Foundations of Machine Learning
AI2000 and AI5000

FoML-04
Maximum Likelihood Principle

Dr. Konda Reddy Mopuri
Department of AI, IIT Hyderabad
July-Nov 2025

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

DiL
Data-driven Intelligence
& Learning Lab

So far in FoML

- What is ML and the learning paradigms
- Probability refresher
  - Random variables, Bayes Theorem, Independence, Expectation, Variance

Maximum Likelihood Principle

Maximum Likelihood Principle

- Widely used technique for optimizing model parameters

Maximum Likelihood Principle

- Given - Dataset of N independent observations D

Maximum Likelihood Principle

- Goal: recover the probability distribution that may have generated this dataset

Maximum Likelihood Principle

- Goal: recover the probability distribution that may have generated this dataset
- Likelihood of the dataset $p(D|w)$

Maximum Likelihood Principle

- The most likely 'explanation' of D is given by w_{ML} that maximizes the likelihood function

$$\mathbf{w}_{ML} =$$

Maximum Likelihood Principle

- The iid assumption - each $x_i \in D$ is independently distributed according to the same distribution conditioned on **w**

Maximum Likelihood Principle

- The iid assumption - each $x_i \in D$ is independently distributed according to the same distribution conditioned on w

The joint distribution

Maximum Likelihood Principle

$$\mathbf{w}_{ML} = \arg\max_{\mathbf{w}} \; p(D|\mathbf{w}) = \arg\max_{\mathbf{w}} \; \prod_{i=1}^{N} p(x_i|\mathbf{w})$$

Maximum Likelihood Principle

$$\mathbf{w}_{ML} = \arg\max_{\mathbf{w}} \; p(D|\mathbf{w}) = \arg\max_{\mathbf{w}} \; \prod_{i=1}^{N} p(x_i|\mathbf{w})$$

- Numerical underflow

Maximum Likelihood Principle

$$\mathbf{w}_{ML} = \arg\max_{\mathbf{w}} \; p(D|\mathbf{w}) = \arg\max_{\mathbf{w}} \; \prod_{i=1}^{N} p(x_i|\mathbf{w})$$

- Numerical underflow
- Maximize the log-likelihood →

Maximum Likelihood Principle

$$\mathbf{w}_{ML} = \arg\max_{\mathbf{w}} \; p(D|\mathbf{w}) = \arg\max_{\mathbf{w}} \; \prod_{i=1}^{N} p(x_i|\mathbf{w})$$

- Numerical underflow
- Maximize the log-likelihood

$$\mathbf{w}_{ML} = \arg\max_{\mathbf{w}} \; \log \prod_{i=1}^{N} p(x_i|\mathbf{w})$$

Error function:

$$E(D;\mathbf{w}) = -\log p(D|\mathbf{w}) = -\sum_{i=1}^{N} \log p(x_i|\mathbf{w})$$

MLE for Gaussian Distributions

- iid Gaussian distributed real variables D =

$$p(x|\mathbf{w}) = \mathcal{N}(x|\mu, \sigma^2) \qquad p(D|\mathbf{w}) = p(D|\mu, \sigma^2) =$$

MLE for Gaussian Distributions

- iid Gaussian distributed real variables D =

$$p(x|\mathbf{w}) = \mathcal{N}(x|\mu, \sigma^2) \qquad p(D|\mathbf{w}) = p(D|\mu, \sigma^2) =$$

log likelihood =

MLE for Gaussian Distributions

- Estimate the model parameters $$\mu_{ML}, \sigma^2_{ML} = \underset{\mu, \sigma^2}{\arg\max} \; \log p(D|\mu, \sigma^2)$$

MLE for Gaussian Distributions

- Estimate the model parameters $$\mu_{ML}, \sigma^2_{ML} = \arg\max_{\mu,\sigma^2} \ \log p(D|\mu, \sigma^2)$$

MLE for Gaussian Distributions

$$\mu_{ML}, \sigma^2_{ML} = \underset{\mu, \sigma^2}{\arg\max} \; \log p(D|\mu, \sigma^2)$$

- How well do these estimates represent the true parameters?

