인공지능 기초/
베이시안 방법과 정보이론

Cheongjae Jang

cjjang@hanyang.ac.kr

The 17th KIAS CAC Summer School on

Artificial Intelligence and Parallel Computing

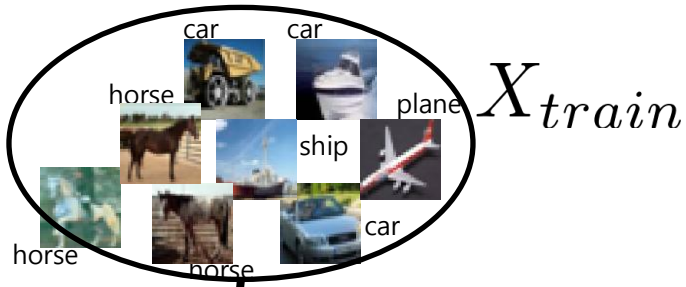
June 22-26, 2026

Contents

- Learning from data
- Bayesian methods
- Information theory for machine learning
- Variational inference

Learning from Data

Goal: Train models from **examples** so that they can make useful predictions for **unseen** data.



X_{train}

$f(\mathbf{x};\theta)$

Learn θ

$f(\mathbf{x},\theta) \in \mathcal{H}$

New unseen data

classifier

$f(\mathbf{x};\theta_{learned})$

label

(a horse)

[Based on the assumption of regularity]

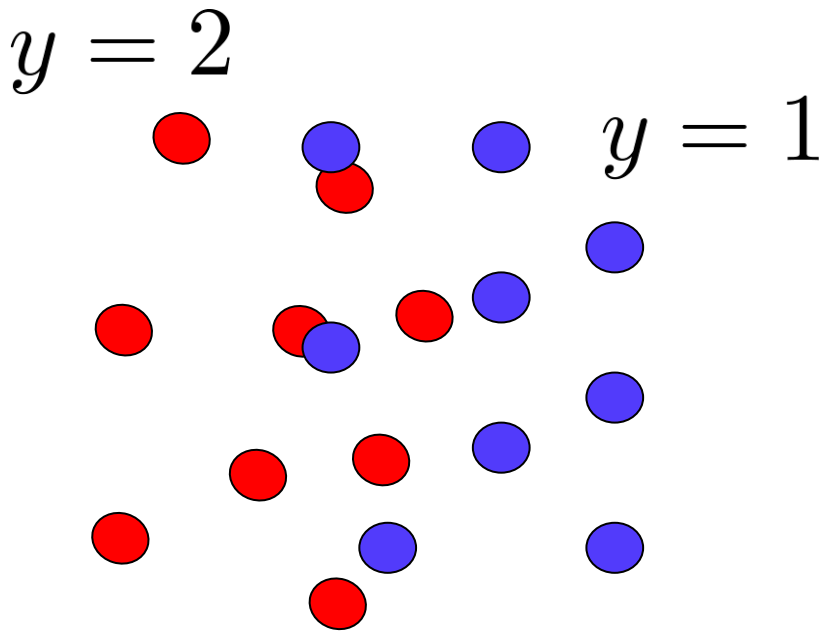
Slide credit: Prof. Yung-Kyun Noh

Learning from Data

- Data

(Regularity)

$$\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{N} \sim P$$



- Prediction

$$\mathbf{x} \in \mathbb{R}^D \xrightarrow{y = f(\mathbf{x})} \begin{array}{l} y \in \{1, 2, \ldots, C\} \\ y \in \mathbb{R} \end{array}$$

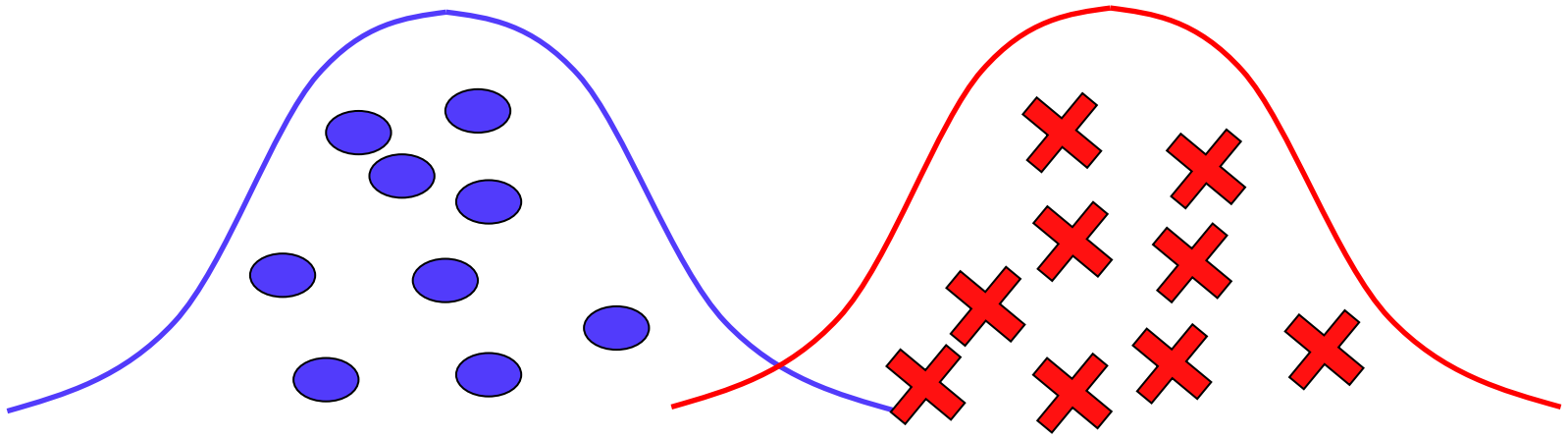
- Learning
  - Learn prediction function $f(x; \theta) \in \mathcal{H}$ from data $\mathcal{D}$

($\mathcal{H}$: Hypothesis set/Candidate set)

Slide credit: Prof. Yung-Kyun Noh

Assumptions for Learning (1/2)

- Trained data and data for prediction are ***randomly*** chosen from the ***same*** population. (= *randomly* generated from the *same* underlying density function. *Regularity* exists.).



Slide credit: Prof. Yung-Kyun Noh

Assumptions for Learning (2/2)

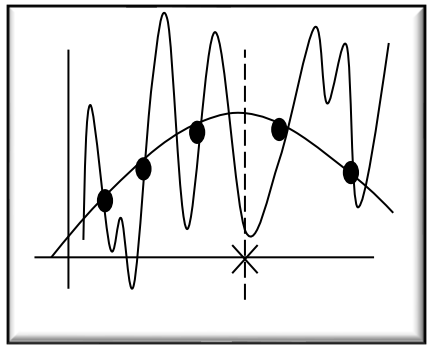
- Consider a set of candidate functions

$$\mathcal{H} = \{h_1(\mathbf{x}), h_2(\mathbf{x}), \ldots, h_{N_\mathcal{H}}(\mathbf{x})\}$$

In general, $N_\mathcal{H}$ is infinite.

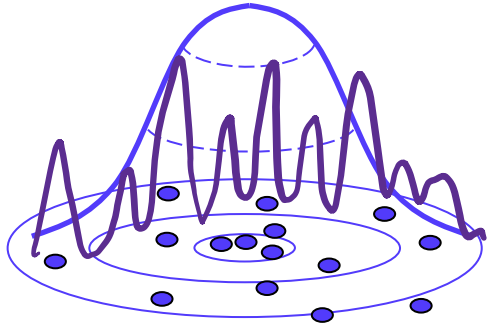
$h_i(\mathbf{x})$ can be a prediction function.

$$h_i(\mathbf{x}) \rightarrow y$$



$h_i(\mathbf{x})$ can be for probability density.

$$h_i(\mathbf{x}) \rightarrow p(\mathbf{x})$$

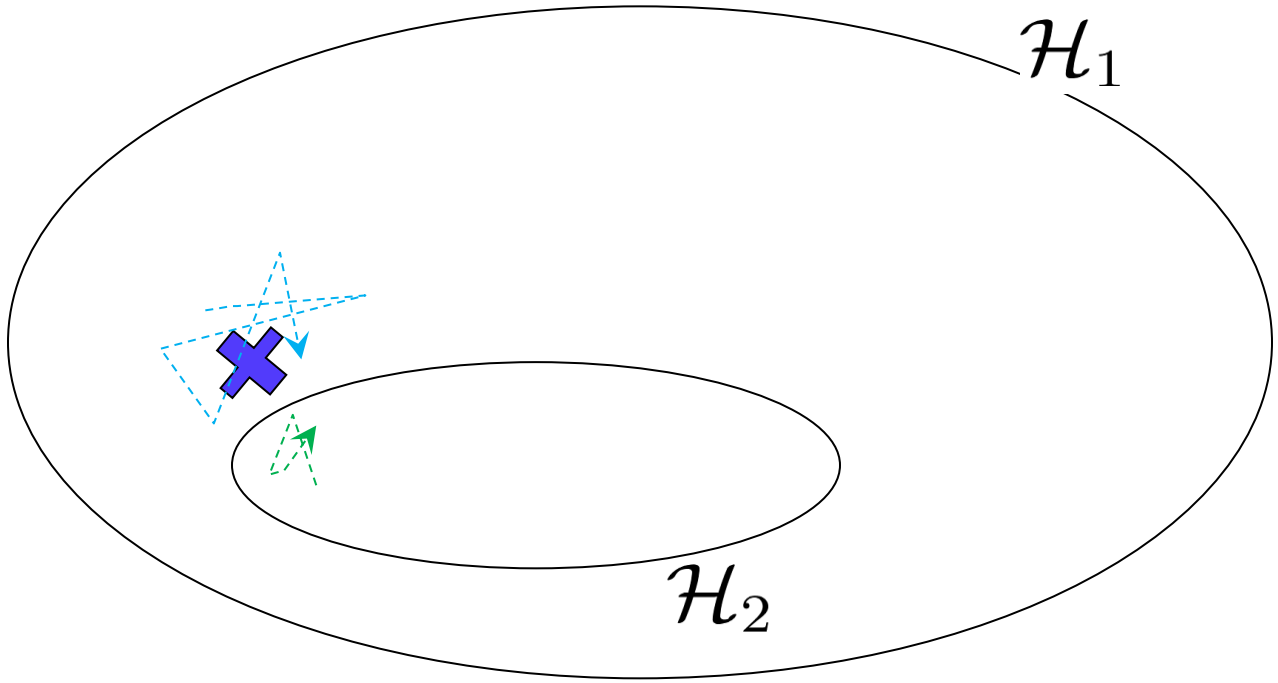


Slide credit: Prof. Yung-Kyun Noh

Assumptions for Learning (2/2)

- Confine the hypothesis space $\mathcal{H}$.

Estimation with a large number of data



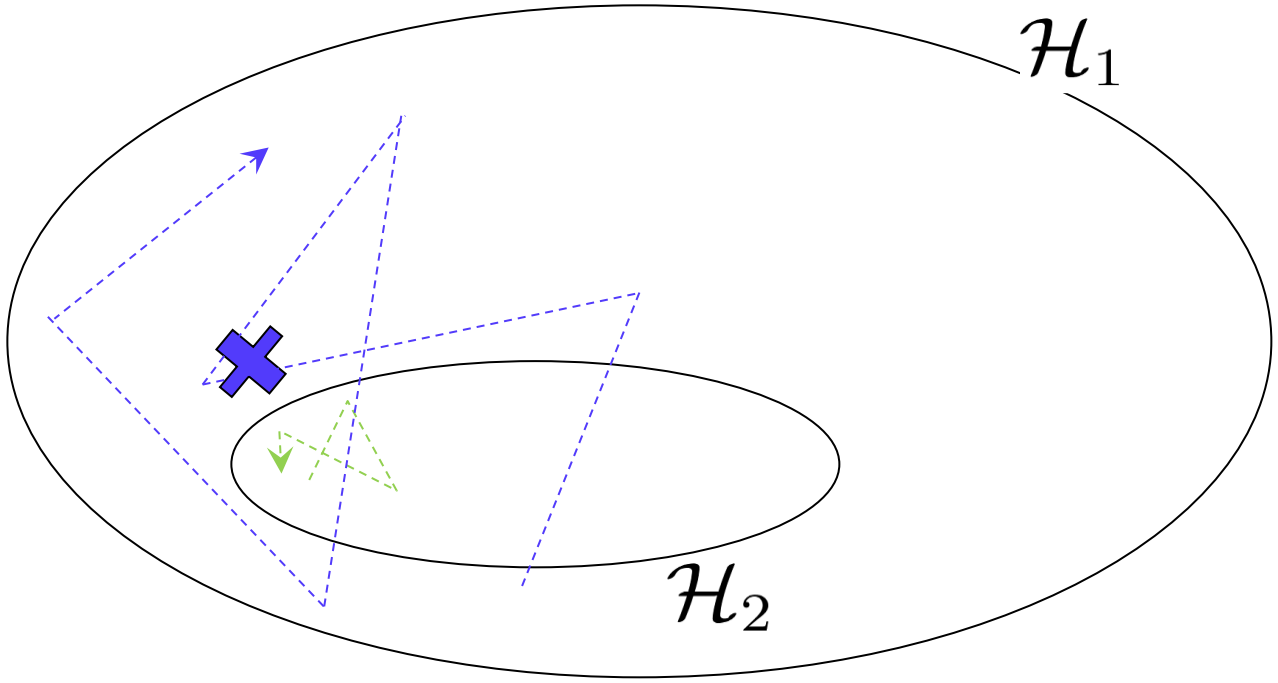
Optimal solution (✖) and the selected solutions of different realizations.

Slide credit: Prof. Yung-Kyun Noh

Assumptions for Learning (2/2)

- Confine the hypothesis space $\mathcal{H}$.

Estimation with a small number of data



Optimal solution (✖) and the selected solutions of different realizations.

Slide credit: Prof. Yung-Kyun Noh

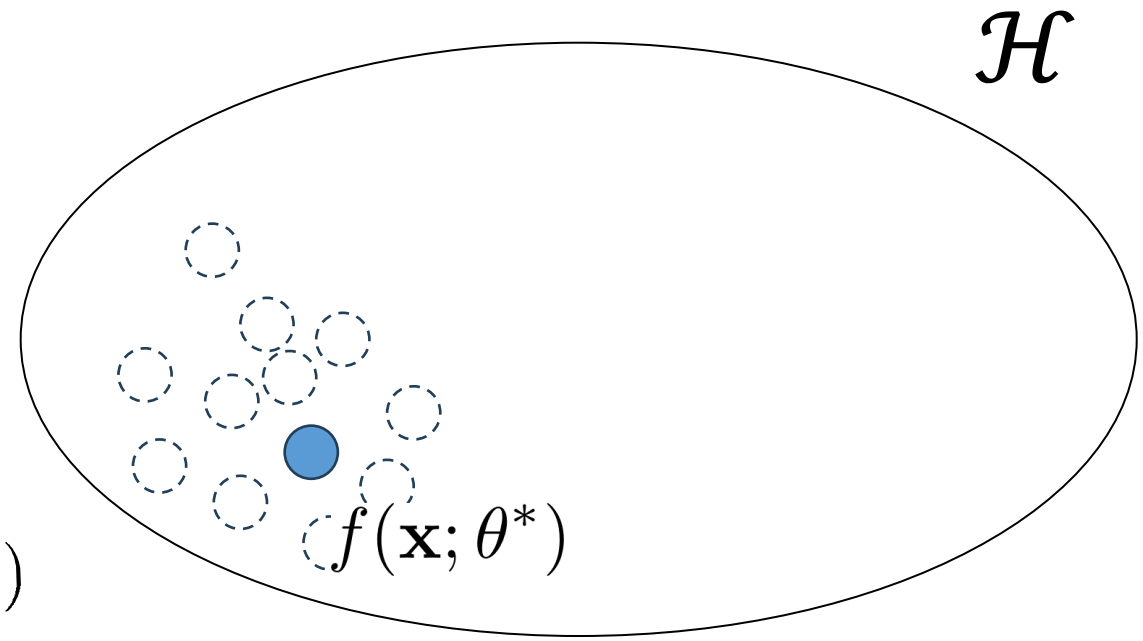
Frequentist Methods (Optimization)

- For given model $\mathcal{H} = \left\{ f(\mathbf{x};\theta) \;\middle|\; \theta \in \Theta \right\}$

Find $f(\mathbf{x};\theta^*)$ s.t.

$$\theta^* = \arg\min_{\theta} L\left(f(\mathbf{x};\theta), \mathcal{D}\right)$$

$$\theta^* = \arg\min_{\theta\in\Theta} \sum_{i=1}^{N} \left(f(\mathbf{x}_i;\theta) - y_i\right)^2 + \lambda\ \Omega(\theta)$$



Slide credit: Prof. Yung-Kyun Noh

Bayesian Methods (Integration)

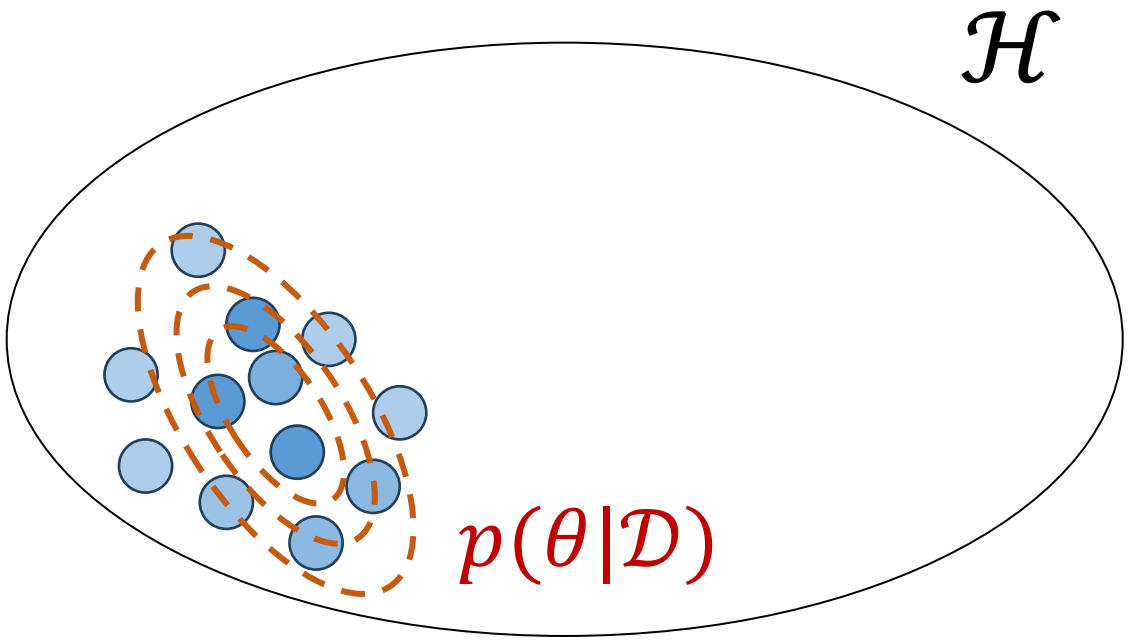
- For given model $\mathcal{H} = \left\{ f(\mathbf{x};\theta) \;\middle|\; \theta \in \Theta \right\}$

$$\widehat{y} = \int y \; p(y|\mathbf{x},\mathcal{D}) dy$$

$$= \iint y \; p(y|\mathbf{x},\theta) p(\theta|\mathcal{D}) d\theta dy$$

$f(\mathbf{x};\theta)$

Weight for each model



Slide credit: Prof. Yung-Kyun Noh

Introduction to Bayesian Inference

Bayesian Framework

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)\, p(\theta)}{p(\mathcal{D})}$$

θ: parameters

$\mathcal{D}$: data

$$posterior = \frac{likelihood \times prior}{evidence}$$

Probabilistic Modeling of Data

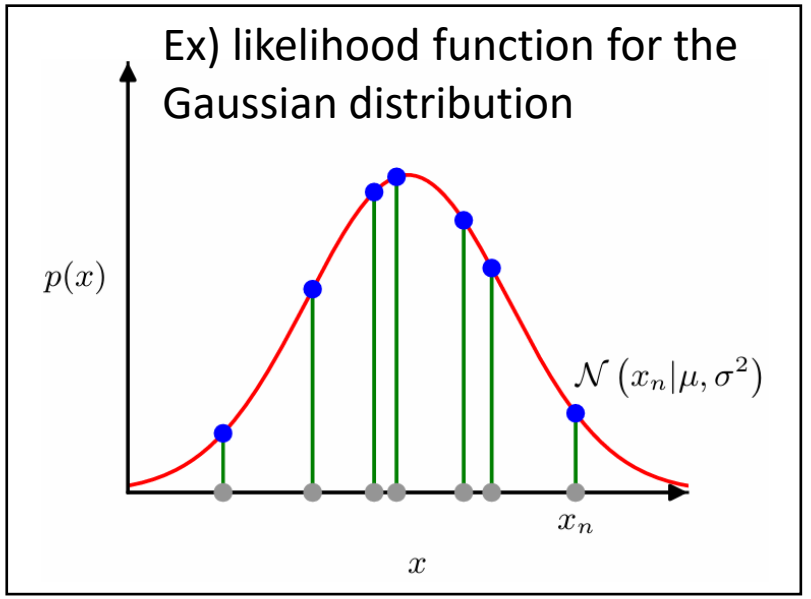
- The likelihood function

$$p(\mathcal{D}|\theta) \in \mathcal{H}$$

(The likelihood of parameter θ given data $\mathcal{D}$.)

- Maximum likelihood estimation (MLE):

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\, p(\mathcal{D}|\theta)$$

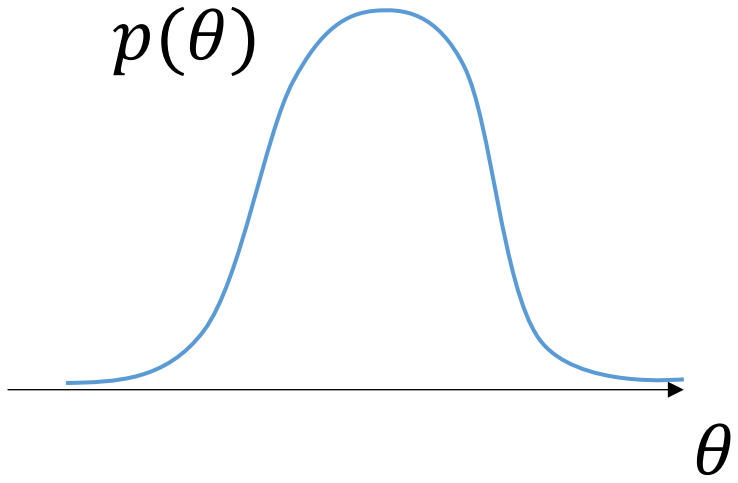


Many loss functions we minimize can be interpreted as $L(\theta) = -\log p(\mathcal{D}|\theta)$ + some regularizer.

Figures from C. Bishop (2023)

Bayesian Model

- Prior: Introduce an **assumption** (prior distribution $p(\theta)$) for parameter θ.



