. Then we have that

$$\log p_\theta(x, y) = \log p_\theta(x|y) + \log p(y)$$

and

$$p_\theta(x_k|y_k) = \frac{1}{(2\pi\sigma^2)^{p/2}} \exp\left\{-\frac{1}{2\sigma^2} \|x_k - f_\theta(y_k)\|^2\right\}$$

So the MLE is given by

$$\begin{aligned}\hat{\theta} &= \arg \min_{\theta} \{-\log p_\theta(x, y)\} \\ &= \arg \min_{\theta} \{-\log p_\theta(x|y) - \log p(y)\} \\ &= \arg \min_{\theta} \left\{ \sum_{k=0}^{K-1} -\log p_\theta(x_k|y_k) \right\} \\ &= \arg \min_{\theta} \left\{ \sum_{k=0}^{K-1} \frac{1}{2\sigma^2} \|x_k - f_\theta(y_k)\|^2 + \frac{p}{2} \log\{2\pi\sigma^2\} \right\} \\ &= \arg \min_{\theta} \left\{ \frac{1}{K} \sum_{k=0}^{K-1} \|x_k - f_\theta(y_k)\|^2 \right\} \\ &= \arg \min_{\theta} L_{MSE}(\theta; x, y)\end{aligned}$$

MSE Loss \Rightarrow maximum likelihood parameter estimate

MLE Training for Classification (1)

- Our goal is to train a DNN by estimating the probability $0 \leq m < M$

$$P\{X = \text{has class } m | Y\} = f_{\theta}(m; Y)$$

To do this, we will estimate $\hat{\theta}$ from training pairs (x_k, y_k) for $0 \leq k < K$ using the maximum likelihood estimator (MLE).

The class x_k is represented using 1-hot encoding, so we have that

$$x_{k,m} = \delta(m - m_k^*)$$

where m_k^* is the class of x_k .

Example:

$$x_k = [0, \dots, 0, 1, 0, \dots, 0]$$

\wedge
 m_k^*

- Notice that x_k and m_k^* contain the same information.

MLE Training for Classification (2)

The DNN estimates the probability of x_k given y_k

$$p_{\theta}(x_k|y_k) = P\{X_k = x_k|Y_k = y_k\} = f_{\theta}(m_k^*; y_k)$$

So then we have that

$$\begin{aligned} -\log p_{\theta}(x_k|y_k) &= -\log f_{\theta}(m_k^*; y_k) \\ &= \sum_{m=0}^{M-1} -x_{k,m} \log f_{\theta}(m; y_k) \end{aligned}$$

Since each training pair is assumed independent,

$$\begin{aligned} -\log p_{\theta}(x|y) &= \sum_{k=1}^K -\log p_{\theta}(x_k|y_k) \\ &= \sum_{k=1}^K \sum_{m=0}^{M-1} -x_{k,m} \log f_{\theta}(m; y_k) \\ &= \sum_{k=0}^{K-1} \rho_{CE}(x_k, f_{\theta}(\cdot; y_k)) \\ &= K \cdot L_{CE}(\theta; x, y) \end{aligned}$$

Only one term of this sum is non-zero due to 1-hot encoding

Cross entropy distortion function

1-hot encoded ground truth

Vector of predicted class probabilities

Cross entropy loss function

$$\rho_{CE}(a, b) = - \sum_i a_i \log b_i$$

MLE Training for Classification (3)

Putting this together, the maximum likelihood estimate is given by

$$\begin{aligned}\hat{\theta} &= \arg \min_{\theta} \{-\log p_{\theta}(x, y)\} \\ &= \arg \min_{\theta} \{-\log p_{\theta}(x|y) - \log p(y)\} \\ &= \arg \min_{\theta} \{-\log p_{\theta}(x|y)\} \\ &= \arg \min_{\theta} \{L_{CE}(\theta; x, y)\}\end{aligned}$$

$$\hat{\theta} = \arg \min_{\theta} \{L_{CE}(\theta; x, y)\}$$

Cross Entropy Loss \Rightarrow maximum likelihood parameter estimate

MLE for Multinomial Distribution

- Example: Let $Y = (Y_0, \dots, Y_{N-1})$ be a sequence of independent and identically distributed (i.i.d.) random variables

$$P_{\theta}\{Y_n = i\} = \theta_i \text{ for } \theta \in \mathcal{S}^M$$

$$p_{\theta}(y) = P_{\theta}\{Y = y\} = \prod_{i=0}^{M-1} \theta_i^{N_i}$$

where

$$N_i = \sum_{n=0}^{N-1} \delta(y_n = i)$$

Then

$$-\frac{1}{N} \log p_{\theta}(y) = -\sum_{i=0}^{M-1} \frac{N_i}{N} \log \theta_i = \text{CrossEntropy}\left(\frac{N_i}{N}, \theta_i\right)$$