MLE for Gaussian Distributions

$$\mu_{ML}, \sigma^2_{ML} = \underset{\mu,\sigma^2}{\arg\max} \; \log p(D|\mu, \sigma^2)$$

- How well do these estimates represent the true parameters?
- Note that these are functions of the data sample

MLE for Gaussian Distributions

$$\mu_{ML}, \sigma^2_{ML} = \arg\max_{\mu, \sigma^2} \; \log p(D|\mu, \sigma^2)$$

- How well do these estimates represent the true parameters?
- Note that these are functions of the data sample
  - → expected values of these estimates

MLE for Gaussian Distributions

- ML estimate of the mean

MLE for Gaussian Distributions

- Bias of the "ML estimate of the mean"

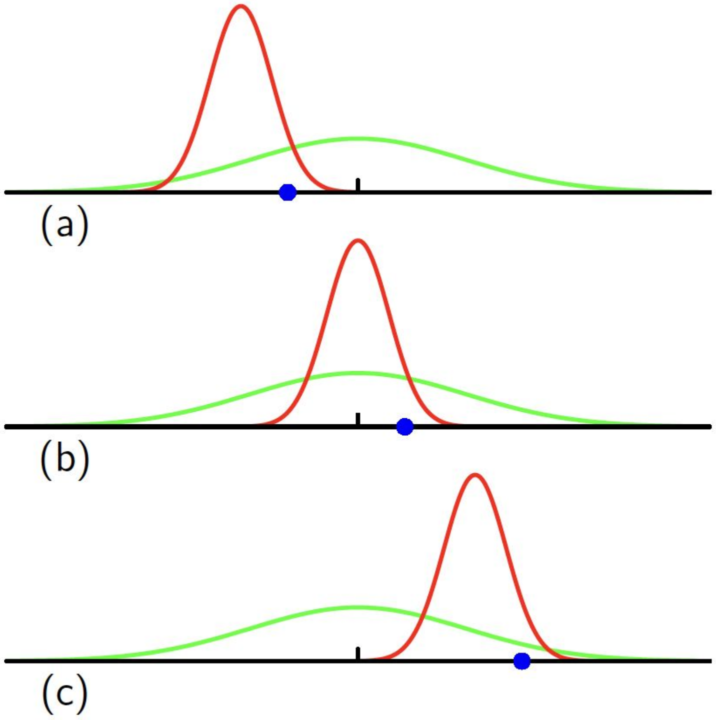
MLE for Gaussian Distributions

- ML estimate of the variance

MLE for Gaussian Distributions

- Bias of the "ML estimate of the variance"

Bias in variance estimate



Regression example

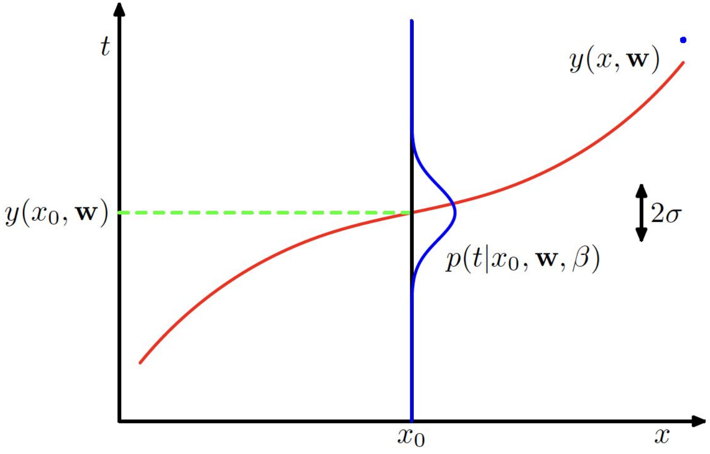
MLE for Regression (curve fitting)

- Given data D $D = \{(x_1, t_1), (x_2, t_2), \ldots (x_N, t_N)\} = \{\mathbf{x}, \mathbf{t}\}$

MLE for Regression (curve fitting)

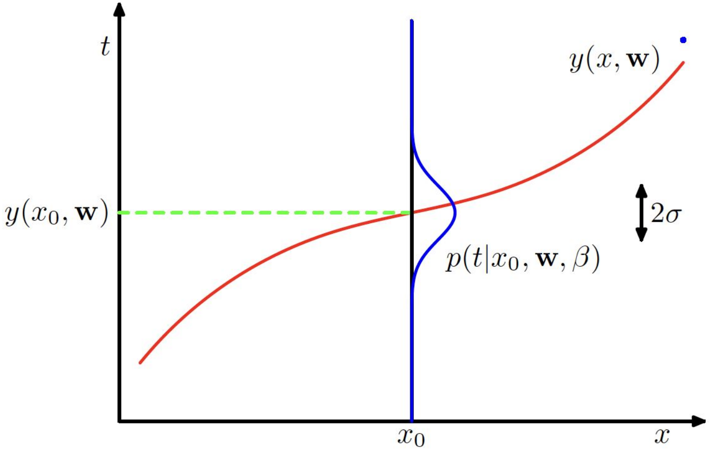
- Given data D $D = \{(x_1, t_1), (x_2, t_2), \ldots (x_N, t_N)\} = \{\mathbf{x}, \mathbf{t}\}$
- Assume the data is generated by