Q. What does it mean by the probability of θ?

A. The **degree of belief** for the hypothesis corresponding to θ.

Bayesian Model

- Bayes theorem:

$$p(\theta|\mathcal{D},\mathcal{H}) = \frac{p(\mathcal{D}|\theta,\mathcal{H})\, p(\theta|\mathcal{H})}{p(\mathcal{D}|\mathcal{H})}$$

$$p(\mathcal{D}|\mathcal{H}) = \int p(\mathcal{D}|\theta,\mathcal{H})\, p(\theta|\mathcal{H}) d\theta$$

θ: unknown parameters

$\mathcal{D} = \{x_i, y_i\}_{i=1}^{N}$: data

$\mathcal{H}$: the hypothesis space (often omitted)

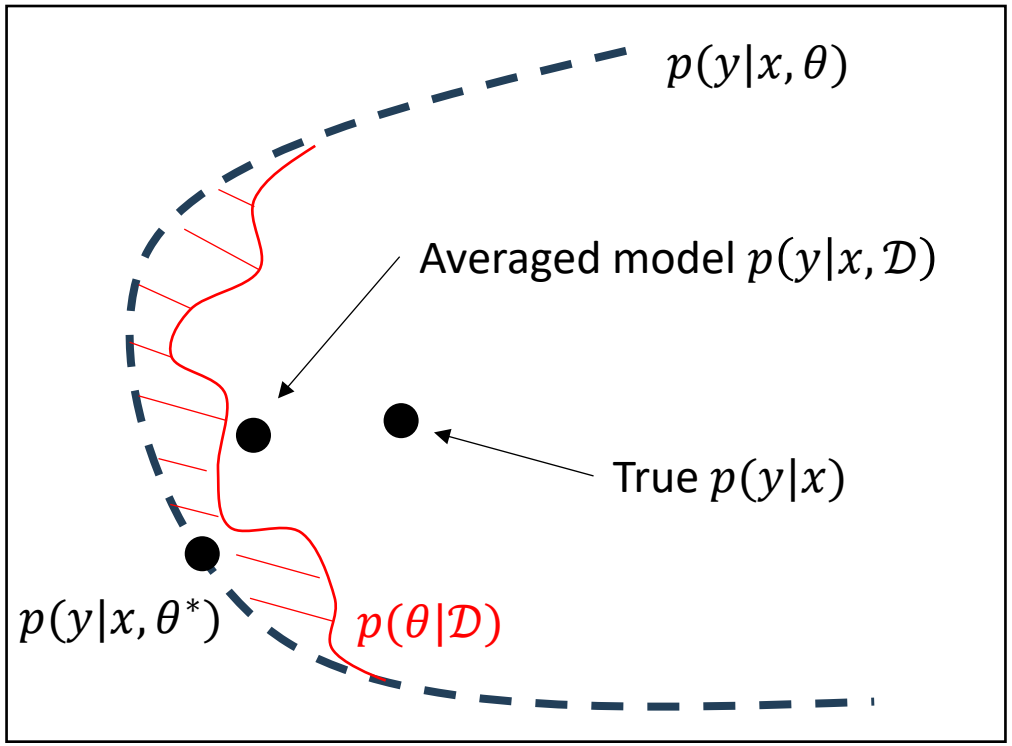
Prior $p(\theta|\mathcal{H})$: our degree of belief for θ **before observing the data.**
Posterior $p(\theta|\mathcal{D},\mathcal{H})$: our degree of belief for θ **after observing the data.**
Evidence $p(\mathcal{D}|\mathcal{H})$ **(a.k.a. marginal likelihood):** measures how well a model set describes the data.

Making Predictions

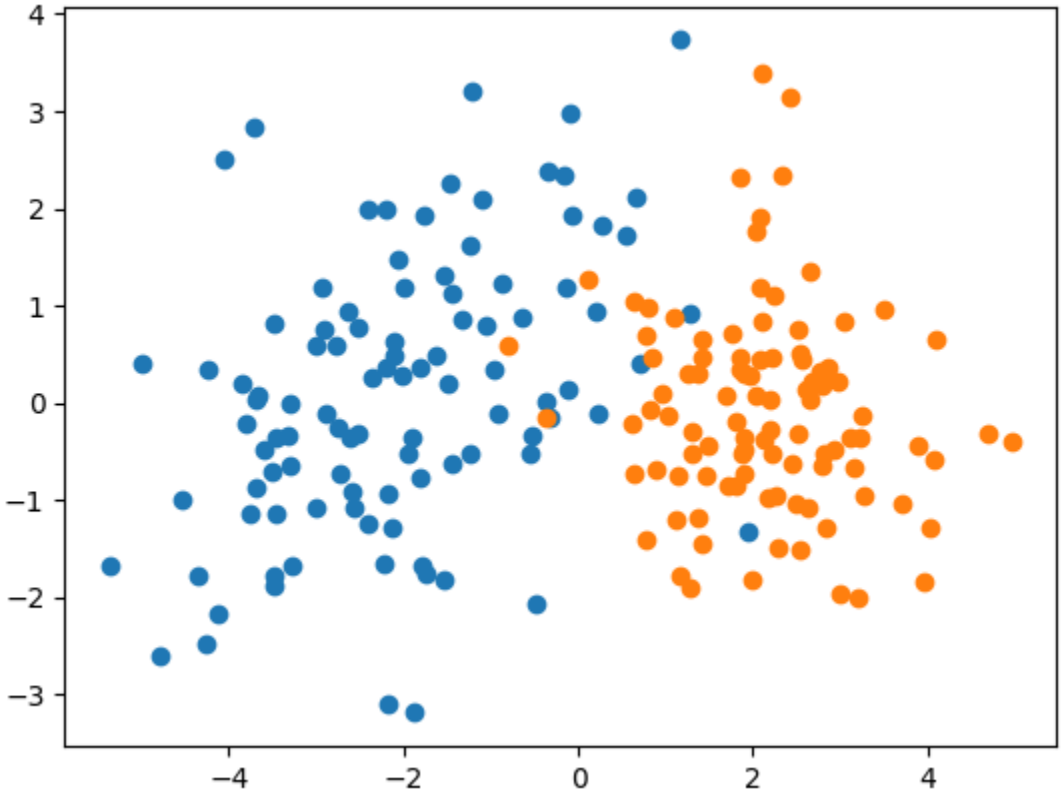
- Weigh the predictions from each model according to the posterior.

$$p(y|x,\mathcal{D}) = \int p(y|x,\theta)p(\theta|\mathcal{D})\,d\theta$$

Space of $p(y|x)$

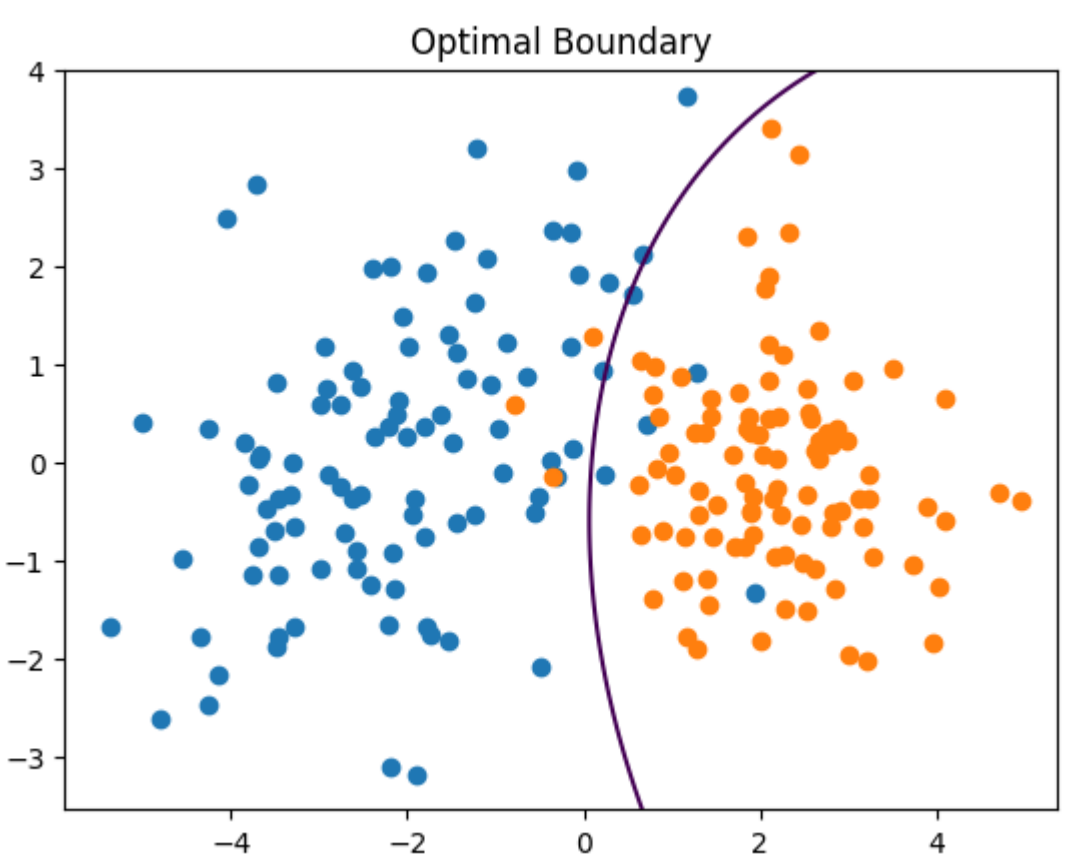


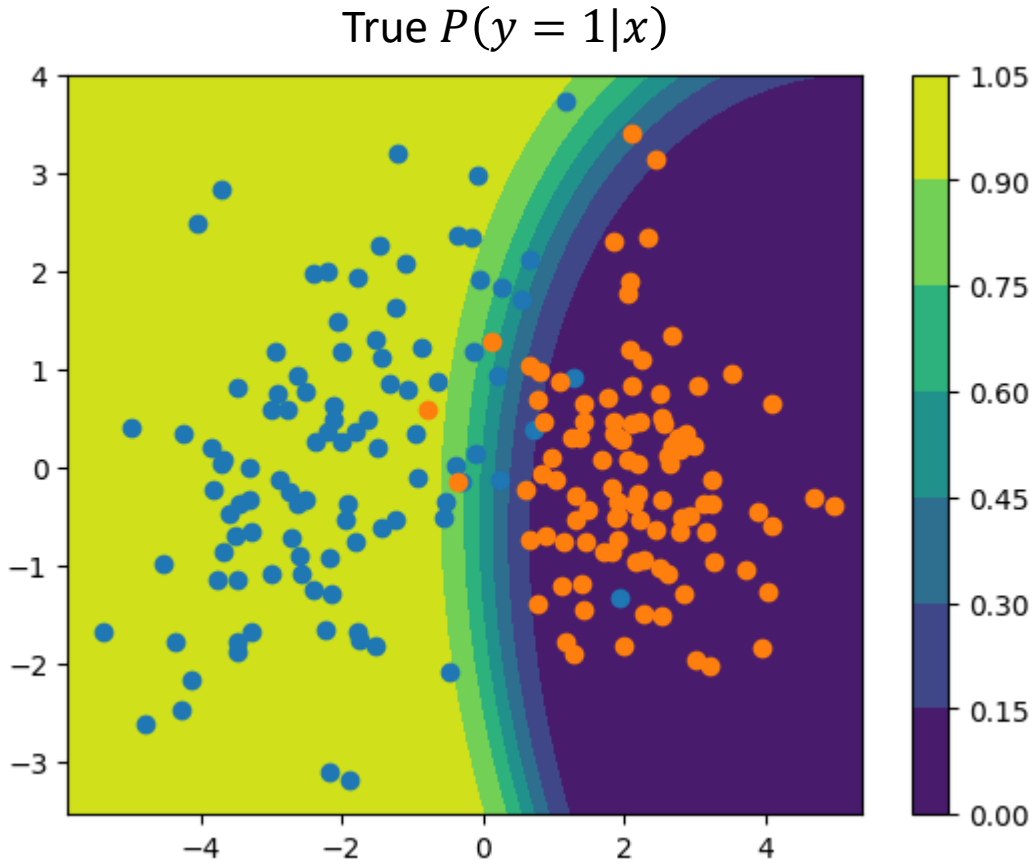
Bayesian Logistic Regression Example



Slide credit: Prof. Yung-Kyun Noh

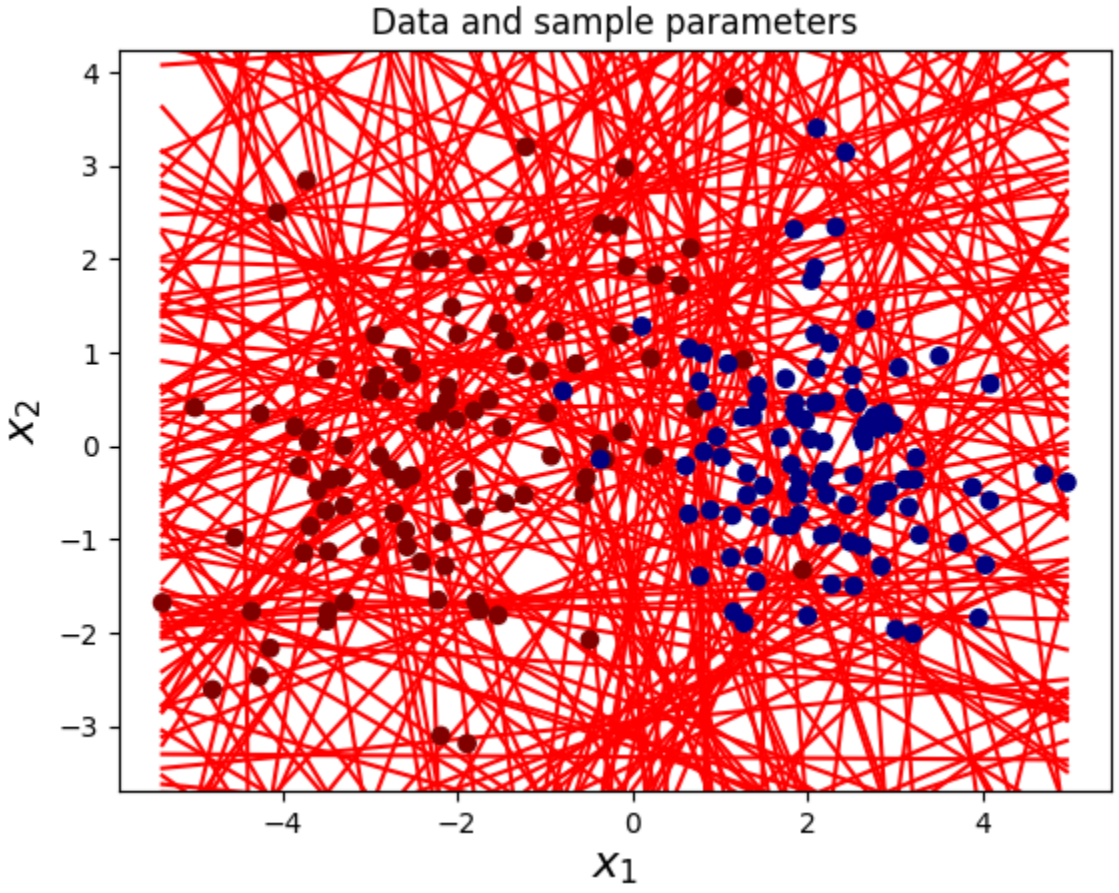
Bayesian Logistic Regression Example

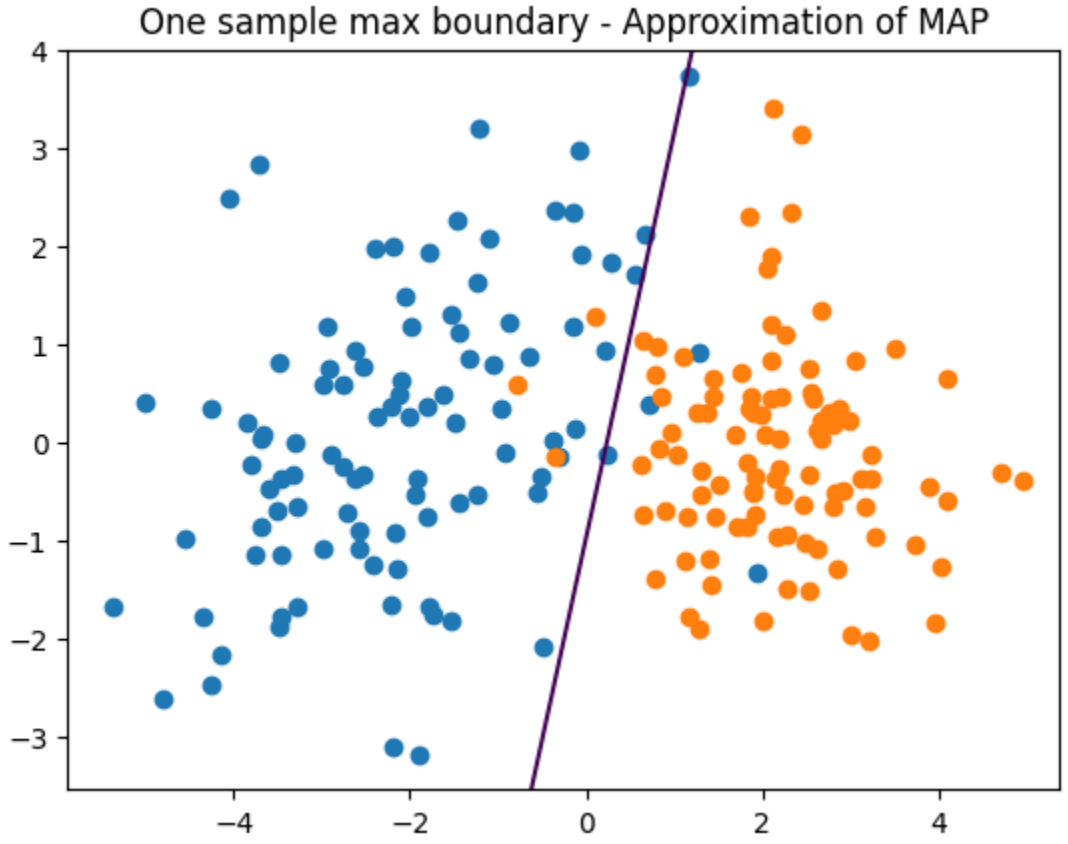




Slide credit: Prof. Yung-Kyun Noh

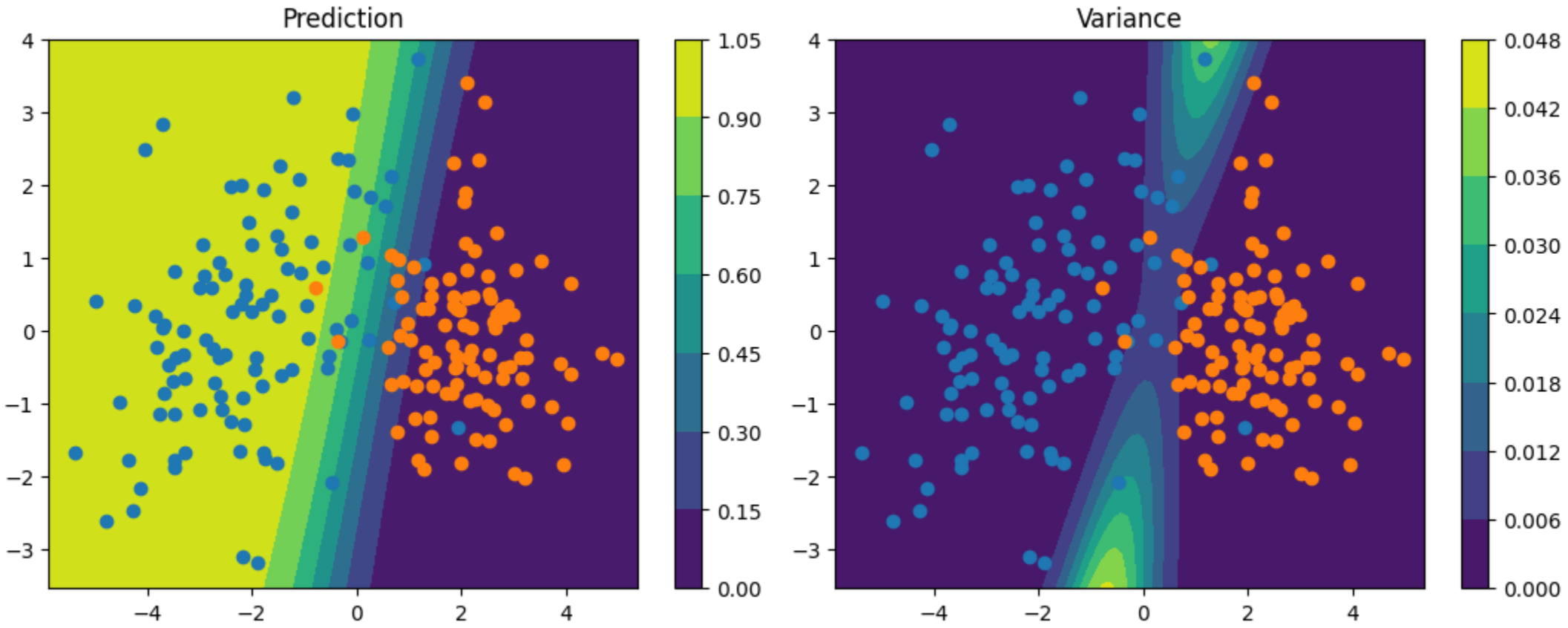
Bayesian Logistic Regression Example





Slide credit: Prof. Yung-Kyun Noh

Bayesian Logistic Regression Example



Slide credit: Prof. Yung-Kyun Noh

Benefits of Bayesian Inference

- Once assumptions (the choice of prior) are made, the inferences are objective and unique, reproducible with complete agreement.
- When the assumptions are explicit, they are easier to criticize, easier to modify.
- When it is not sure which assumption is the most appropriate, we can treat it as another inference task.

$$p(\mathcal{H}|\mathcal{D}, I) = \frac{p(\mathcal{D}|\mathcal{H}, I)\, p(\mathcal{H}|I)}{p(\mathcal{D}|I)}$$

I: the highest assumption, which we are not questioning.