Then MLE is given by

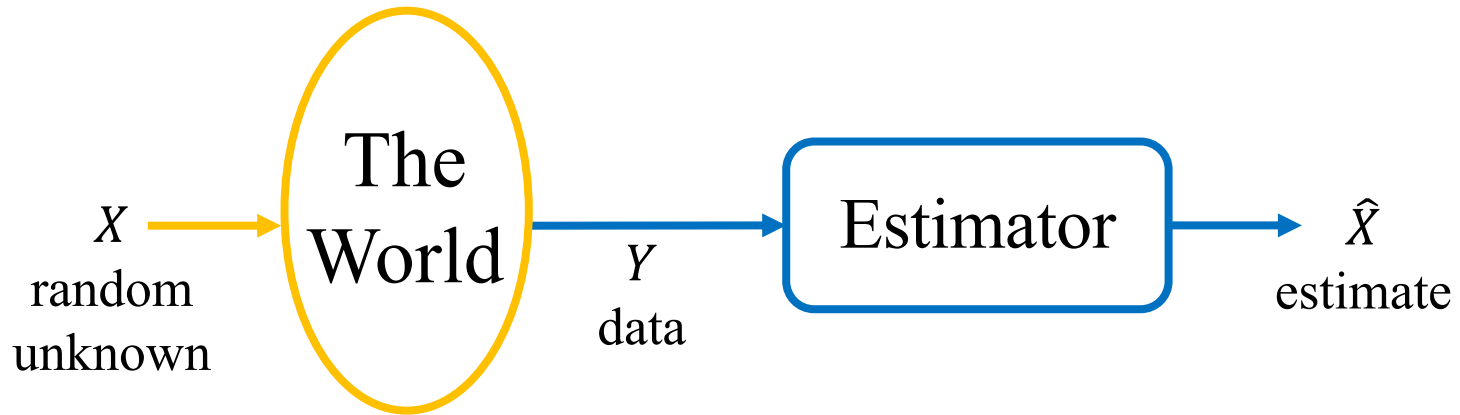
$$\hat{\theta} = \arg \min_{\theta} \left\{ -\sum_{i=0}^{M-1} \frac{N_i}{N} \log \theta_i \right\} = \frac{N_i}{N}$$

The Big Idea About ML Estimate

- The ML Estimate is...
 - Pure and ideal. It only uses the data.
 - It is (mostly) unbiased.*
 - It is asymptotically efficient, which means it achieves the Cramer Rao bound asymptotically.
 - It's mostly used when there is plenty of data.
- But...
 - It tends to overfit when there is not enough data.

*This isn't really true, but it's mostly true, and the truth is too complicated to explain right now.

Bayesian View of the World



- Model of world

$$Y|X \sim p_{y|x}(y|x) = P\{Y = y|X = x\} \Leftarrow \text{forward model}$$

$$X \sim p_x(x) = P\{X = x\} \Leftarrow \text{prior model}$$

- Estimate of unknown

$$\hat{X} = f(Y)$$

Bayes Law and the Posterior Distribution

- Model of world

$$Y|X \sim p_{y|x}(y|x) = P\{Y = y|X = x\} \Leftarrow \text{forward model}$$

$$X \sim p_x(x) = P\{X = x\} \Leftarrow \text{prior model}$$

- Bayes Law

$$p_{x|y}(x|y) = \frac{p_{y|x}(y|x) p_x(x)}{p_y(y)} = \frac{p_{y|x}(y|x) p_x(x)}{\sum_{k=0}^{M-1} p_{y|x}(y|k) p_x(k)}$$