$$t = y(x, \mathbf{w}) + \sigma \cdot \epsilon, \quad \epsilon \in \mathcal{N}(0, 1)$$



MLE for Regression (curve fitting)

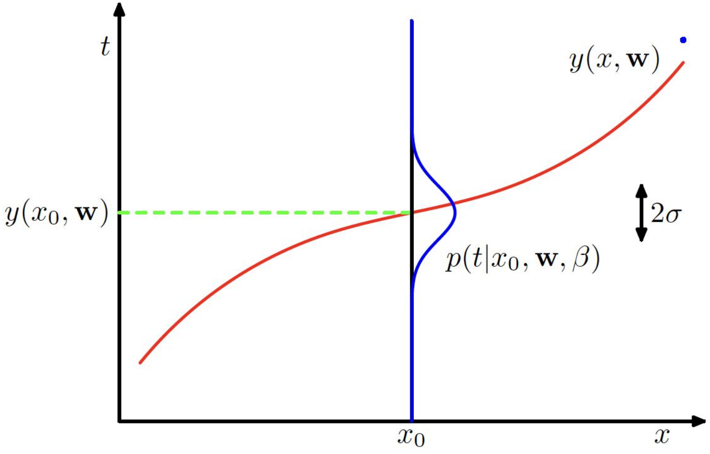
- Target distribution $$p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1})$$



MLE for Regression (curve fitting)

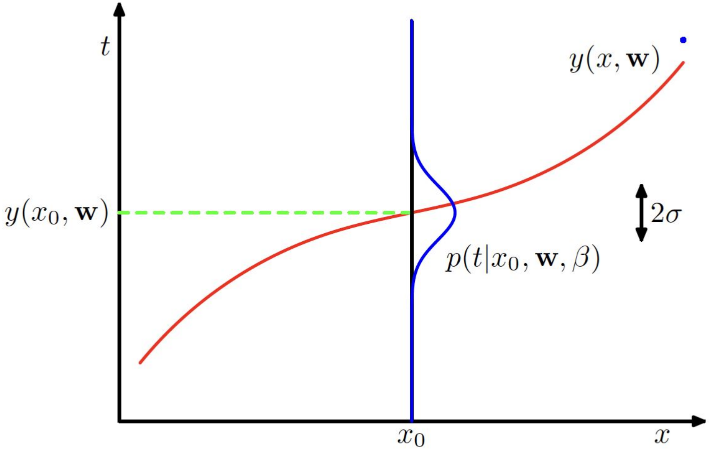
- Target distribution $$p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1})$$
- log likelihood

$$\log p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta^{-1})$$



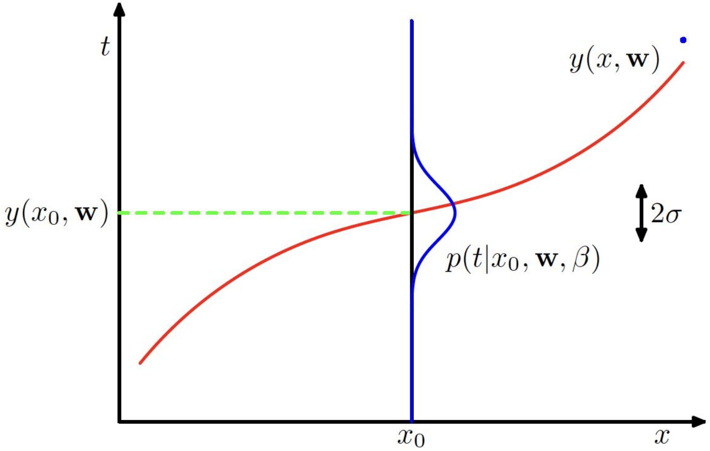
MLE for Regression (curve fitting)

- Minimize the NLL w.r.t the parameters w and β



MLE for Regression (curve fitting)

- The predictive distribution



Rough work

Next
MAP