- We can take into account our uncertainty regarding such assumptions when we make subsequent predictions.

$$p(y|\mathcal{D}, I) = \sum_{\mathcal{H}} p(y|\mathcal{D}, \mathcal{H}, I) p(\mathcal{H}|\mathcal{D}, I)$$

D. MacKay, Information theory, inference, and learning algorithms (2003)

Gaussian Distribution and Parameter Estimation

Multivariate Gaussian Distribution

$$p(x;\mu,\Sigma) = \frac{1}{\sqrt{2\pi}^{D}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x-\mu)^{\top}\Sigma^{-1}(x-\mu)\right)$$

$x = (x_1, x_2, \ldots, x_D) \in \mathbb{R}^D$
$\mu \in \mathbb{R}^D$
$\Sigma \in \mathbb{R}^{D\times D}$: symmetric, positive-definite.

$$\mathbb{E}[x_j] = \mu_j$$

$$Cov[x_i, x_j] = \mathbb{E}[x_i x_j] - \mathbb{E}[x_i]\mathbb{E}[x_j] = \Sigma_{ij}$$

Two-dimensional case:

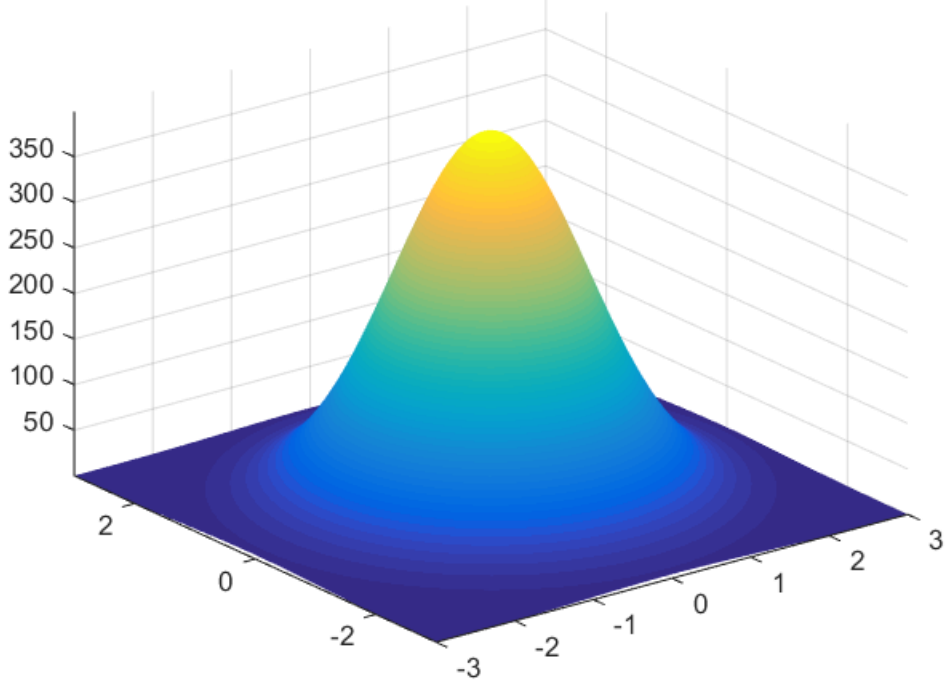


Figure from https://brilliant.org/wiki/multivariate-normal-distribution/

Multivariate Gaussian Distribution

$$p(x;\mu,\Sigma) = \frac{1}{\sqrt{2\pi}^{D}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^{\top}\Sigma^{-1}(x-\mu)\right)$$

The eigendecomposition of Σ:

$$\Sigma u_i = \lambda_i u_i$$

Red ellipse: The contour of

$$p(x;\mu,\Sigma) = p(\mu;\mu,\Sigma)\cdot \exp\left(-\frac{1}{2}\right)$$

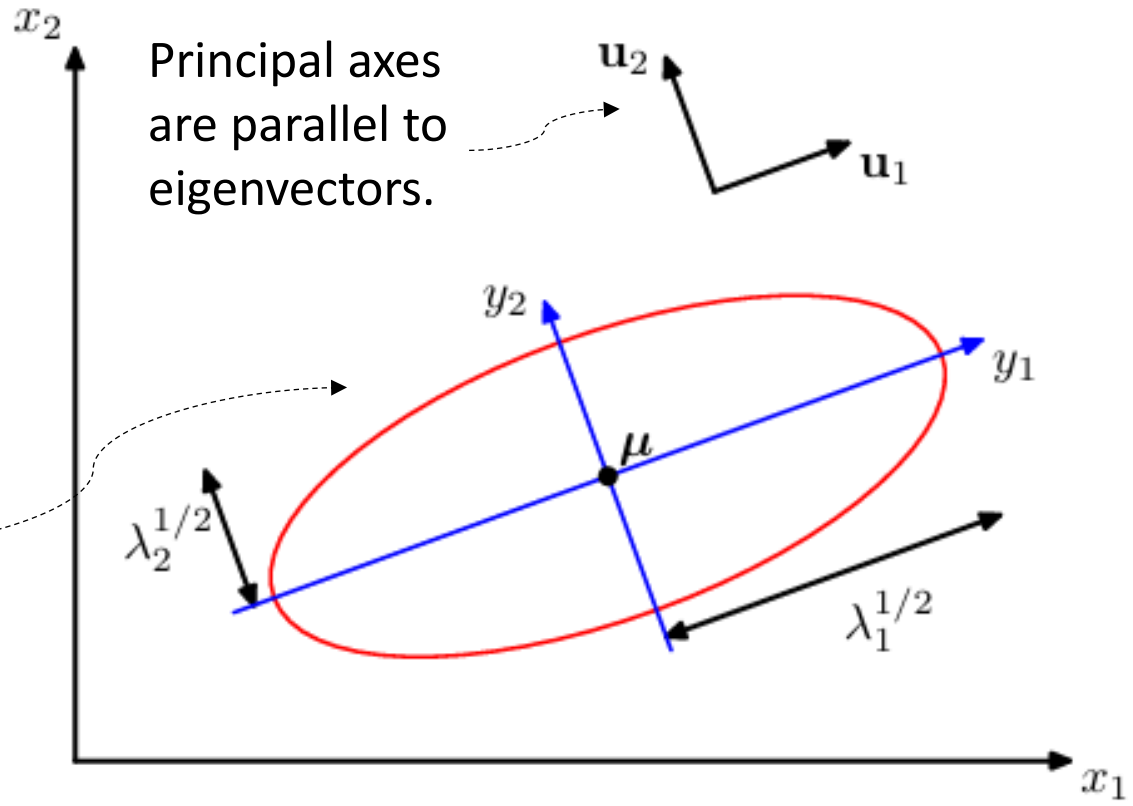
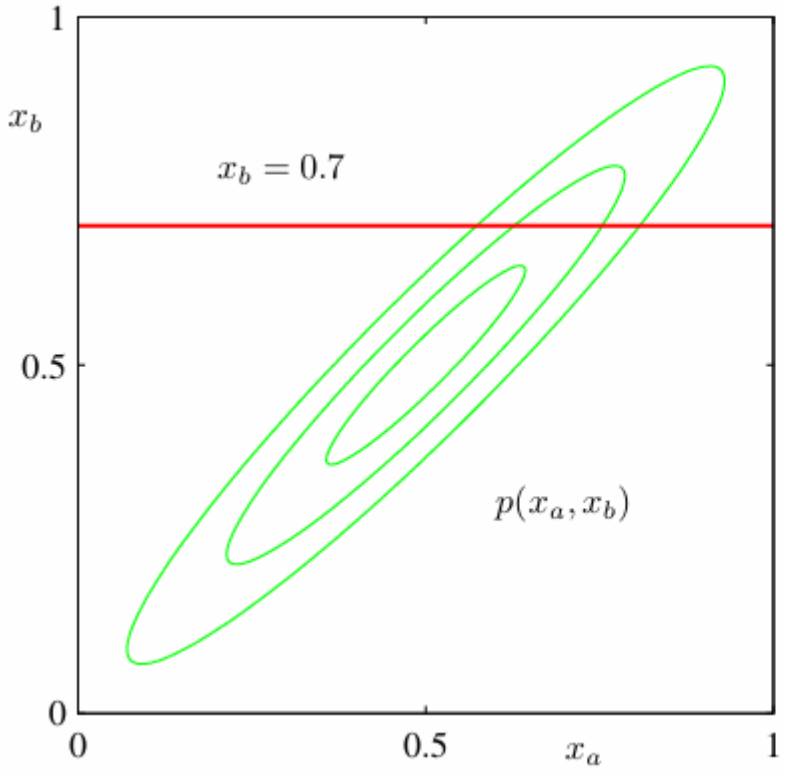
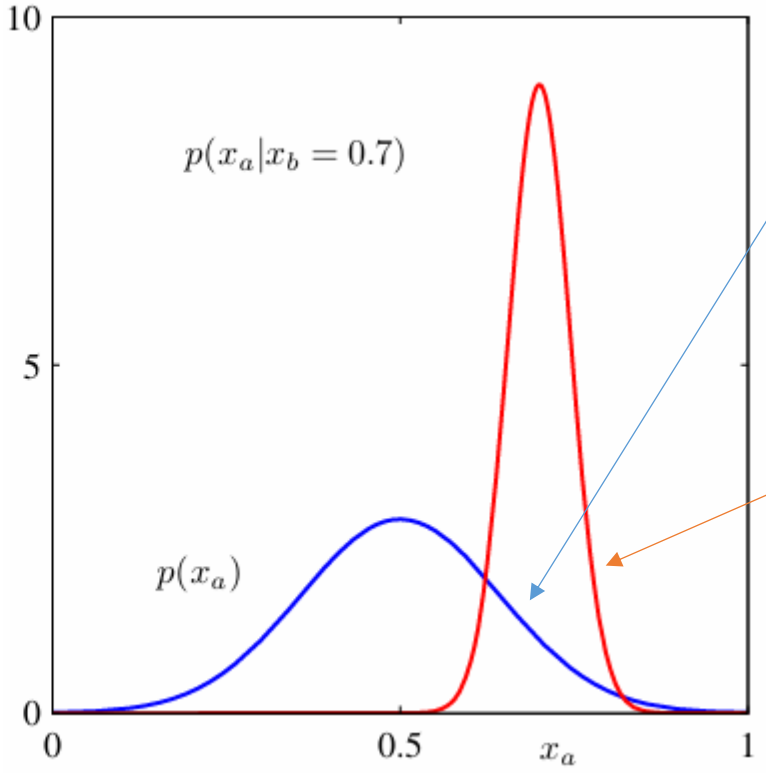


Figure from C. Bishop (2006)

Marginal and Conditional Gaussian

- For a Gaussian random variable $x = (x_a, x_b) \in \mathbb{R}^D, x_a \in \mathbb{R}^{Da}, x_b \in \mathbb{R}^{Db}$, the **marginal** and **conditional** distributions are **also Gaussian.**





$$p(x_a) = \int p(x_a, x_b) dx_b$$

$$p(x_a|x_b) = \frac{p(x_a, x_b)}{p(x_b)}$$

Figures from C. Bishop (2006)

Marginal and Conditional Gaussian

- For a Gaussian random variable $x = (x_a, x_b) \in \mathbb{R}^D, x_a \in \mathbb{R}^{Da}, x_b \in \mathbb{R}^{Db}$, the **marginal** and **conditional** distributions are **also Gaussian.**

$$p(x_a, x_b) = N(\mu, \Sigma)$$
$$\propto \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$

$$\mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} \in \mathbb{R}^D \qquad \Sigma = \begin{pmatrix} \Sigma_{aa} \Sigma_{ab} \\ \Sigma_{ba} \Sigma_{bb} \end{pmatrix} \in \mathbb{R}^{D \times D}$$

A useful fact (for deriving the above results) which can be obtained from **Schur's complement**:

$$-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)$$
$$= -\frac{1}{2}(x_b - \mu_b)^\top \Sigma_{bb}^{-1}(x_b - \mu_b) - \frac{1}{2}\left(x_a - \mu_{a|b}\right)^\top \Sigma_{a|b}^{-1}(x_a - \mu_{a|b})$$

Marginal and Conditional Gaussian

- Let $\Sigma^{-1} = \begin{bmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{bmatrix}$.

$$-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu) = -\frac{1}{2}\begin{pmatrix} x_a-\mu_a \\ x_b-\mu_b \end{pmatrix}^\top \begin{bmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{bmatrix} \begin{pmatrix} x_a-\mu_a \\ x_b-\mu_b \end{pmatrix}$$

$$= -\frac{1}{2}(x_a-\mu_a)^\top \Lambda_{aa}(x_a-\mu_a) - \frac{1}{2}(x_a-\mu_a)^\top \Lambda_{ab}(x_b-\mu_b) - \frac{1}{2}(x_b-\mu_b)^\top \Lambda_{ba}(x_a-\mu_a)$$
$$-\frac{1}{2}(x_b-\mu_b)^\top \Lambda_{bb}(x_b-\mu_b)$$

$$= -\frac{1}{2}\left(x_a-\mu_a+\Lambda_{aa}^{-1}\Lambda_{ab}(x_b-\mu_b)\right)^\top \Lambda_{aa}\left(x_a-\mu_a+\Lambda_{aa}^{-1}\Lambda_{ab}(x_b-\mu_b)\right) + \frac{1}{2}(x_b-\mu_b)^\top \Lambda_{ba}\Lambda_{aa}^{-1}\Lambda_{ab}(x_b-\mu_b)$$
$$-\frac{1}{2}(x_b-\mu_b)^\top \Lambda_{bb}(x_b-\mu_b)$$

$$= -\frac{1}{2}(x_a-\mu_{a|b})^\top \Sigma_{a|b}^{-1}(x_a-\mu_{a|b}) - \frac{1}{2}(x_b-\mu_b)^\top \Sigma_{bb}^{-1}(x_b-\mu_b)$$

From **Schur's complement,**

$$\mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b-\mu_b)$$
$$\Sigma_{a|b} = \Lambda_{aa}^{-1} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$$
$$\Sigma_{bb} = (\Lambda_{bb} - \Lambda_{ba}\Lambda_{aa}^{-1}\Lambda_{ab})^{-1}$$

Marginal and Conditional Gaussian

- For a Gaussian random variable $x = (x_a, x_b) \in \mathbb{R}^D, x_a \in \mathbb{R}^{Da}, x_b \in \mathbb{R}^{Db}$, the **marginal** and **conditional** distributions are **also Gaussian.**

$$-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu) = -\frac{1}{2}(x_b-\mu_b)^\top \Sigma_{bb}^{-1}(x_b-\mu_b) - \frac{1}{2}\left(x_a-\mu_{a|b}\right)^\top \Sigma_{a|b}^{-1}(x_a-\mu_{a|b})$$

Marginal distribution:

$$p(x_b) = \int p(x_a, x_b)dx_a = N(x_b; \mu_b, \Sigma_b) \qquad \Sigma_b = \Sigma_{bb}$$

$$\propto \exp\left(-\frac{1}{2}(x_b-\mu_b)^\top \Sigma_{bb}^{-1}(x_b-\mu_b)\right)$$

Conditional distribution:

$$p(x_a|x_b) = \frac{p(x_a, x_b)}{p(x_b)} = N(x_a; \mu_{a|b}, \Sigma_{a|b}) \qquad \mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b-\mu_b)$$

$$\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$$

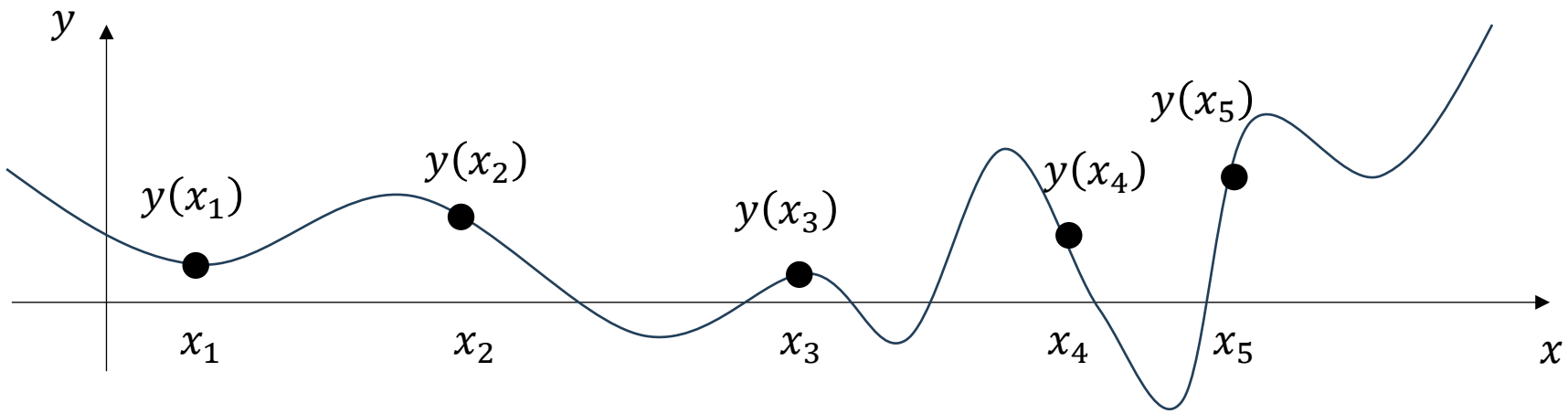
$$\propto \exp\left(-\frac{1}{2}\left(x_a-\mu_{a|b}\right)^\top \Sigma_{a|b}^{-1}(x_a-\mu_{a|b})\right)$$

Gaussian Processes

- A function space view

$$y \sim GP(m, k)$$

$m(x)$: mean function
$k(x_1, x_2)$: covariance function
$y(x)$: a random function

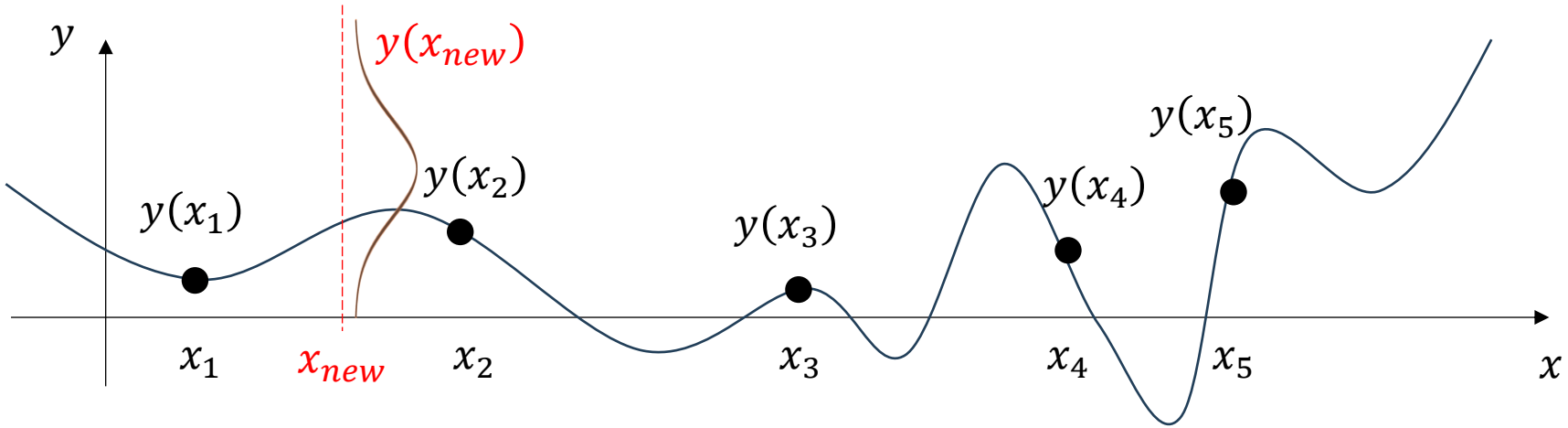


$y(x_1), \ldots, y(x_N)$ are jointly Gaussian with mean $\big(m(x_1), \ldots, m(x_N)\big)^\top \in \mathbb{R}^N$

and covariance $K = \begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_N) \\ \vdots & \ddots & \vdots \\ k(x_N, x_1) & \cdots & k(x_N, x_N) \end{pmatrix} \in \mathbb{R}^{N \times N}$.

Gaussian Processes

- Gaussian process regression



Let $m(x) = 0$. **Joint PDF:** $\mathbb{E}\left[\left(y(x_{new}), y(x_1), \dots, y(x_N)\right)\right] = (0, \dots, 0) \in \mathbb{R}^{N+1}$

$$Cov\left(y(x_{new}), y(x_1), \dots, y(x_N)\right) = \begin{pmatrix} \kappa & k^\top \\ k & K \end{pmatrix} \in \mathbb{R}^{(N+1)\times(N+1)}$$

Recall $\mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b)$

$$\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$$

➡ **Conditional PDF:**

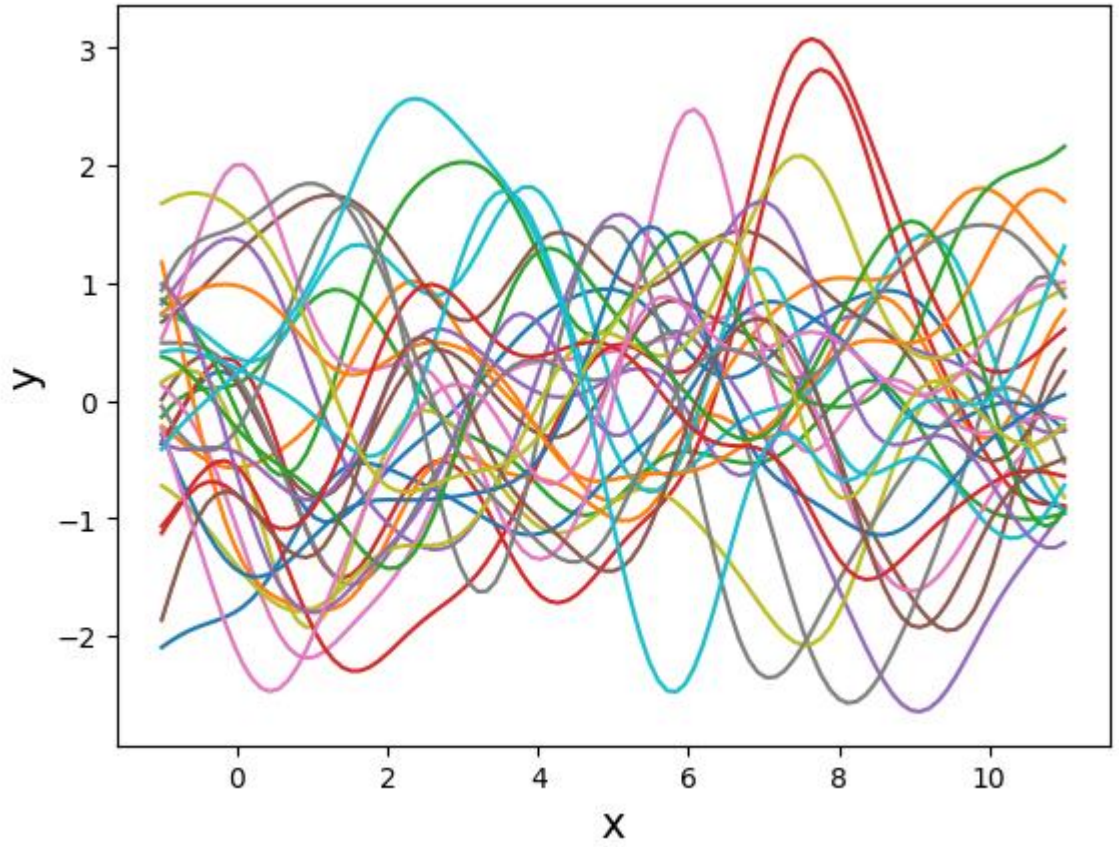
$$\mathbb{E}[y(x_{new}) \mid y(x_1), \dots, y(x_N)] = k^\top K^{-1} y \in \mathbb{R}$$

$$Cov[y(x_{new}) \mid y(x_1), \dots, y(x_N)] = \kappa - k^\top K^{-1} k \in \mathbb{R}$$

$$p(y(x_{new}) | y(x_1), \dots, y(x_N)) = N(k^\top K^{-1} y, \kappa - k^\top K^{-1} k)$$