*Posterior
Distribution*

Common Bayesian Estimators

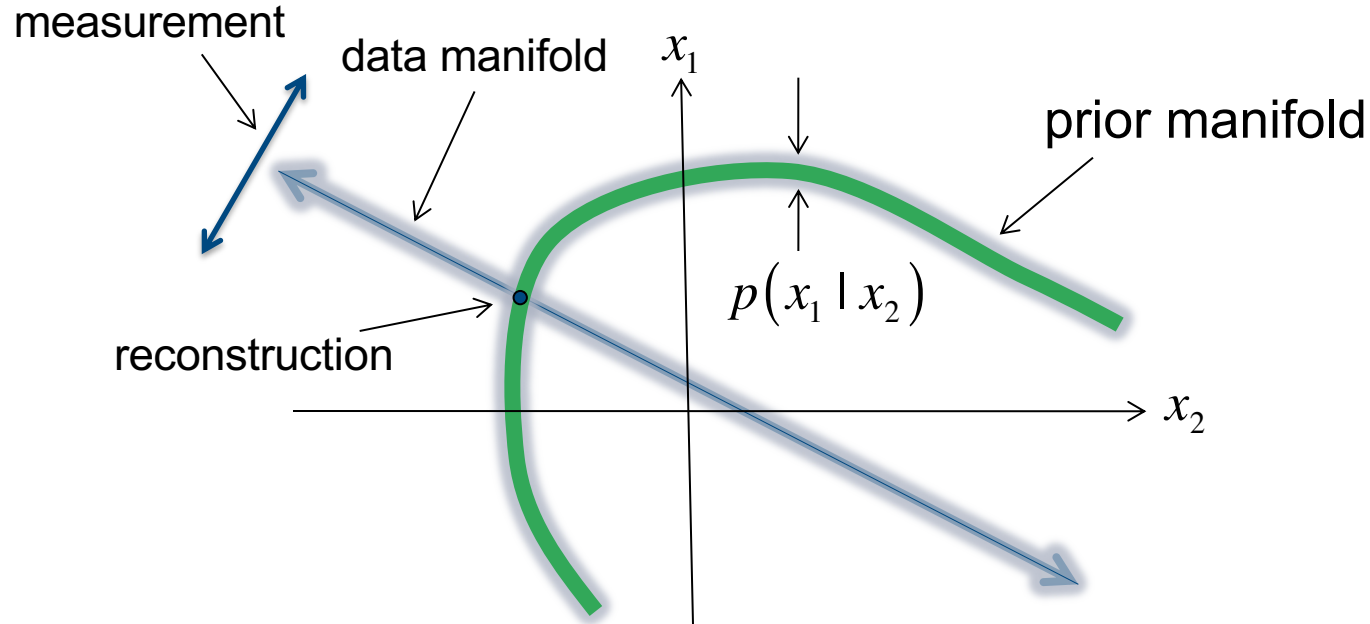
- The Minimum Mean Squared Error (MMSE) estimate

$$\hat{X} = E[X|Y] = \sum_{x=0}^{M-1} x p_{x|y}(x|Y)$$

- Maximum A Posteriori (MAP) estimate

$$\begin{aligned}\hat{x} &= \arg \max_x \{p_{x|y}(x|y)\} \\ &= \arg \min_x \{-\log p_{x|y}(x|y)\} \\ &= \arg \min_x \{-\log p_{y|x}(y|x) - \log p_x(x) + \log p_y(y)\} \\ &= \arg \min_x \{-\log p_{y|x}(y|x) - \log p_x(x)\}\end{aligned}$$

Manifold Interpretation of Bayesian Estimators



$$\hat{x} = \arg \min_x \{-\log p_{y|x}(y|x) - \log p_x(x)\}$$

- Notice that prior manifold fills the space but is not a linear manifold
 - But it has thickness
 - Dimension of measurement > dimension of manifold
- Prior information can dramatically simplify the estimation problem.

The Bias and Variance

■ Definitions*

– Frequentist:

$$\text{bias}_\theta = E[\hat{\theta}|\theta] - \theta$$

$$\text{Var}_\theta = E[\|\hat{\theta} - E[\hat{\theta}|\theta]\|^2|\theta]$$

$$\text{MSE}_\theta = \text{Var}_\theta + (\text{bias}_\theta)^2$$

– Bayesian*:

$$\text{bias}_x = E[\hat{X}|X = x] - x$$

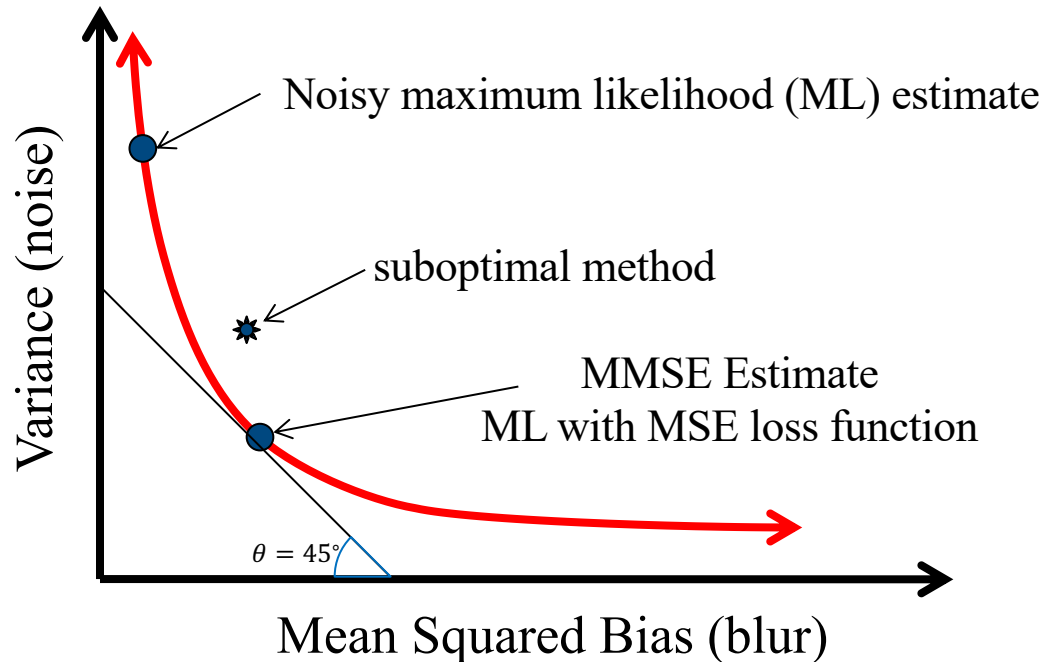
$$\overline{\text{bias}^2} = E[\|E[\hat{X}|X] - X\|^2]$$

$$\text{Var} = E[\|\hat{X} - E[\hat{X}|X]\|^2]$$

$$\text{MSE} = \text{Var} + \overline{\text{bias}^2}$$

- Bias is usually less desirable than variance

The Bias Variance Tradeoff



- The tradeoff
 - Bias is usually worse than variance
- Interpretation
 - Variance \Leftrightarrow noise
 - Bias \Leftrightarrow excessive smoothing, regularization or blur
 - As bias $\Rightarrow 0$, then variance $\Rightarrow \infty$
 - Not possible to estimate a solution with infinite resolution using a finite amount of data

Bayesian Versus Frequentist Estimation

- Bayesian Estimation

- Usually used for ML inference
- Typically high-bias low-variance estimates
- Most appropriate when prior information is strong
- Most appropriate when the prior information is strong; or when the amount of data is small and/or the quality of data is poor

- Frequentist Estimation

- Usually used for parameter estimation
- Typically low-bias high-variance estimates
- Most appropriate when prior information is weak; or when the amount of data is large and/or the quality of data is high.

Regularized Maximum Likelihood

- Regularize ML estimate:

$$\hat{\theta} = \arg \min_{\theta} \{-\log p_{\theta}(x, y) + \beta S(\theta)\}$$

where $S(\theta)$ is a “regularizing” function, and β is the regularization weight.

Typical choices are

$$S(\theta) = -\log p(\theta) \quad \leftarrow \text{MAP estimate}$$

$$S(\theta) = \|\theta\|^2 \quad \leftarrow \begin{array}{l} \text{Like a Gaussian Prior} \\ \text{Reduces amplitude of weights} \end{array}$$

$$S(\theta) = \|\theta\|_1 \quad \leftarrow \begin{array}{l} \text{Like a Laplacian Prior} \\ \text{Encourages weights to go to zero} \end{array}$$

- Modified Loss function

$$\tilde{L}(\theta) = L(\theta) + \beta S(\theta)$$

- Can be interpreted as MAP estimate with $p(\theta) = \frac{1}{z} \exp\left\{-\frac{\beta}{2} S(\theta)\right\}$
- Introduces bias into the estimate of θ
- Reduces overfitting
- Use regularization if training error \ll validation error