prediction uncertainty

Gaussian Process Examples

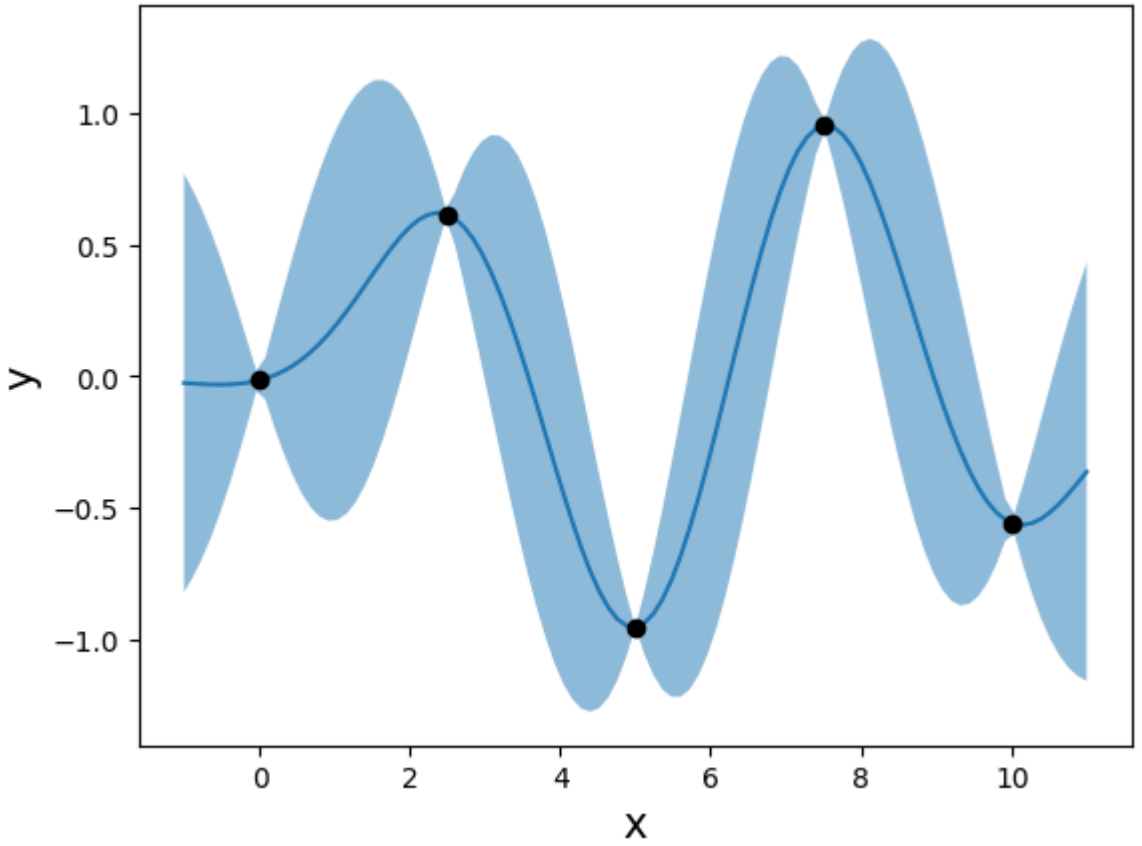


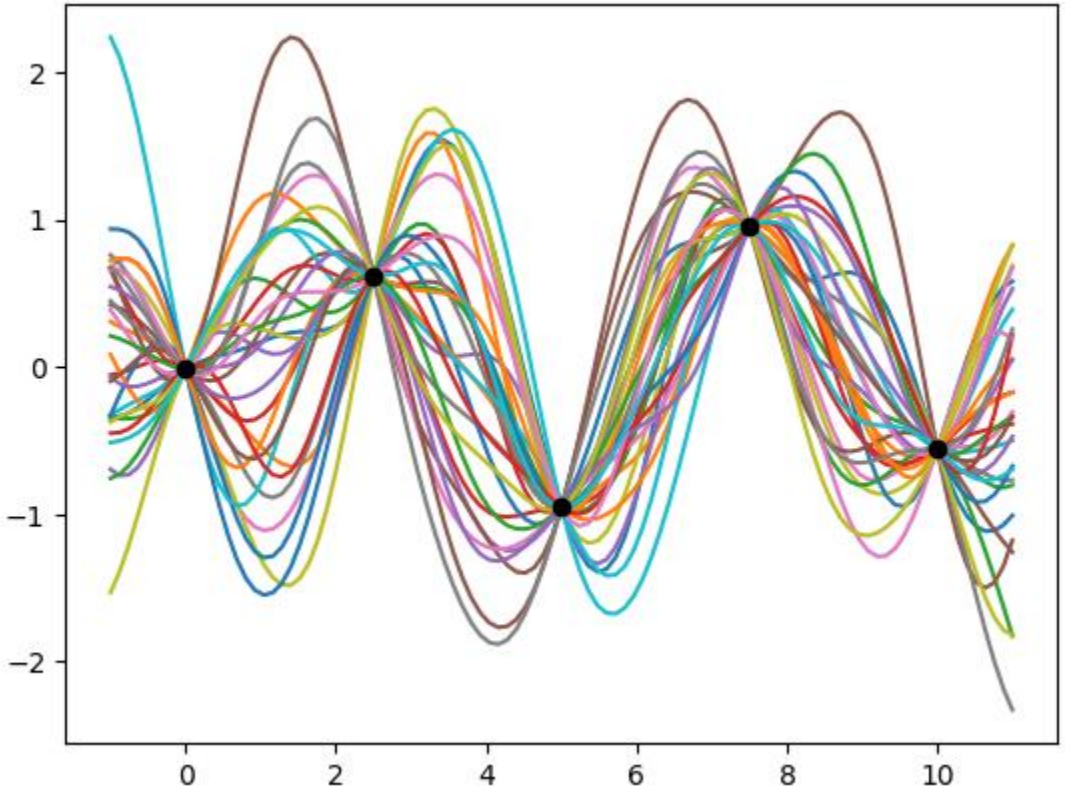
$$k(x, y) = \exp\left(-\frac{(x-y)^2}{2h^2}\right)$$

$$h = 1.0$$

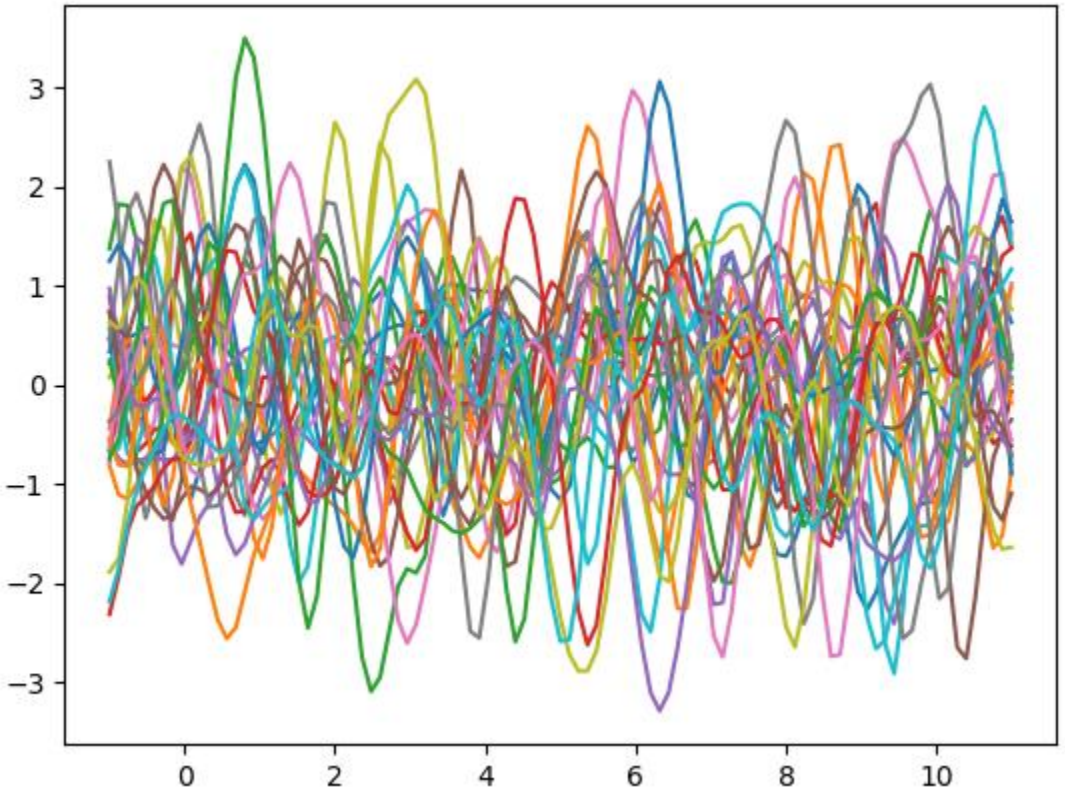
Gaussian Process Examples

- Regression





Gaussian Process Examples

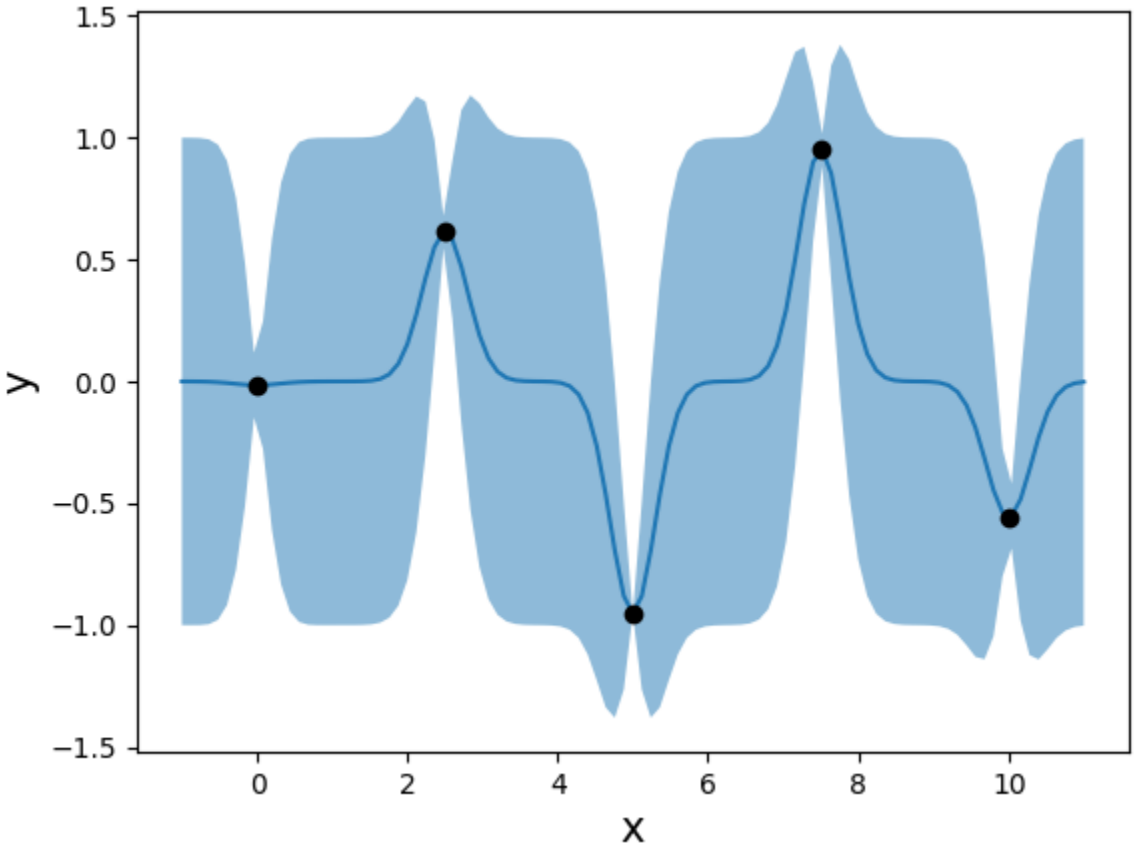


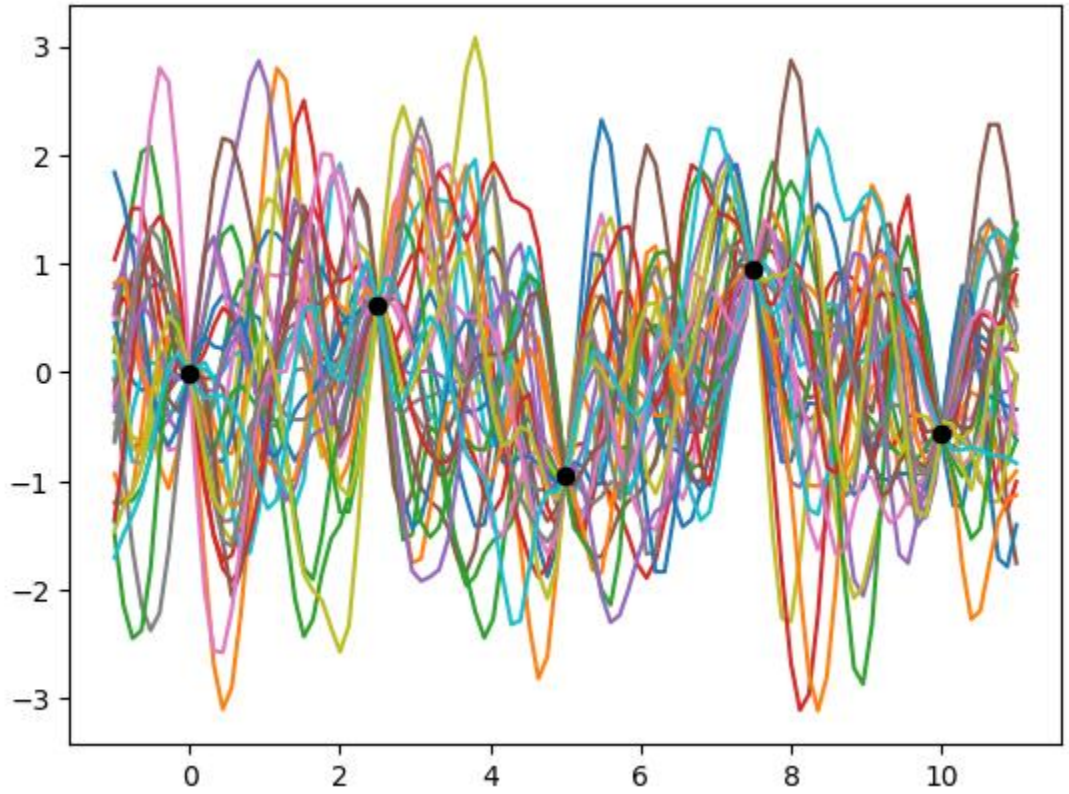
$$k(x, y) = \exp\left(-\frac{(x-y)^2}{2h^2}\right)$$

$$h = 0.3$$

Gaussian Process Examples

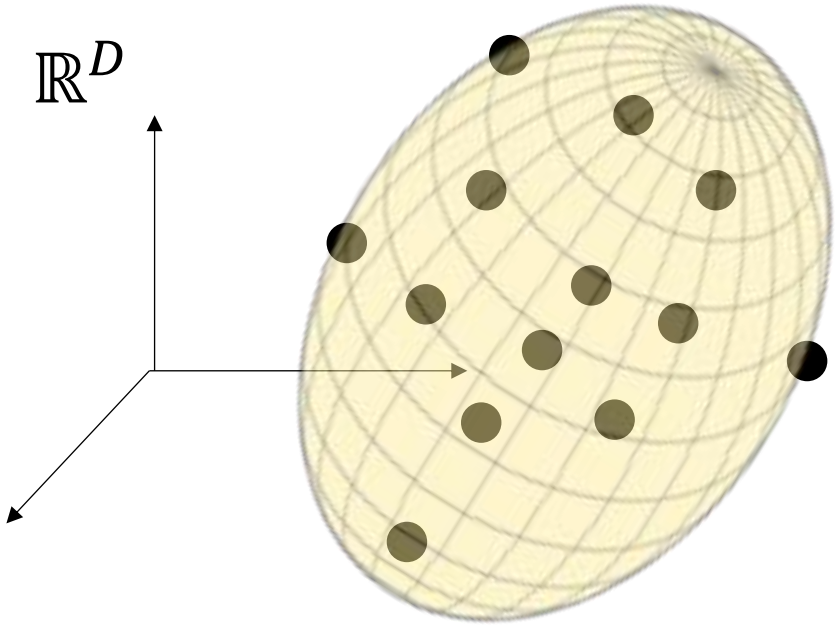
- Regression





Parameter Estimation

- Suppose data points $\mathcal{D} = \{x_i\}_{i=1}^N, x_i \in \mathbb{R}^D$ are given.



Let's consider using a parametric model $p(x;\theta)$, e.g., Gaussian model, to fit the data!

$\theta \in \mathbb{R}^p$: parameters

Maximum Likelihood Estimation

- Likelihood (i.i.d. assumption for data):

$$L(\theta) = p(\{x_i\}_{i=1}^N; \theta) = \prod_{i=1}^{N} p(x_i; \theta)$$

- Maximum likelihood estimation:

$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}} \log L(\theta) = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} \log p(x_i; \theta)$$

The necessary condition:

$$\left.\frac{\partial \log L(\theta)}{\partial \theta}\right|_{\theta=\widehat{\theta}_{ML}} = 0$$

Maximum Likelihood Estimation

- Example: Gaussian model with $\theta = (\mu, \Sigma)$

$$\max_{\theta} \log L(\theta) = \sum_{i=1}^{N} \log p(x_i; \mu, \Sigma)$$

$$= \sum_{i=1}^{N} -\frac{1}{2}(x_i - \mu)^\top \Sigma^{-1}(x_i - \mu) - \frac{D}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma|$$

$$p(x; \mu, \Sigma) = \frac{1}{\sqrt{2\pi}^D |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu)\right)$$

The necessary conditions:

$$\frac{\partial \log L}{\partial \mu} = \sum_{i=1}^{N} \Sigma^{-1}(x_i - \mu) = 0 \quad \Rightarrow \quad \hat{\mu}_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\frac{\partial \log L}{\partial \Sigma} = 0$$

Derive this as an exercise ☺

$$\Rightarrow \quad \hat{\Sigma}_{ML} = \frac{1}{N}\sum_{i=1}^{N} (x_i - \hat{\mu}_{ML})(x_i - \hat{\mu}_{ML})^\top$$

Useful facts for the derivation: $\frac{\partial}{\partial \Sigma}(x^\top \Sigma^{-1} x) = -\Sigma^{-1} x x^\top \Sigma^{-1}$, $\frac{\partial}{\partial \Sigma}(\log|\Sigma|) = \Sigma^{-1}$.

Maximum a Posteriori Estimation

- Bayes rule: $p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)\, p(\theta)$

- Maximum a posteriori (MAP) estimation:

$$\hat{\theta}_{MAP} = \operatorname*{argmax}_{\theta} p(\theta|\mathcal{D})$$

$$= \operatorname*{argmax}_{\theta} \log p(\theta|\mathcal{D})$$

$$= \operatorname*{argmax}_{\theta} \sum_{i=1}^{N} \log p(x_i|\theta) + \log p(\theta)$$

Plays a similar role to the **regularizer** in the frequentist formulations.

Maximum a Posteriori Estimation

- Example: one-dimensional Gaussian model with $\theta = \mu$ (σ^2 is fixed).

Prior: $p(\mu) = N(\mu; \mu_0, \sigma_0^2)$ Likelihood: $\prod_{i=1}^{N} p(x_i|\mu)$

$$\max_{\mu} \sum_{i=1}^{N} \log p(x_i|\mu) + \log p(\mu)$$

$$\max_{\mu} L(\mu) = \sum_{i=1}^{N} \left(-\frac{1}{2\sigma^2}(x_i - \mu)^2 \right) - \frac{1}{2\sigma_0^2}(\mu - \mu_0)^2$$

The necessary condition:

$$\left.\frac{\partial L}{\partial \mu}\right|_{\mu=\hat{\mu}_{MAP}} = 0 \quad \Rightarrow \quad \hat{\mu}_{MAP} = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right) + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0$$

Posterior Distribution

- Example: one-dimensional Gaussian model with $\theta = \mu$ (σ^2 is fixed).

  Here, we can calculate the exact posterior distribution.

$$p(\mu|D) \propto p(D|\mu)\, p(\mu) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2\right)\exp\left(-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right)$$

$$\propto \exp\left(-\frac{1}{2}\left(\left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{i=1}^{N}x_i+\frac{\mu_0}{\sigma_0^2}\right)\mu\right)\right)$$

$$\propto \exp\left(-\frac{1}{2\sigma_N^2}(\mu-\mu_N)^2\right)$$

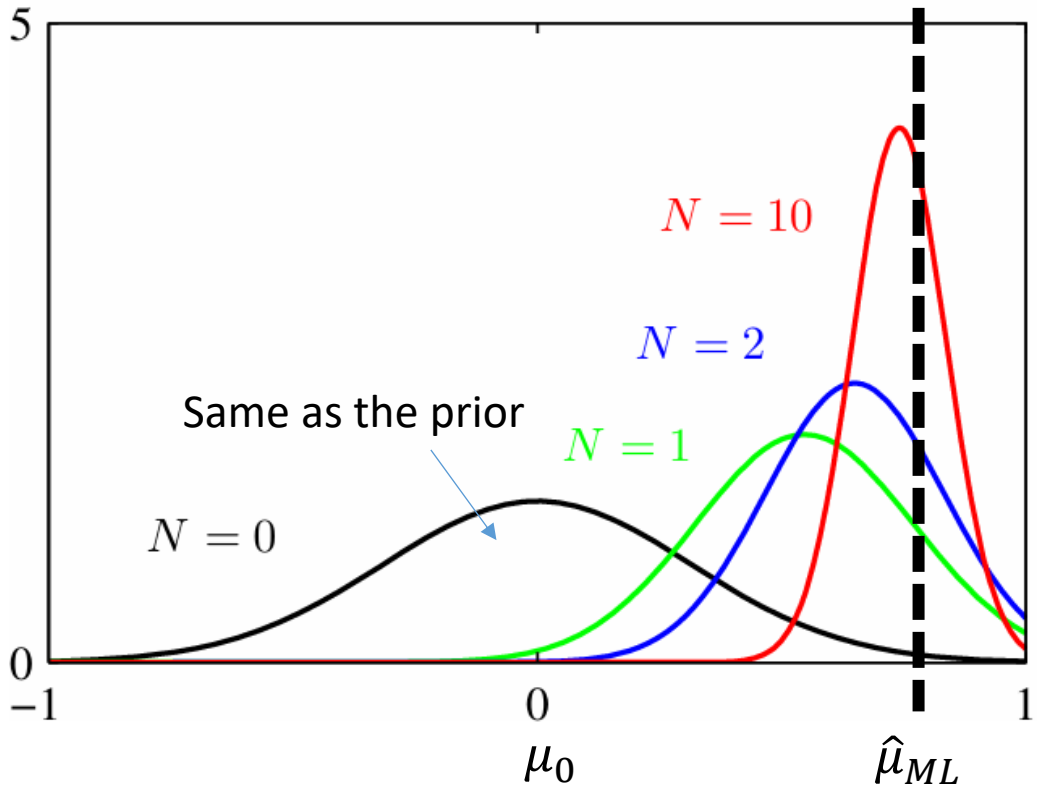
A useful fact for derivation:

$$a\mu^2 - 2b\mu = a\left(\mu-\frac{b}{a}\right)^2 + const.$$

$$p(\mu|D) = N(\mu;\mu_N,\sigma_N^2) \quad \text{where} \quad \mu_N = \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\hat{\mu}_{ML} + \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 = \hat{\mu}_{MAP}$$

$$\sigma_N^2 = \frac{\sigma^2\sigma_0^2}{N\sigma_0^2+\sigma^2}.$$

Posterior Distribution



$$\mu_N = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\hat{\mu}_{ML} + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0$$

$$\sigma_N^2 = \frac{\sigma^2\sigma_0^2}{N\sigma_0^2 + \sigma^2}$$

Some special cases:

If $N = 0$ or $\sigma_0^2 = 0$, $\mu_N = \mu_0, \sigma_N^2 = \sigma_0^2$ → Prior.

If $\sigma^2 \gg 1$, $\mu_N \approx \mu_0, \sigma_N^2 \approx \sigma_0^2$ → Close to the prior.

If $N \to \infty$, $\mu_N = \hat{\mu}_{ML}, \sigma_N^2 = 0$

→ Posterior is infinitely peaked at $\hat{\mu}_{ML}$.

If $\sigma_0^2 \to \infty$, $\mu_N = \hat{\mu}_{ML}, \sigma_N^2 = \frac{\sigma^2}{N}$ → Uninformative prior.

Figures from C. Bishop (2006)

Predictive Distribution for Gaussian

$$p(x|\mathcal{D}) = \int p(x|\mu)p(\mu|\mathcal{D})\, d\mu$$

$$p(x|\mu)p(\mu|\mathcal{D}) \propto \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)\exp\left(-\frac{1}{2\sigma_N^2}(\mu-\mu_N)^2\right)$$

$$p(x|\mathcal{D}) \propto \exp\left(-\frac{(x-\mu_N)^2}{2(\sigma^2+\sigma_N^2)}\right)$$

Uncertainty inherent in the data

Uncertainty from the model

$$\boxed{p(x|\mathcal{D}) = N(x; \mu_N, \sigma^2 + \sigma_N^2)}$$

See Appendix for details.

Bayesian Linear Regression

Linear Regression

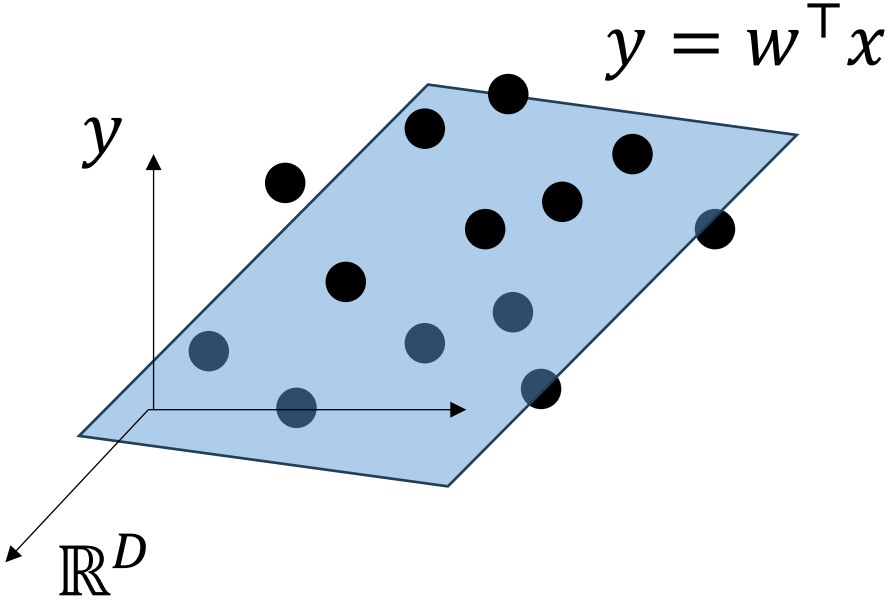
- Given the dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N},\ x_i \in \mathbb{R}^D,\ y_i \in \mathbb{R}$, find a linear model to fit data.

- Gaussian model:

$$p(y|x, w, \sigma^2) = N(y; w^\top x, \sigma^2), \qquad w \in \mathbb{R}^D$$

or

$$y = w^\top x + \epsilon, \qquad \epsilon \sim N(0, \sigma^2)$$



- Likelihood:

$$p(y|X, w, \sigma^2) = \prod_{i=1}^{N} p(y_i|x_i, w, \sigma^2)$$

$$X = (x_1 \quad \cdots \quad x_N) \in \mathbb{R}^{D\times N}$$
$$y = (y_1 \quad \cdots \quad y_N)^\top \in \mathbb{R}^N$$

Linear Regression

- Maximum likelihood estimation:

$$\max_{w} \log L(w) = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - w^\top x_i)^2 - \frac{N}{2}\log\sigma^2 - \frac{N}{2}\log(2\pi)\,.$$

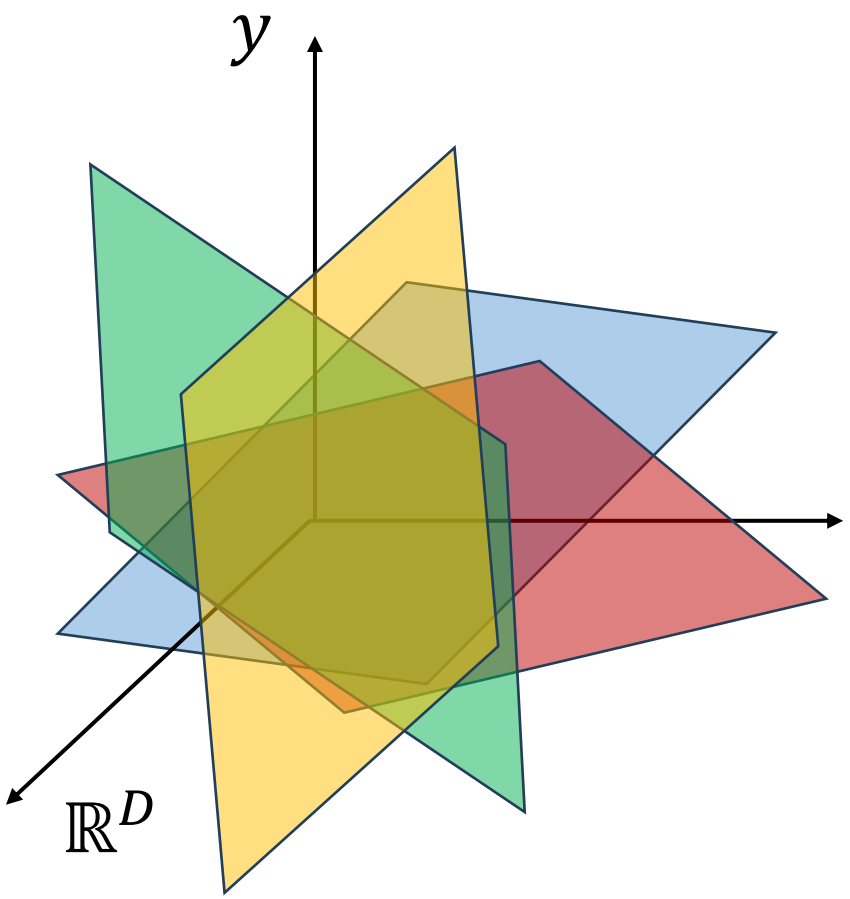
The necessary condition: $$\frac{\partial}{\partial w}\sum_{i=1}^{N}(y_i - w^\top x_i)^2 = -2\sum_{i=1}^{N}(y_i - w^\top x_i)x_i = 0.$$

➡ $$\widehat{w}_{ML} = \left(\sum_{i=1}^{N} x_i x_i^\top\right)^{-1}\sum_{i=1}^{N} y_i x_i = (XX^\top)^{-1}Xy \in \mathbb{R}^D$$

$$X = (x_1 \quad \cdots \quad x_N) \in \mathbb{R}^{D\times N}$$
$$y = (y_1 \quad \cdots \quad y_N)^\top \in \mathbb{R}^N$$

Bayesian Linear Regression

- We assume a prior $p(w) = N(w; 0, \sigma_0^2 I)$.

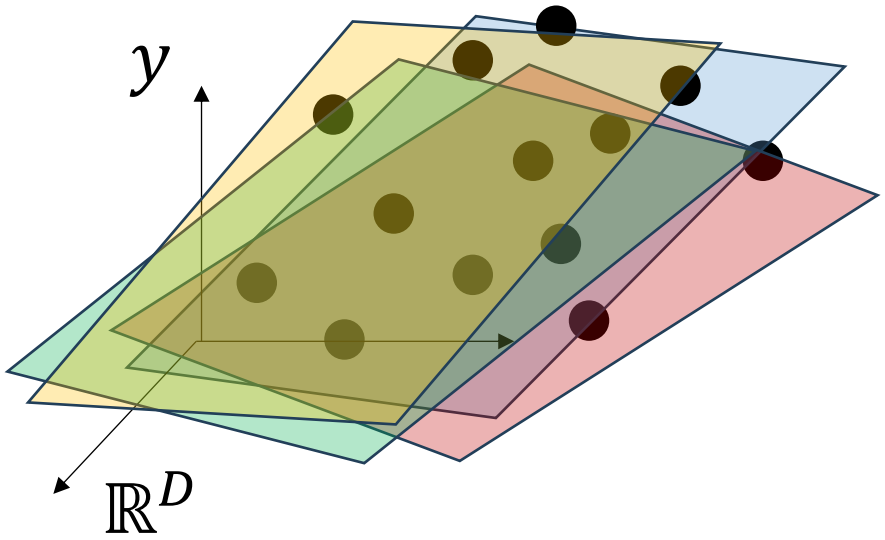


Bayesian Linear Regression

- Posterior:

$$p(w|\mathcal{D}) \propto \prod_{i=1}^{N} p(y_i|x_i, w)p(w)$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}^{ND}} \frac{1}{\sqrt{2\pi\sigma_0^2}^{D}} \exp\left(-\sum_{i=1}^{N} \frac{(y_i - w^\top x_i)^2}{2\sigma^2} - \frac{w^\top w}{2\sigma_0^2}\right)$$



Let's make the exponent to be $-\frac{1}{2}(w - \mu_w)^\top \Sigma_w^{-1}(w - \mu_w)$ to determine the posterior!

Bayesian Linear Regression

- Derive the posterior mean and variance:

$$-\sum_{i=1}^{N}\frac{(y_i - w^\top x_i)^2}{2\sigma^2} - \frac{w^\top w}{2\sigma_0^2} = -\frac{1}{2}(w-\mu_w)^\top \Sigma_w^{-1}(w-\mu_w) + const.$$

$$\mu_w = \left(XX^\top + \frac{\sigma^2}{\sigma_0^2}I\right)^{-1} Xy$$

$$X = (x_1 \quad \cdots \quad x_N) \in \mathbb{R}^{D\times N}$$
$$y = (y_1 \quad \cdots \quad y_N)^\top \in \mathbb{R}^N$$

$$\Sigma_w = \sigma^2\left(XX^\top + \frac{\sigma^2}{\sigma_0^2}I\right)^{-1} = \frac{\sigma^2}{N}\left(\frac{1}{N}XX^\top + \frac{\sigma^2}{N\sigma_0^2}I\right)^{-1}$$

σ^2: inherent noise in the data assumed in $p(y|x,w,\sigma^2)$

As N gets large, Σ_w becomes small.

Regularization becomes small with large N and σ_0^2.

Bayesian Linear Regression

- Predictive distribution

$$p(y|x,\mathcal{D}) = \int p(y|x,w)\, p(w|\mathcal{D})\, dw$$

where

$$p(y|x,w) = N\left(y; w^\top x, \sigma^2\right)$$

$$p(w|\mathcal{D}) = N(w; \mu_w, \Sigma_w).$$

$$p(y|x,w)\, p(w|\mathcal{D}) = C \cdot \exp\left(-\frac{1}{2\sigma^2}(y - w^\top x)^2 - \frac{1}{2}(w-\mu_w)^\top \Sigma_w (w-\mu_w)\right)$$

Marginalizing out w will leave a quadratic term $-\frac{1}{2\sigma^2_{y|x,\mathcal{D}}}\left(y - \mu_{y|x,\mathcal{D}}\right)^2$.

➡ $p(y|x,\mathcal{D})$ is a Gaussian!

Let's calculate the **conditional mean and variance** of $y = w^\top x + \epsilon,\ \epsilon \sim N(0,\sigma^2)$ to figure out $\mu_{y|x,\mathcal{D}}$ and $\sigma^2_{y|x,\mathcal{D}}$.

Bayesian Linear Regression

- Predictive distribution $p(y|x,\mathcal{D}) = N(y;\mu_{y|x,\mathcal{D}},\sigma^2_{y|x,\mathcal{D}})$

$$\mu_{y|x,\mathcal{D}} = \mathbb{E}_{w,\epsilon}[y|x,\mathcal{D}] = \mathbb{E}_{w,\epsilon}[w^\top x + \epsilon|x,\mathcal{D}] = x^\top \mathbb{E}[w|\mathcal{D}]$$

$$= x^\top \mu_w = x^\top \left(XX^\top + \frac{\sigma^2}{\sigma_0^2} I\right)^{-1} Xy$$

$$\sigma^2_{y|x,\mathcal{D}} = Var_{w,\epsilon}[y|x,\mathcal{D}] = \mathbb{E}_{w,\epsilon}[(w^\top x + \epsilon)^2|x,\mathcal{D}] - \mathbb{E}_{w,\epsilon}[w^\top x + \epsilon|x,\mathcal{D}]^2$$

$$= x^\top \mathbb{E}[ww^\top|\mathcal{D}]x + \mathbb{E}[\epsilon^2] - x^\top \mathbb{E}[w|\mathcal{D}]\mathbb{E}[w|\mathcal{D}]^\top x = \sigma^2 + x^\top Cov[w|\mathcal{D}]x$$

$$= \sigma^2 + x^\top \Sigma_w x = \sigma^2 + \frac{\sigma^2}{N} \cdot x^\top \left(\frac{1}{N} XX^\top + \frac{\sigma^2}{N\sigma_0^2} I\right)^{-1} x$$

σ^2: aleatoric uncertainty (inherent noise in the data)

Epistemic uncertainty (associated with model parameter w) which becomes zero as $N \to \infty$.

Bayesian Linear Regression

- Predictive distribution $p(y|x,\mathcal{D}) = N(y; \mu_{y|x,\mathcal{D}}, \sigma^2_{y|x,\mathcal{D}})$

$$\mu_{y|x,\mathcal{D}} = x^\top \mu_w = x^\top \left(XX^\top + \frac{\sigma^2}{\sigma_0^2} I \right)^{-1} Xy = x^\top X \left(X^\top X + \frac{\sigma^2}{\sigma_0^2} I \right)^{-1} y$$

Results are expressed using inner products
→ Extend to **kernel methods!**

$$\sigma^2_{y|x,\mathcal{D}} = \sigma^2 + x^\top \Sigma_w x = \sigma^2 + \sigma^2 \cdot x^\top \left(XX^\top + \frac{\sigma^2}{\sigma_0^2} I \right)^{-1} x$$

σ^2: aleatoric uncertainty (inherent noise in the data)

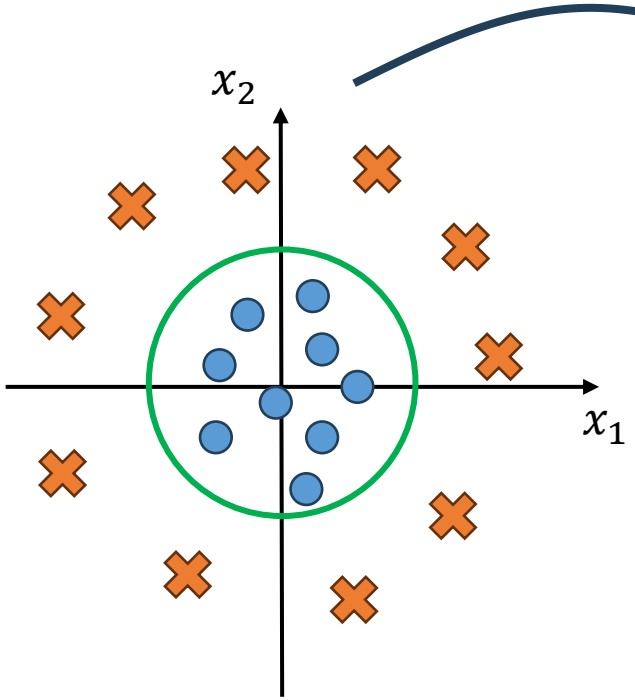
Epistemic uncertainty (associated with model parameter w) which becomes zero as $N \to \infty$.

$$= \sigma^2 + \sigma_0^2 x^\top x - \sigma_0^2 x^\top X \left(X^\top X + \frac{\sigma^2}{\sigma_0^2} I \right)^{-1} X^\top x$$

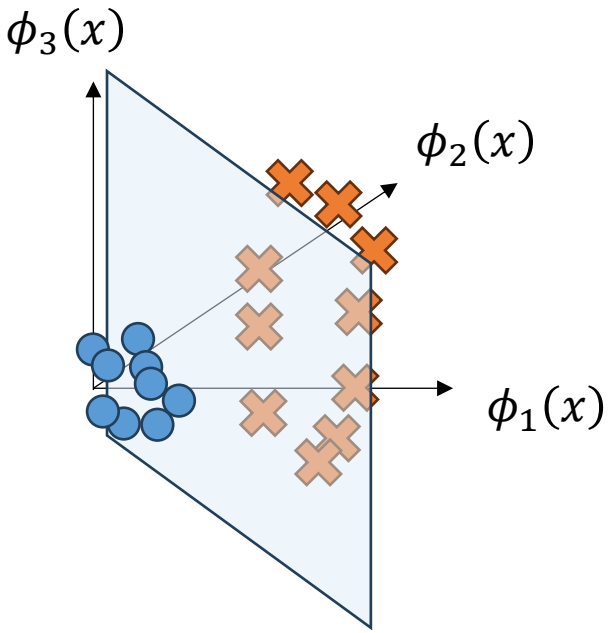
See Appendix for details.

Kernel Methods

$$x \in \mathbb{R}^2 \mapsto \phi(x) = \left(x_1^2, x_2^2, \sqrt{2}x_1x_2\right) \in \mathbb{R}^3$$



Linearly non-separable



Linearly separable in higher dimension

Kernel function:

$$k(x, y) = \phi(x)^\top \phi(y)$$

$$= x_1^2 y_1^2 + x_2^2 y_2^2 + 2x_1x_2y_1y_2$$

$$= (x^\top y)^2$$

Kernel Methods

- Mercer theorem

$k(x, y)$ is **positive semi-definite**, i.e.,

$$K = \begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_N) \\ \vdots & \ddots & \vdots \\ k(x_N, x_1) & \cdots & k(x_N, x_N) \end{pmatrix} \in \mathbb{R}^{N \times N} \text{ is } \Leftrightarrow$$

Feature mapping $\phi(x)$ that satisfies $k(x, y) = \phi(x)^\top \phi(y)$ exists.

positive semi-definite for any $x_1, \ldots, x_N$.

- Examples of kernel functions
  - Polynomial kernel $k(x, y) = (x^\top y + 1)^d$

    When $d = 2$, corresponding $\phi(x) = \begin{pmatrix} x_1^2 & x_2^2 & 1 & \sqrt{2}x_1x_2 & \sqrt{2}x_1 & \sqrt{2}x_2 \end{pmatrix} \in \mathbb{R}^6$.
  - Exponential kernel $k(x, y) = \exp\left(-\frac{\|x-y\|^2}{2}\right)$

    Corresponding $\phi(x) = \exp\left(-\frac{x^2}{2}\right)\begin{pmatrix} 1 & x_1 & \cdots & x_D & x_1^2/\sqrt{2} & \cdots & x_D^2/\sqrt{2} & x_1x_2 & \cdots \end{pmatrix} \in \mathbb{R}^\infty$.

Kernelization of Bayesian Linear Regression

Feature mapping: $x \in \mathbb{R}^D \mapsto \phi(x) \in \mathbb{R}^F$

$$X = (x_1 \quad \cdots \quad x_N) \in \mathbb{R}^{D \times N} \mapsto \Phi = (\phi(x_1) \quad \cdots \quad \phi(x_N)) \in \mathbb{R}^{F \times N}$$

$$X^\top x \mapsto k \equiv \Phi^\top \phi(x) = (\phi(x)^\top \phi(x_1) \quad \cdots \quad \phi(x)^\top \phi(x_N))^\top = (k(x_1, x) \quad \cdots \quad k(x_N, x))^\top \in \mathbb{R}^{N \times 1}$$

$$X^\top X \mapsto K \equiv \Phi^\top \Phi = \begin{pmatrix} \phi(x_1)^\top \phi(x_1) & \cdots & \phi(x_1)^\top \phi(x_N) \\ \vdots & \ddots & \vdots \\ \phi(x_N)^\top \phi(x_1) & \cdots & \phi(x_N)^\top \phi(x_N) \end{pmatrix} = \begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_N) \\ \vdots & \ddots & \vdots \\ k(x_N, x_1) & \cdots & k(x_N, x_N) \end{pmatrix} \in \mathbb{R}^{N \times N}$$

$$x^\top X \left(X^\top X + \frac{\sigma^2}{\sigma_0^2} I \right)^{-1} y$$

$$\sigma^2 + \sigma_0^2 x^\top x - \sigma_0^2 x^\top X \left(X^\top X + \frac{\sigma^2}{\sigma_0^2} I \right)^{-1} X^\top x$$

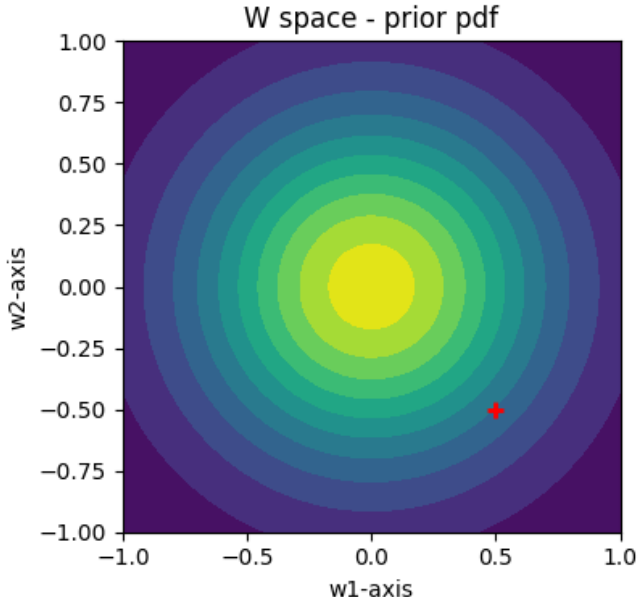
$$\Rightarrow$$

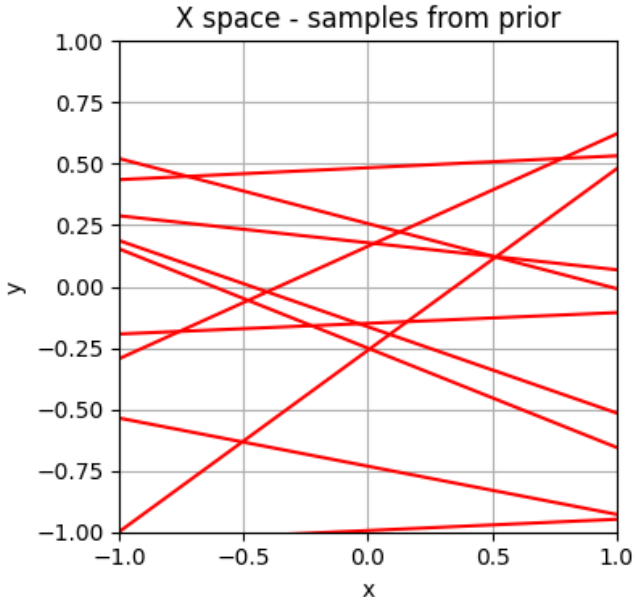
$$\mu_{y|x,\mathcal{D}} = k^\top \left(K + \frac{\sigma^2}{\sigma_0^2} I \right)^{-1} y$$

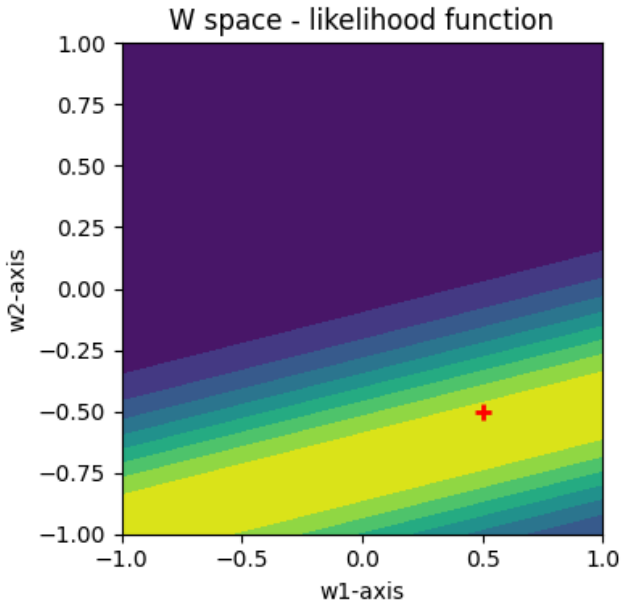
$$\sigma^2_{y|x,\mathcal{D}} = \sigma^2 + \sigma_0^2 k(x, x) - \sigma_0^2 k^\top \left(K + \frac{\sigma^2}{\sigma_0^2} I \right)^{-1} k$$

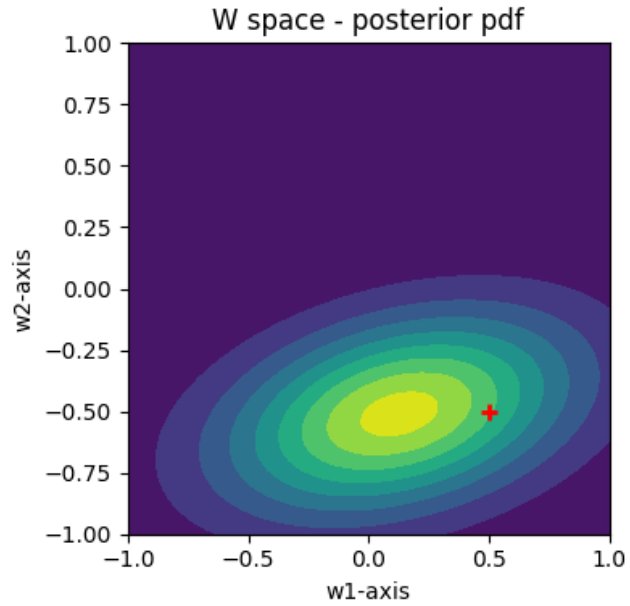
Bayesian Linear Regression Examples

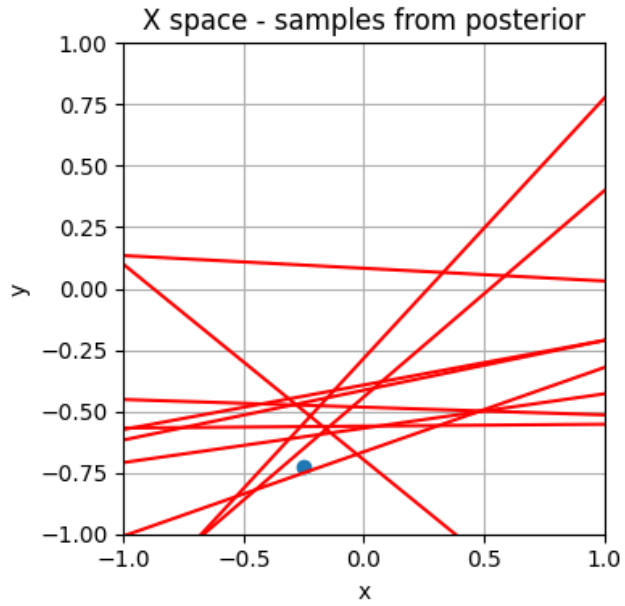
✚ : true parameter

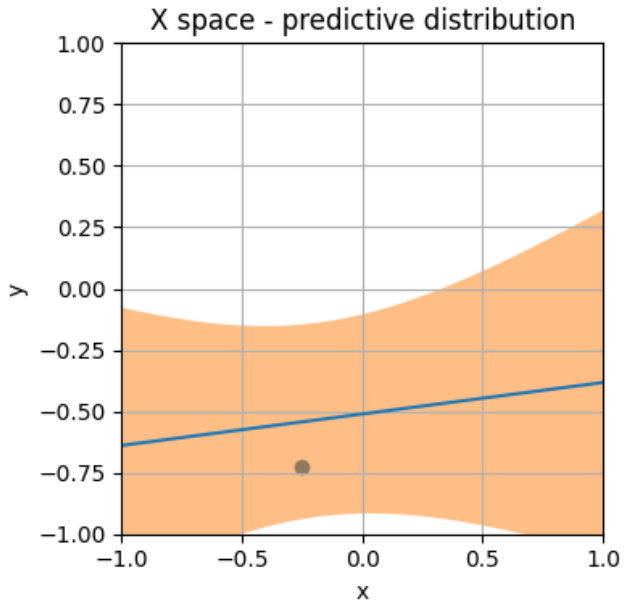




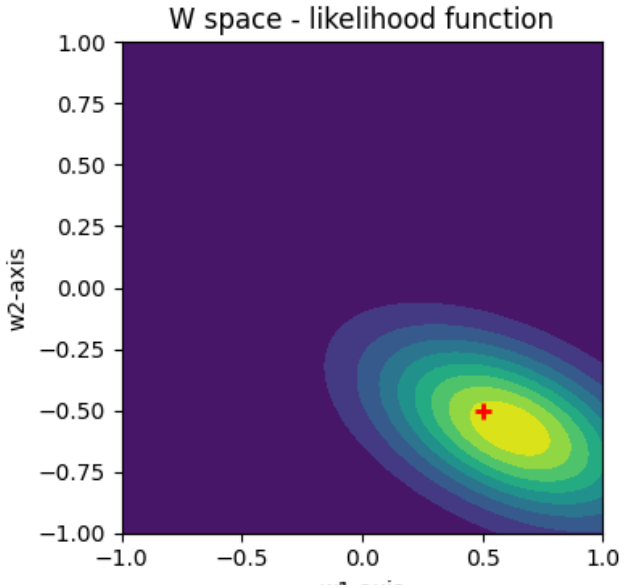


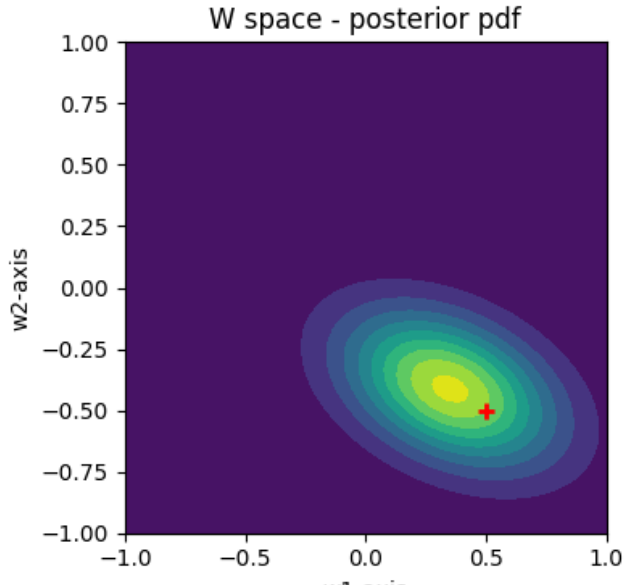


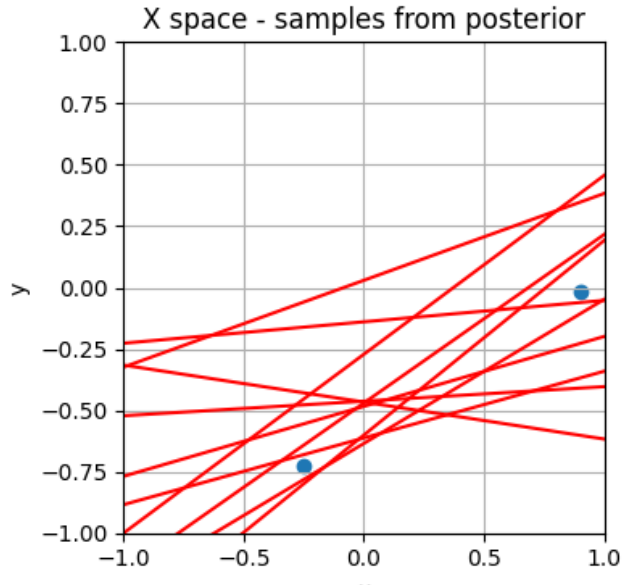


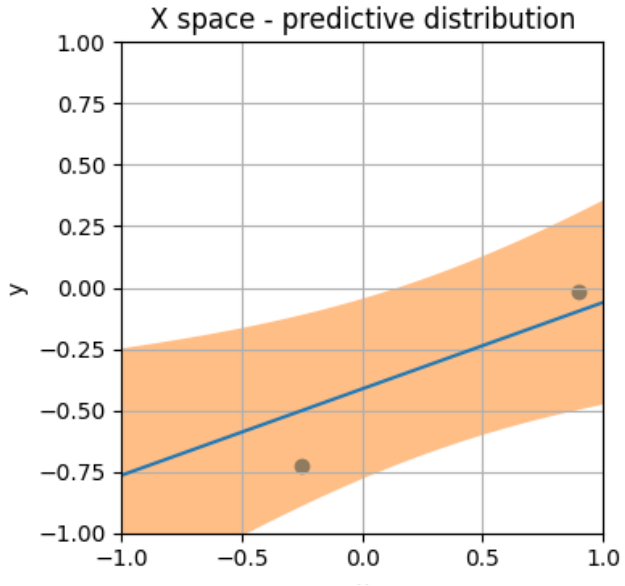


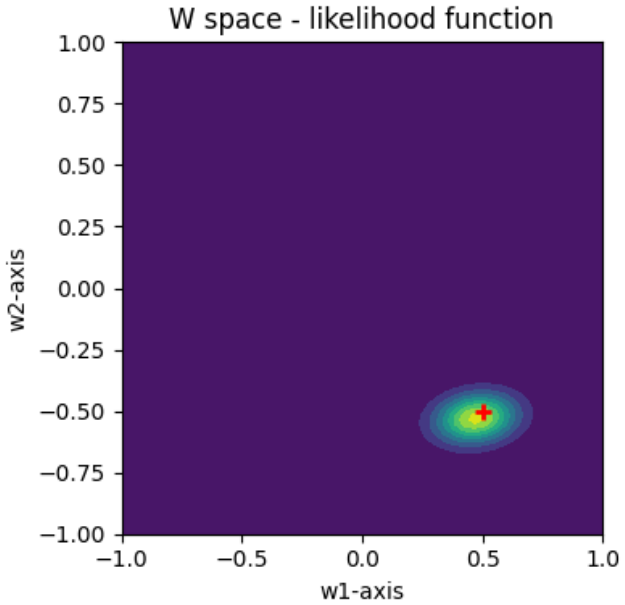
Bayesian Linear Regression Examples

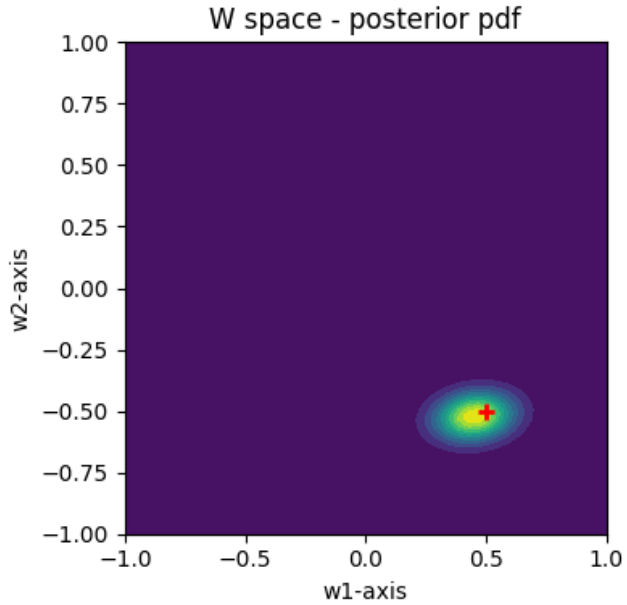


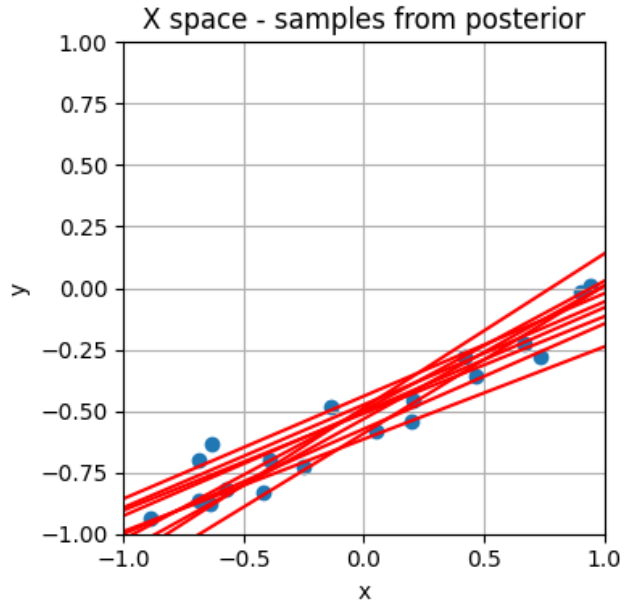


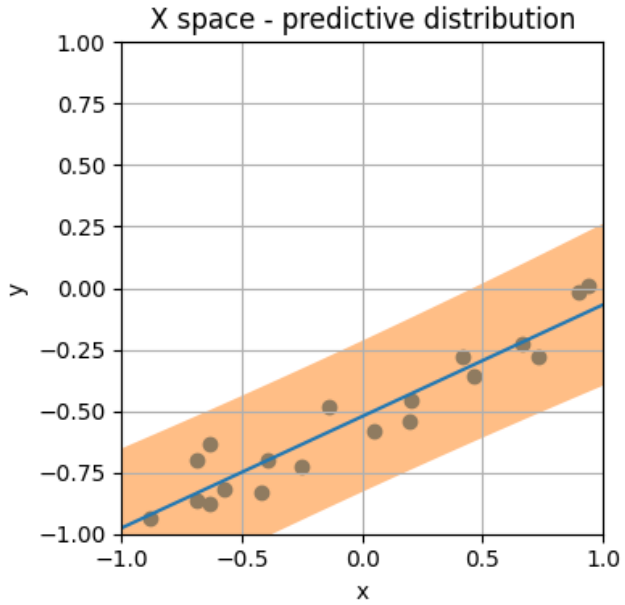




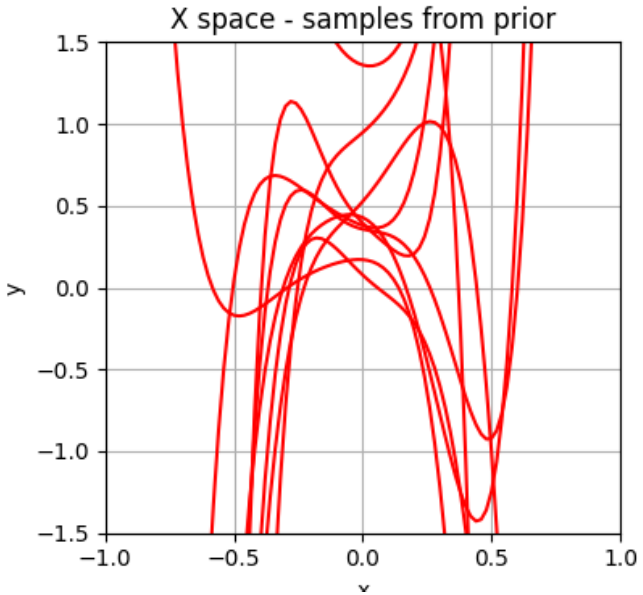






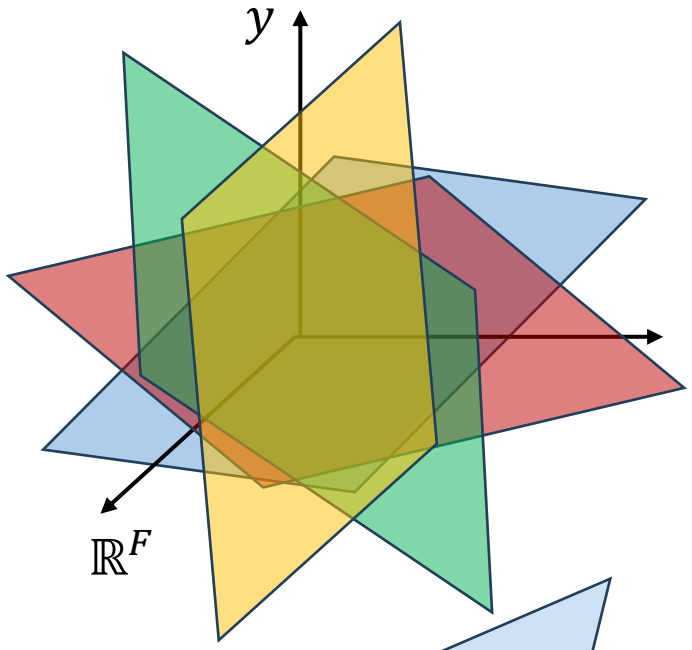


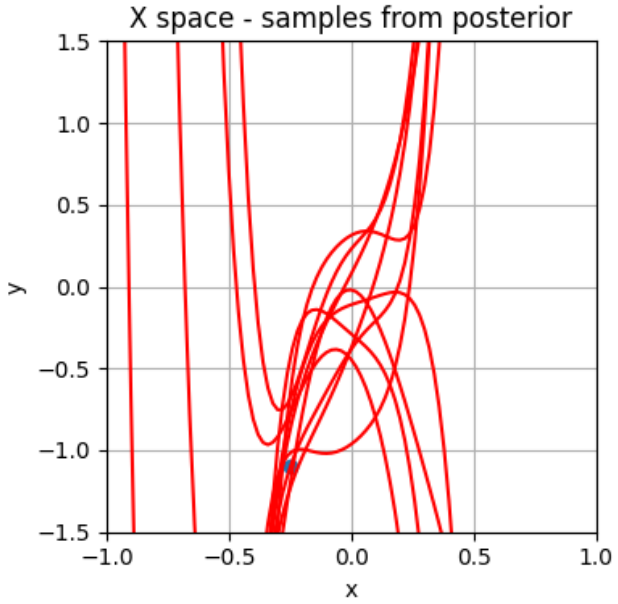
Kernel Bayesian Regression Examples

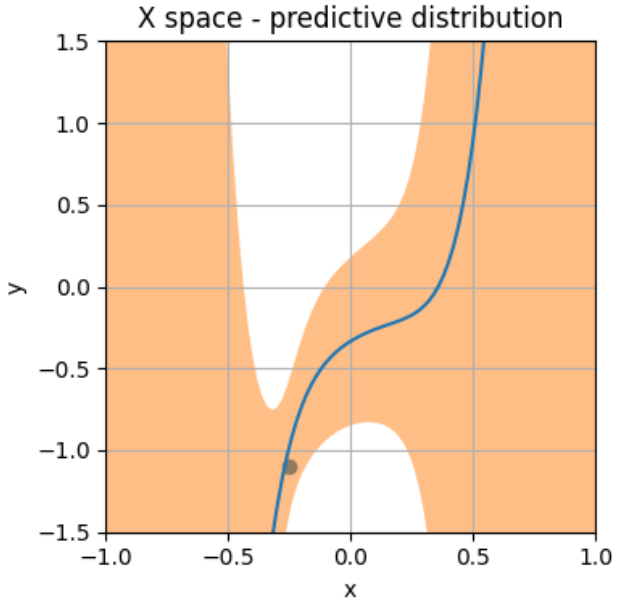


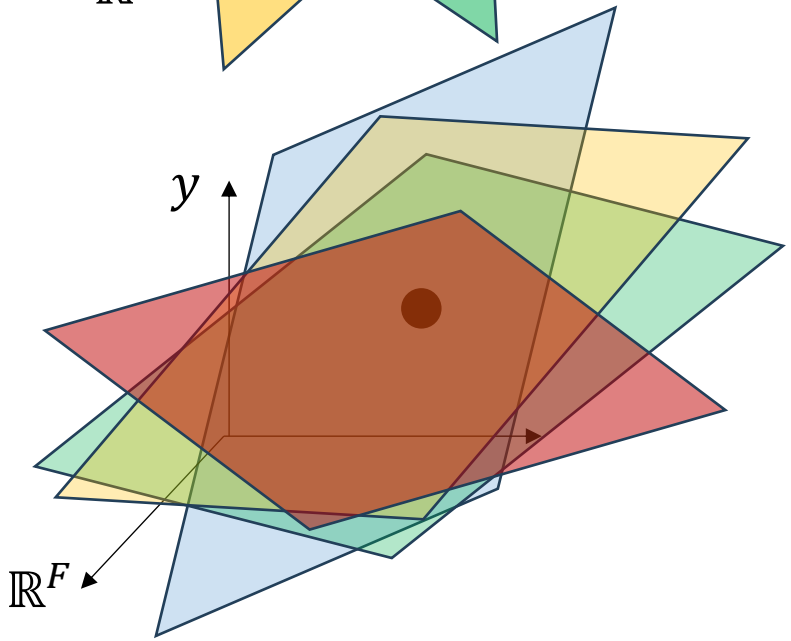
Polynomial feature mapping:

$$\phi(x) = \left(\left(\frac{x}{s}\right)^p, \left(\frac{x}{s}\right)^{p-1}, \ldots, \frac{x}{s}, 1 \right)^\top \in \mathbb{R}^F$$

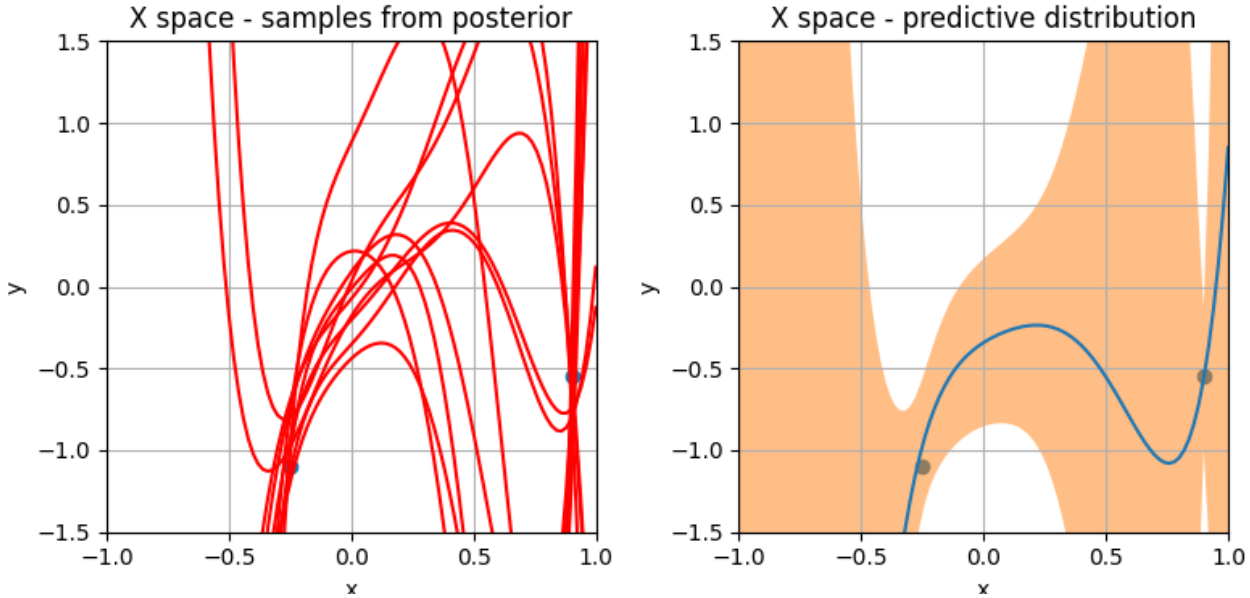


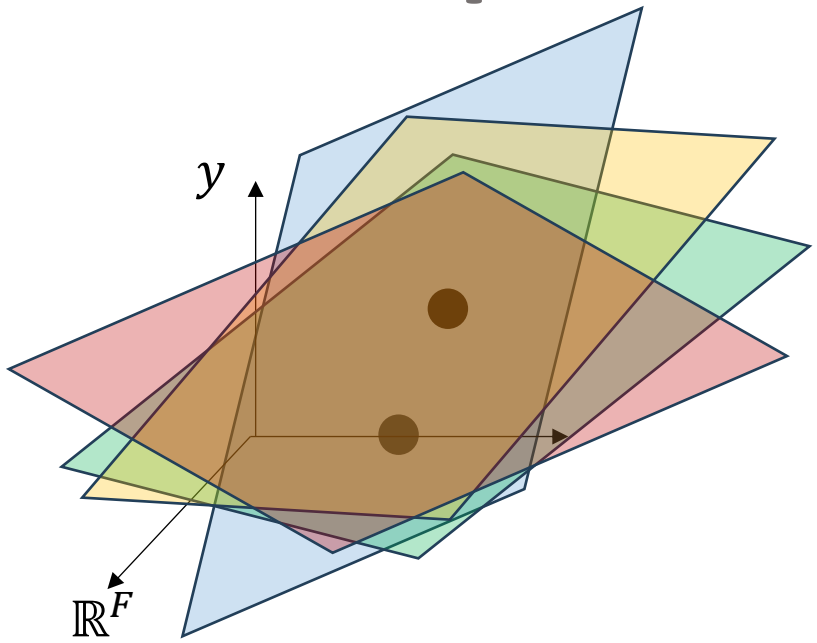


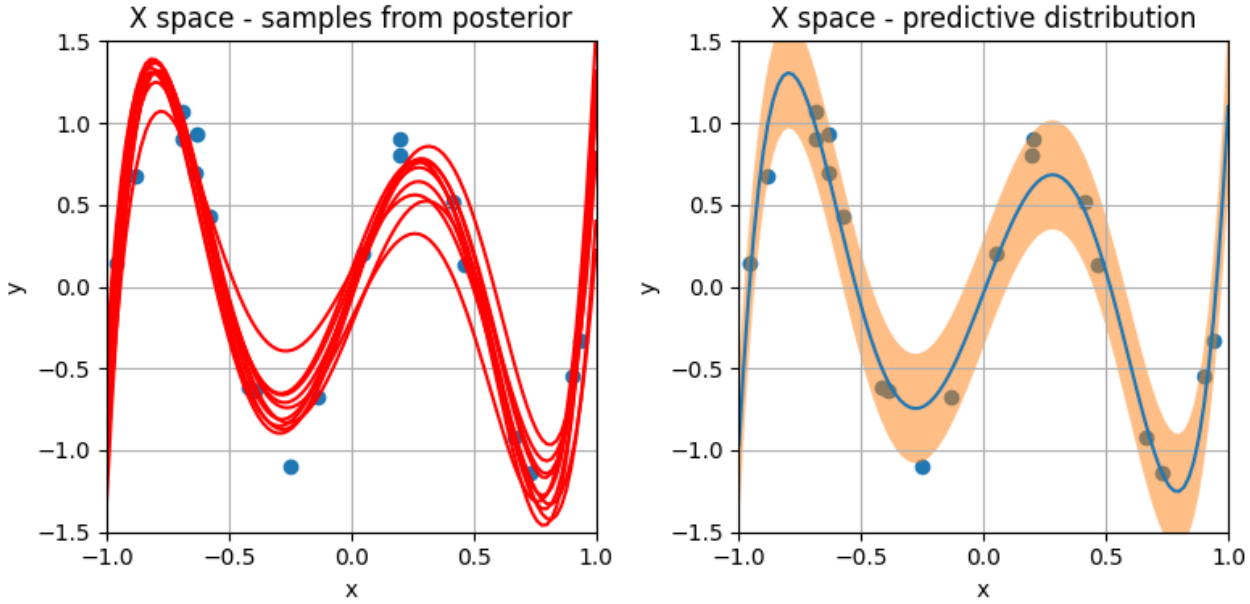


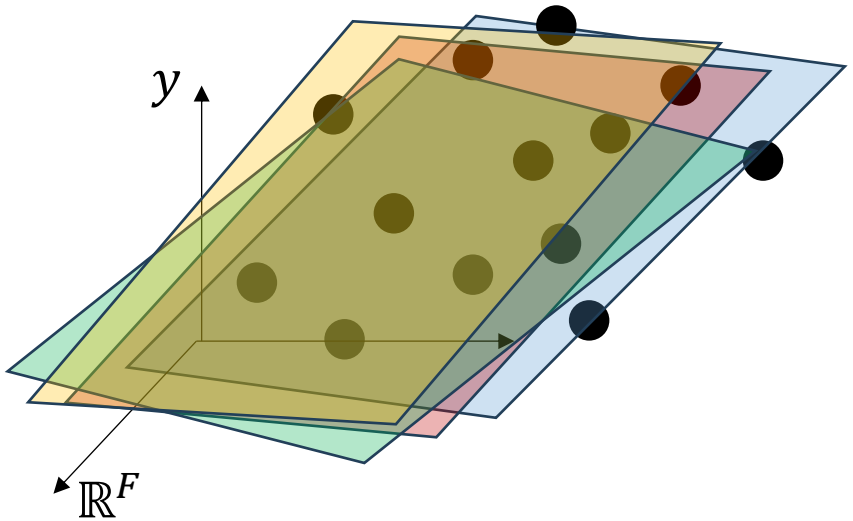


Kernel Bayesian Regression Examples



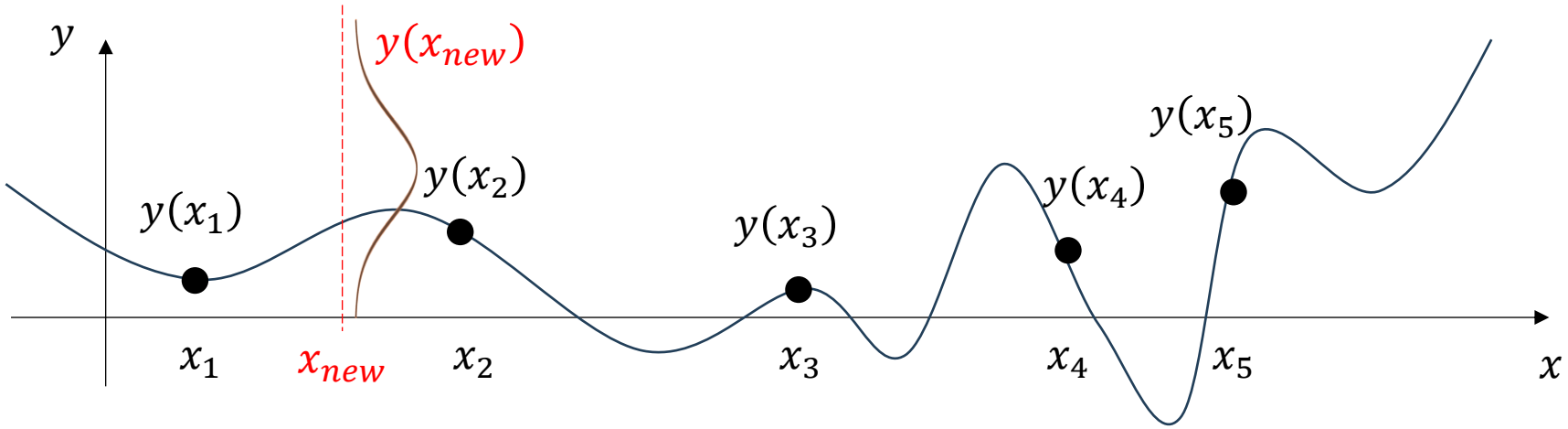






Connections to Gaussian Processes

- Gaussian process regression (parameter space view)



Choosing $m(x) = 0$ and $k_{GP}(x_i, x_j) = \sigma^2 \delta_{ij} + \sigma_0^2 k(x_i, x_j)$ recovers the results from **Bayesian linear regression**.

$$p(y(x_{new})|y(x_1), \ldots, y(x_N)) = N(\mu_{y|x,\mathcal{D}}, \sigma^2_{y|x,\mathcal{D}})$$ where

$$\mu_{y|x,\mathcal{D}} = k^\top \left(K + \frac{\sigma^2}{\sigma_0^2} I \right)^{-1} y$$

$$\sigma^2_{y|x,\mathcal{D}} = \sigma^2 + \sigma_0^2 k(x,x) - \sigma_0^2 k^\top \left(K + \frac{\sigma^2}{\sigma_0^2} I \right)^{-1} k$$

Laplace Approximation

Intractable Marginal Likelihood and Posterior

- **Problem:**

Let $f(\theta) = \prod_{i=1}^{N} p(x_i|\theta)\ p(\theta)$ be the unnormalized posterior.

$p(\mathcal{D}) = \int f(\theta) d\theta$: marginal likelihood

$p(\theta|\mathcal{D}) = f(\theta)/p(\mathcal{D})$: posterior

➡ $p(\mathcal{D})$ and $p(\theta|\mathcal{D})$ are intractable unless the integration is possible.

- **Sol 1.** Let's approximate $\log f(\theta)$ quadratically locally (at the mode) so that the posterior is a Gaussian.

  ➡ **Laplace approximation**

- **Sol 2.** Let's approximate the posterior with a tractable distribution $q(\theta)$.

  ➡ **Variational inference** (will be introduced later)

Laplace Approximation

- Perform a **Gaussian approximation** of the posterior at the **mode** of the posterior.
- The second order approximation of $\log f(\theta)$ near $\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\, f(\theta)$:

$$\log \hat{f}(\theta) = \log f(\hat{\theta}_{MAP}) - \frac{1}{2}(\theta - \hat{\theta}_{MAP})^{\top} A (\theta - \hat{\theta}_{MAP})$$

$$\hat{f}(\theta) = f(\hat{\theta}_{MAP}) \exp\left(-\frac{1}{2}(\theta - \hat{\theta}_{MAP})^{\top} A (\theta - \hat{\theta}_{MAP})\right)$$

From the necessary condition of $\hat{\theta}_{MAP}$,

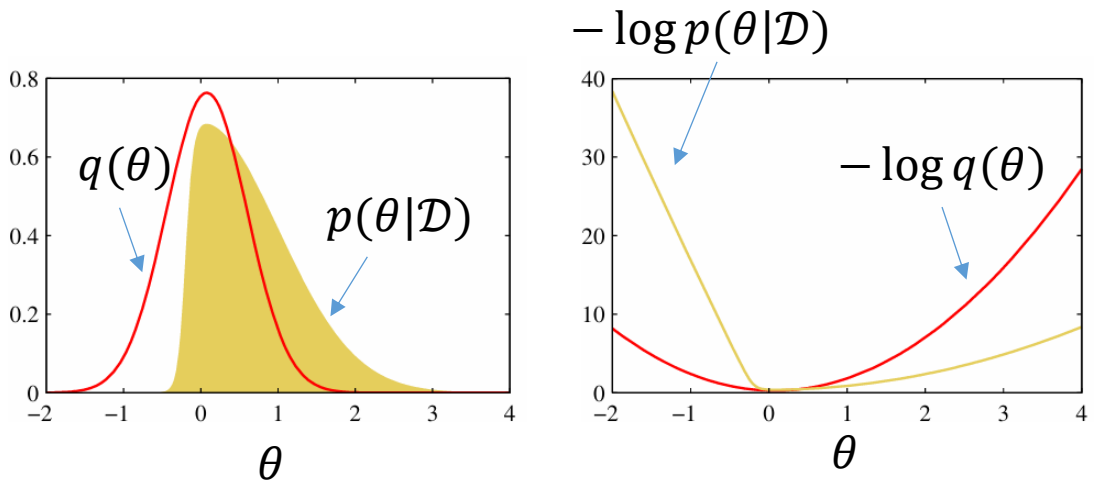
$$\frac{\partial}{\partial \theta} \log f(\theta)\Big|_{\theta=\hat{\theta}_{MAP}} = 0.$$

$$A = -\frac{\partial^2}{\partial \theta^2} \log f(\theta)\Big|_{\theta=\hat{\theta}_{MAP}} \in \mathbb{R}^{p \times p}:$$

the negative Hessian matrix.

➡ $$\int \hat{f}(\theta)\, d\theta = f(\hat{\theta}_{MAP}) \cdot \sqrt{\frac{(2\pi)^p}{|A|}}$$

➡ $$q(\theta) = \frac{\hat{f}(\theta)}{\int \hat{f}(\theta)\, d\theta} = N(\hat{\theta}_{MAP}, A^{-1})$$



Laplace Approximation

- Perform a **Gaussian approximation** of the posterior at the **mode** of the posterior.

$$q(\theta) = N\left(\hat{\theta}_{MAP}, A^{-1}\right)$$

$$A = -\frac{\partial^2}{\partial \theta^2} \log f(\theta)\Big|_{\theta=\hat{\theta}_{MAP}} \in \mathbb{R}^{p\times p}:$$
the negative Hessian matrix.

Q. When would this approximation be accurate?
→When $\log f(\theta)$ is nearly quadratic.
→When $p(\theta|\mathcal{D})$ is close to the Gaussian!
→When N is sufficiently large, $p(\theta|\mathcal{D})$ is close to Gaussian (due to the central limit theorem).

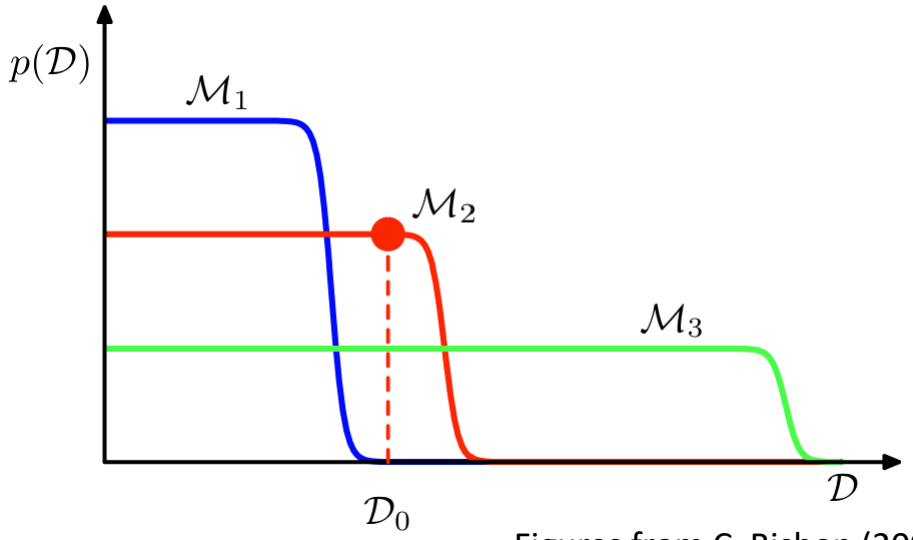
Application to Model Comparison

- Bayesian inference utilizes the **hypothesis set, prior distribution** to make predictions. How can we compare different hypotheses, priors?

  An idea: Use data $\mathcal{D} = \{x_i\}_{i=1}^{N}$ to objectively compare the models.

  Marginal likelihood: $$p(\mathcal{D}|\mathcal{M}) = \int \prod_{i=1}^{N} p(x_i|\theta, \mathcal{M})\ p(\theta|\mathcal{M})\, d\theta$$

- Bayes factor: $$\frac{p(\mathcal{D}|\mathcal{M}_1)}{p(\mathcal{D}|\mathcal{M}_2)}$$



Figures from C. Bishop (2006)

Application to Model Comparison

$$f(\theta) = \prod_{i=1}^{N} p(x_i|\theta, \mathcal{M})\ p(\theta|\mathcal{M})$$

$$p(\mathcal{D}|\mathcal{M}) = \int f(\theta) d\theta \approx \int \hat{f}(\theta)\, d\theta = f\left(\hat{\theta}_{MAP}\right) \cdot \sqrt{\frac{(2\pi)^p}{|A|}}$$

$$\log p(\mathcal{D}|\mathcal{M})$$

$$\approx \underbrace{\sum_{i=1}^{N} \log p\left(x_i|\hat{\theta}_{MAP}, \mathcal{M}\right)}_{\text{Log-likelihood}} + \underbrace{\log p(\hat{\theta}_{MAP}|\mathcal{M}) + \frac{p}{2}\log 2\pi - \frac{1}{2}\log|A|}_{\text{Occam factor}}$$

Bayesian Information Criterion (BIC)

- If $N \gg 1$, we can approximate further.

$$A = -\frac{\partial^2}{\partial\theta^2}\left(\sum_{i=1}^{N} \log p(x_i|\theta,\mathcal{M}) + \log p(\theta|\mathcal{M})\right)\Bigg|_{\theta=\widehat{\theta}_{MAP}} \approx N \cdot H, \quad \text{where } H = -\frac{\partial^2}{\partial\theta^2}\,\mathbb{E}_{p(x)}[\log p(x|\theta,\mathcal{M})]\Big|_{\theta=\widehat{\theta}_{MAP}}.$$

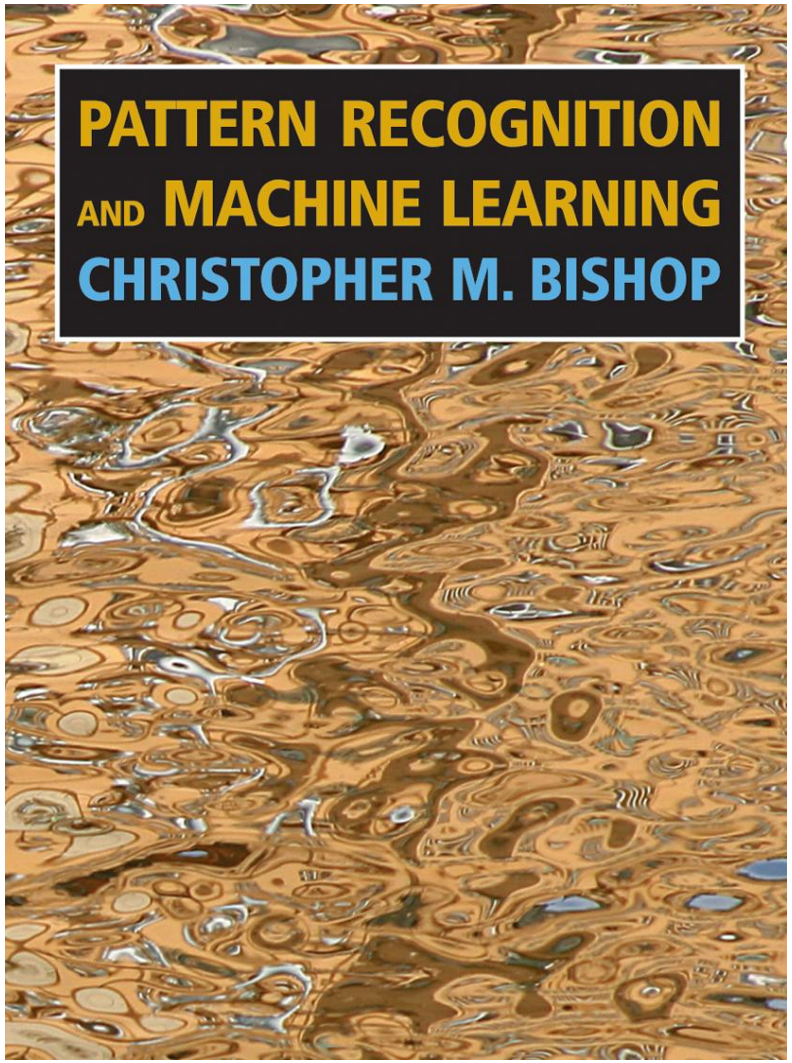
$$|A| \approx N^p |H|, \qquad \log|A| \approx p \log N + \log|H|.$$

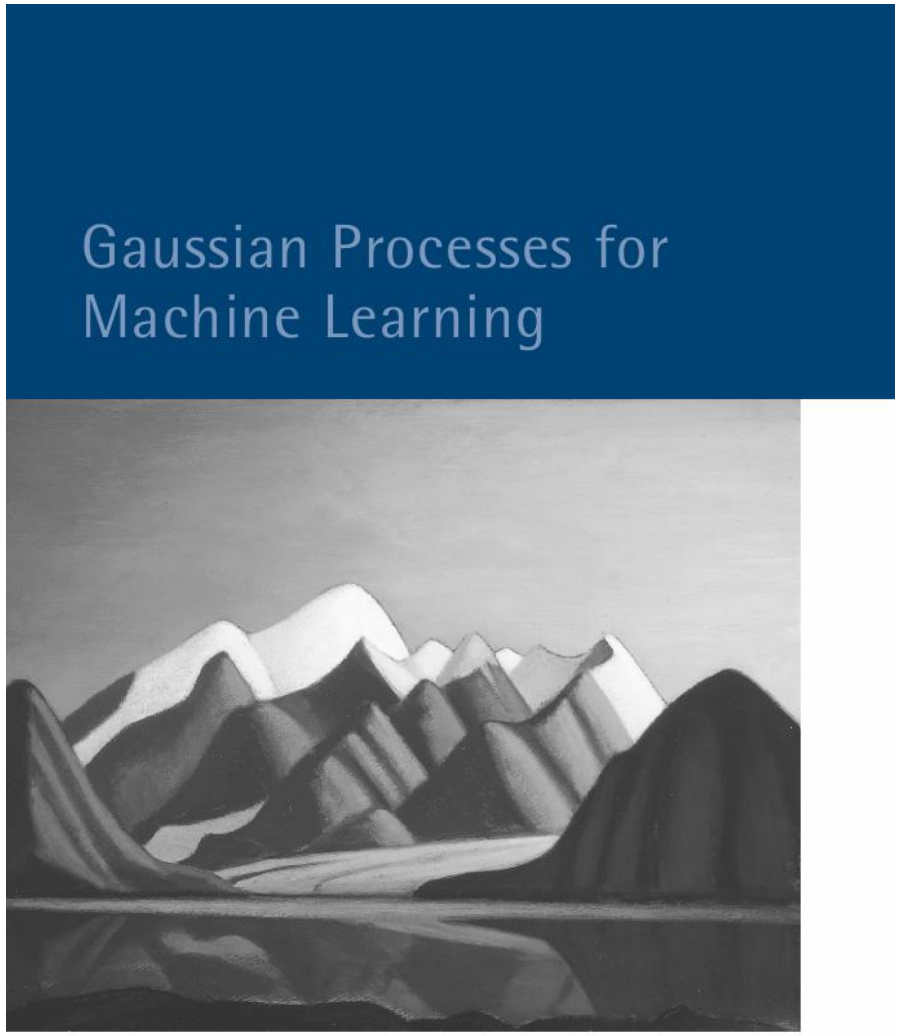
Bayesian information criterion (BIC):

$$-\log p(\mathcal{D}|\mathcal{M}) \approx -\sum_{i=1}^{N} \log p\left(x_i\middle|\hat{\theta}_{MAP}\right) + \frac{p}{2}\log N$$

When we **select models with a lower BIC**, models of the same likelihood using more parameters are not selected **(Occam's razor).**

References





Carl Edward Rasmussen and Christopher K. I. Williams

Information Theory for Machine Learning

Information Content (Degree of Surprise)

- How surprising is it to observe x (a discrete random variable)?

$$h(x) = -\log P(x)$$

Which event's information should be larger?

rare event (small P) → large surprise; certain event ($P = 1$) → zero surprise.

Ex) $x \in \{pass, fail\}$: if $P(pass) = 1$, observing $x = pass$ tells us nothing;
if $P(pass) = 0.1$, observing $x = pass$ is much **informative.**

Why the log?

If two unrelated events x and y occur, the information gain observing both should be the sum of the gain from each of them:

$$p(x, y) = p(x)p(y) \rightarrow h(x, y) = h(x) + h(y)$$

Entropy

- Averaged information content of an outcome (a.k.a. uncertainty)

Entropy for a **discrete** random variable X:

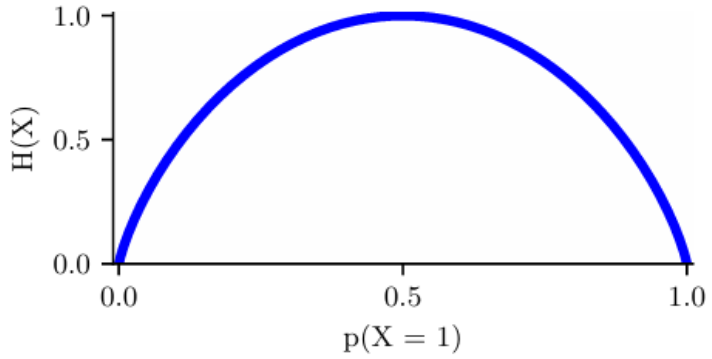
$$H(X) = -\sum_x P(x) \log P(x)$$

(We let $P(x) \log P(x) = 0$ when $P(x) = 0$.)

Differential entropy for a **continuous** random variable X:

$$H(X) = -\int p(x) \log p(x) \; dx$$

Ex) Entropy of a binary random variable



(maximum at $p = 0.5$)

Figures from K. Murphy (2023)

Entropy: Properties

- Suppose X takes values $x_1, \ldots, x_M$.
- Non-negative: $H(X) \geq 0$, with equality if and only if X is deterministic, i.e., $P(x_i) = 1$ for one i.
- Maximal at uniform: $H(X) \leq \log M$ (equality if and only if $P(x_1) = \cdots = P(x_M) = 1/M$).
- Joint entropy: $H(X,Y) = -\sum_{x,y} P(x,y) \log P(x,y)$

Derive these properties as an exercise ☺

- Additive under independence:

$H(X,Y) = H(X) + H(Y)$ if and only if $P(x,y) = P(x)P(y)$.

Entropy as Code Length

- $-\log_2 P(x) =$ optimal code length (bits) for symbol x; rare symbols get longer codes.
- **Entropy = minimum average code length** to encode data from P (Shannon's source coding theorem):

$$H(X) = -\sum_x P(x) \log_2 P(x)$$

Ex) a non-uniform 8-state distribution

x	$P(x)$	Length = $-\log_2 P(x)$	Code
a	1/2	1	0
b	1/4	2	10
c	1/8	3	110
d	1/16	4	1110
e	1/64	6	111100
f	1/64	6	111101
g	1/64	6	111110
h	1/64	6	111111

Average code length

= ½·1 + ¼·2 + ⅛·3 + 1/16·4 + 1/64·6·4

= 2 bits = $H(X)$

Cf) Uniform case

Average code length = 3 bits = $H(X)$

Cross Entropy

- What if we encode data $x \sim P$ with a code based on the **wrong** distribution Q?

Cross entropy: $$H(P, Q) = -\sum_{x} P(x) \log Q(x)$$

= average code length using a code based on Q, on data from P.

The cross entropy is **minimized** when $Q = P$ for a fixed P (will be addressed shortly).

KL-Divergence

- The **Kullback-Leibler (KL) divergence** (or the relative entropy) between two PDFs $p(x)$ and $q(x)$:

$$KL(p\|q) = -\int p(x)\log q(x)\ dx + \int p(x)\log p(x)\ dx = \int p(x)\log\frac{p(x)}{q(x)}\ dx$$

$$= H(p,q) - H(p)$$

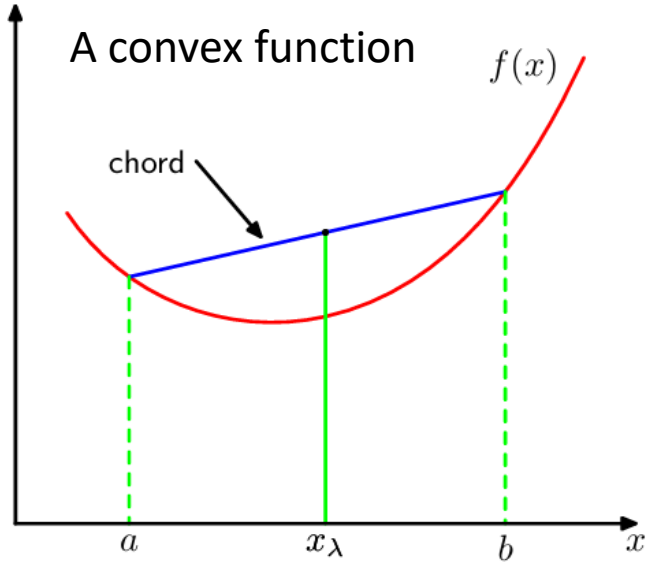
= average **additional** code length required to encode data from p due to using a code based on q (instead of p).

KL-Divergence

$$KL(p||q) = \int p(x) \log \frac{p(x)}{q(x)} \, dx$$

- Gibbs' inequality: $KL(p||q) \geq 0$ with equality if and only if $p(x) = q(x)$ for all x.

proof idea) apply the **Jensen's inequality** for convex functions, i.e., $\mathbb{E}[f(x)] \geq f(\mathbb{E}[x])$ for a convex function f.



KL-divergence is a very useful quantitative means of **measuring the difference** between two probability distributions.

Note: not a distance metric (asymmetric, triangular inequality not satisfied).

Figures from C. Bishop (2006)

KL-Divergence

- Connections to the **maximum likelihood**:

$$q^* = \underset{q}{\operatorname{argmin}}\, KL(p\|q) = \underset{q}{\operatorname{argmin}} -\int p(x)\log q(x)\ dx + \int p(x)\log p(x)\ dx$$

Suppose p is the **empirical distribution** from $\{x_i\}_{i=1}^N$:

$$p_{\mathcal{D}}(x) = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}(x) \qquad (\delta_{x_i}(x) = 1 \text{ if } x = x_i \text{ and zero otherwise.})$$

$$KL(p_{\mathcal{D}}\|q) = -\int p_{\mathcal{D}}(x)\log q(x)\ dx + C = -\int \left(\frac{1}{N}\sum_{i=1}^N \delta_{x_i}(x)\right)\log q(x)\ dx + C$$

$$= -\frac{1}{N}\sum_{i=1}^N \log q(x_i) + C.$$

Minimizing $KL(p_{\mathcal{D}}\|q)$ is **equivalent** to maximizing the log likelihood.

KL-Divergence

- Asymmetry

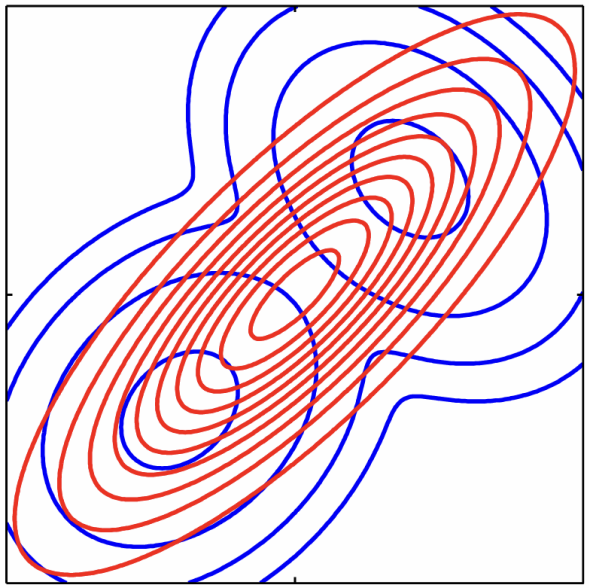
$$\min_q KL(p||q) = \int p(x) \log \frac{p(x)}{q(x)} \, dx$$

$q(x)$ should not be zero when $p(x) > 0$.

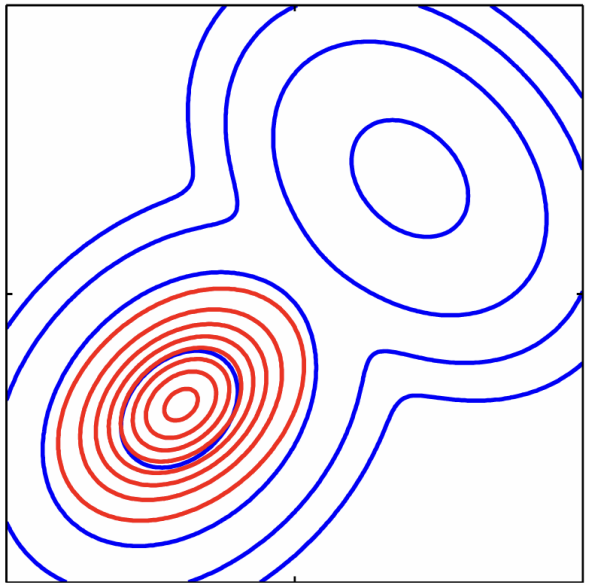
$$\min_q KL(q||p) = \int q(x) \log \frac{q(x)}{p(x)} \, dx$$

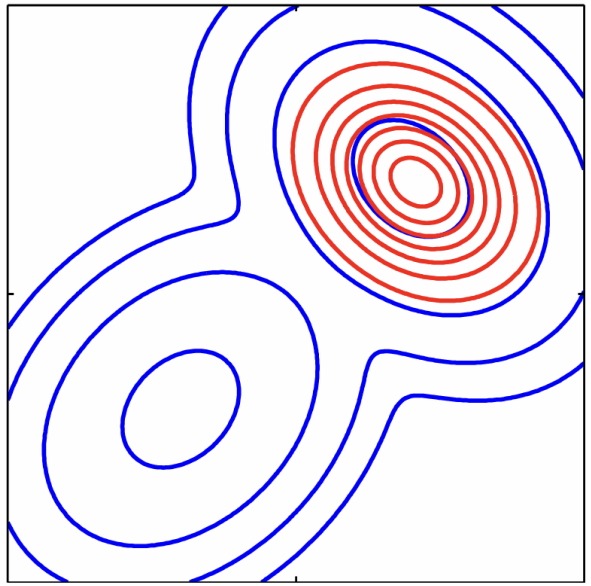
$p(x)$ should not be zero when $q(x) > 0$.

Ex) p is a mixture of two Gaussians, q is a single Gaussian.



Zero-avoiding





Mode-seeking (or zero-forcing)

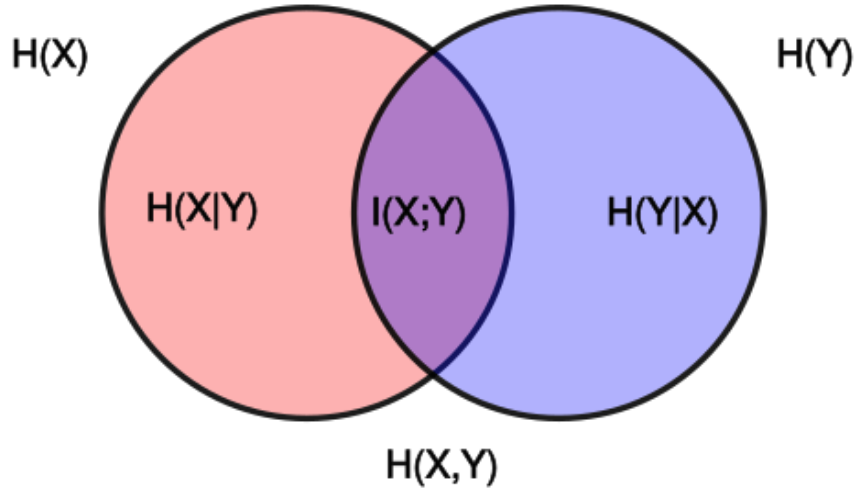
Figures from C. Bishop (2006)

Mutual Information

Conditional entropy:

$$H(X|Y) = -\int\int p(x,y)\log p(x|y)\,dx\,dy$$

$$H(X,Y) = H(Y) + H(X|Y)$$



Mutual information:

$$I(X;Y) = KL(p(x,y)\|p(x)p(y)) = \iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)}\,dx\,dy$$

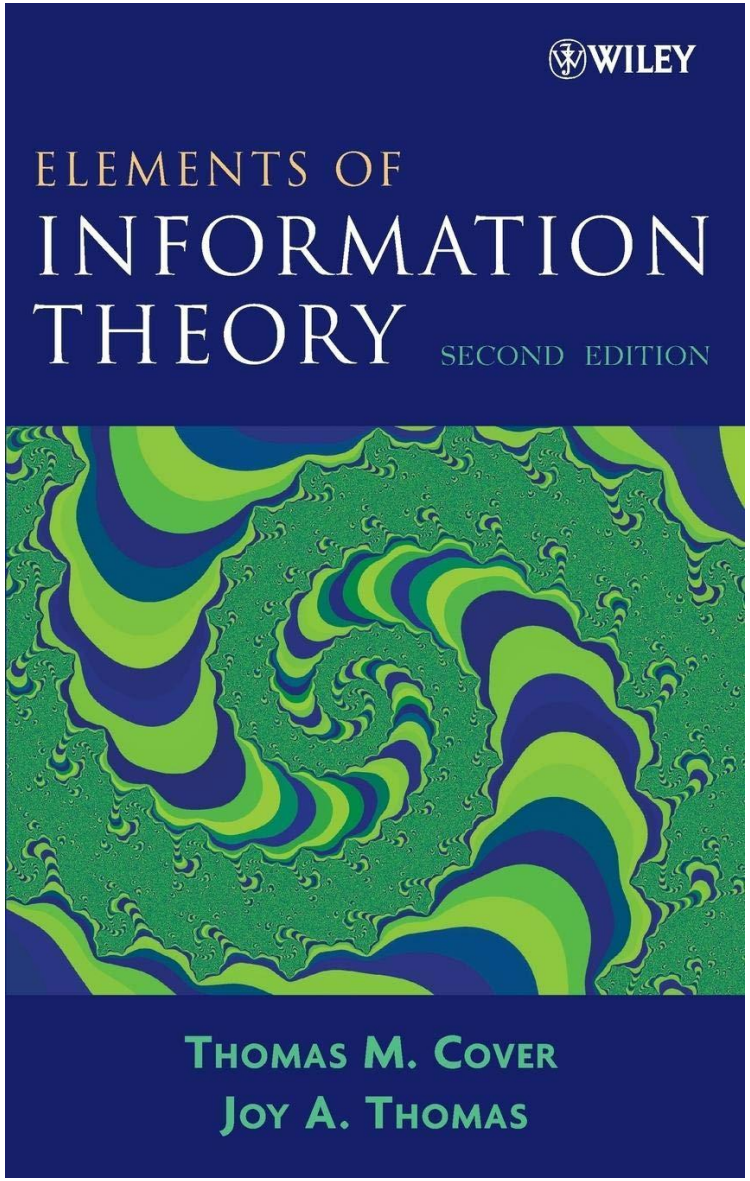
$$= H(Y) - H(Y|X)$$

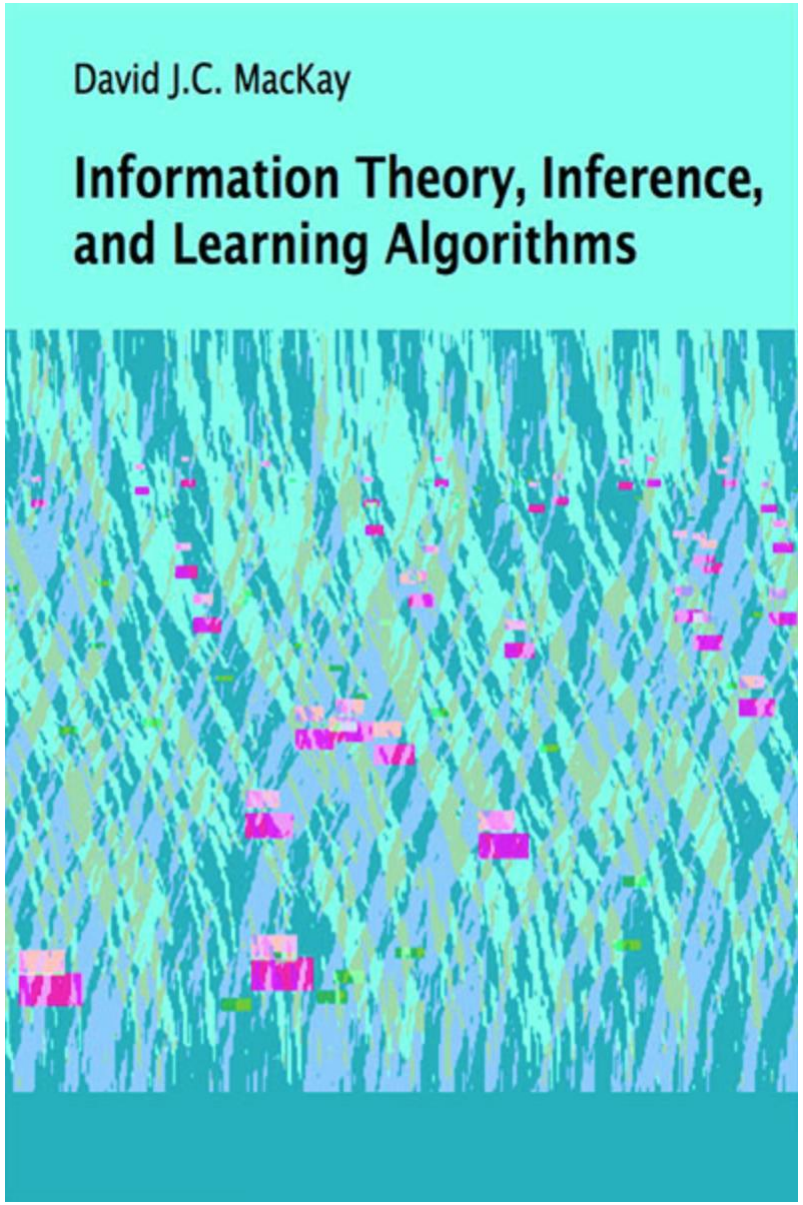
$$= H(X) - H(X|Y)$$

$I(X;Y) \geq 0$ and becomes zero if and only if X and Y are **independent**, i.e., $p(x,y) = p(x)p(y)$.

Figures from Wikipedia

References


WILEY
ELEMENTS OF
INFORMATION
THEORY
SECOND EDITION
THOMAS M. COVER
JOY A. THOMAS


David J.C. MacKay
Information Theory, Inference,
and Learning Algorithms

Variational Inference

Practical Difficulties in Bayesian Methods

- In many cases, it is infeasible to evaluate the posterior or to compute expectations using the posterior.

Posterior: $$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)\, p(\theta)}{p(\mathcal{D})}$$

Marginal likelihood: $$p(\mathcal{D}) = \int p(\mathcal{D}|\theta)\, p(\theta)\, d\theta$$

Predictive distribution: $$p(x|\mathcal{D}) = \int p(x|\theta) p(\theta|\mathcal{D})\, d\theta$$

A Variational Problem

$$\min_{q(\theta)\in\mathcal{P}} \mathbb{E}_{q(\theta)}[-\log p(\mathcal{D}|\theta)] + KL(q(\theta)||p(\theta))$$

$\mathcal{P}$: the space of probability distributions

The optimal solution to this variational problem is

$$q^*(\theta) = p(\theta|\mathcal{D}) \in \mathcal{P}.$$

A Variational Problem

$$\mathbb{E}_{q(\theta)}[-\log p(\mathcal{D}|\theta)] + KL(q(\theta)||p(\theta))$$

$$= \mathbb{E}_{q(\theta)}[-\log p(\mathcal{D}|\theta)p(\theta)] - H(q)$$

$$= \mathbb{E}_{q(\theta)}[-\log p(\mathcal{D})p(\theta|\mathcal{D})] - H(q)$$

$$= \underline{-\log p(\mathcal{D})} + \underline{\mathrm{KL}(q(\theta)||p(\theta|\mathcal{D}))}$$

Independent of $q(\theta)$.

Minimized when $q(\theta) = p(\theta|\mathcal{D})$.

This functional is minimized when $q^*(\theta) = p(\theta|\mathcal{D}) \in \mathcal{P}$.

Variational Inference

- Consider the following variational problem for $q \in \mathcal{Q}$ in a tractable family $\mathcal{Q}$.

$$\min_{q(\theta)\in\mathcal{Q}} \mathbb{E}_{q(\theta)}[-\log p(\mathcal{D}|\theta)] + KL(q(\theta)||p(\theta))$$

We do not need intractable $p(\theta|\mathcal{D})$ or $p(\mathcal{D})$ but $p(\mathcal{D}|\theta), p(\theta), q(\theta)$ → tractable!

The optimal solution $q^*(\theta) \in \mathcal{Q}$ approximates the posterior $p(\theta|\mathcal{D})$.

Different choices of $\mathcal{Q}$ are possible: multivariate Gaussian, $q(\theta) = \prod_{\mathrm{i}=1}^{\mathrm{M}} q_i(\theta_i)$, …

The Evidence Lower Bound (ELBO)

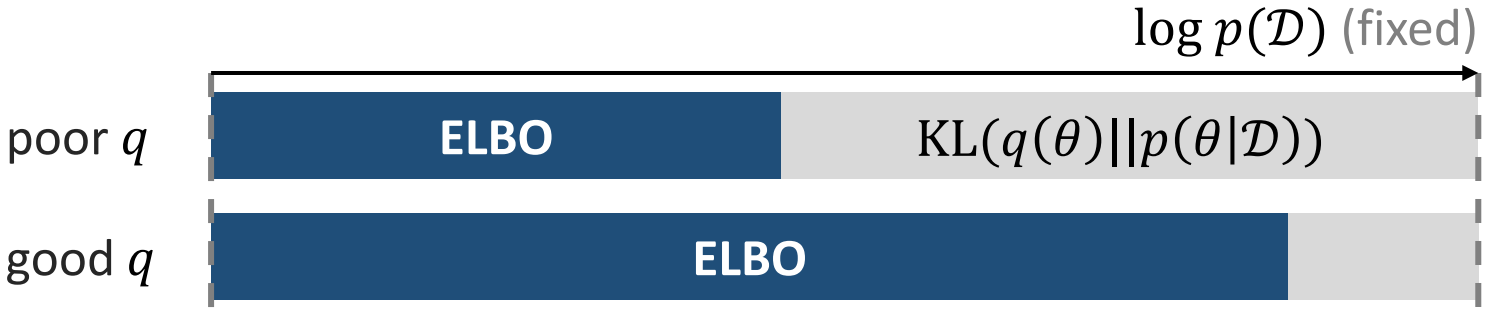
- For any distribution $q(\theta)$, the log-evidence splits exactly:

$$\log p(\mathcal{D}) = \mathbb{E}_{q(\theta)}[\log p(\mathcal{D}|\theta)] - KL(q(\theta)||p(\theta)) + KL(q(\theta)||p(\theta|\mathcal{D}))$$

$$\geq \mathbb{E}_{q(\theta)}[\log p(\mathcal{D}|\theta)] - KL(q(\theta)||p(\theta))$$

The evidence lower bound (ELBO)

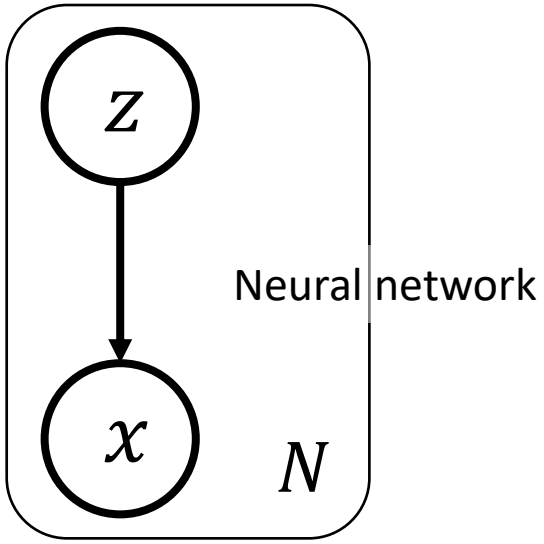
The former variational problem is equivalent to **maximizing the ELBO** with respect to $q \in \mathcal{Q}$:



Optimizing q grows the bound, shrinks the gap.

Variational Autoencoder (VAE)

- Consider a latent variable model.



x: observed data

z: unobserved latent variables

$p(x|z,w) = N\left(g(z,w), \sigma^2 I\right)$: likelihood model

$p(z) = N(0, I)$: prior

To fit the data, we want to maximize the (marginal) likelihood

$$p(x|w) = \int p(x|z,w)p(z)dz\,.$$

However, it is usually intractable. → Variational inference!

Variational Autoencoder (VAE)

$$\log p(\mathcal{D}|w) = \sum_{i=1}^{N} \mathcal{L}_i + \sum_{i=1}^{N} KL(q_i(z_i)||p(z_i|x_i, w))$$

The evidence lower bound (ELBO)

$\mathcal{D} = \{x_i\}_{i=1}^N$: Data (i.i.d.)
$\{z_i\}_{i=1}^N$: Latent variables

$$\mathcal{L}_i = \int q_i(z_i) \log p(x_i|z_i, w) dz_i - KL(q_i(z_i)||p(z_i))$$

Note: Each latent variable has its own distribution $q_i(z_i)$ to be optimized separately.

Variational Autoencoder (VAE)

- Amortized inference

It is inefficient to optimize every $q_i(z_i)$ separately.

Introduce an **inference model** $q(z|x,\phi)$, a parametric model to approximate all $p(z_i|x_i,w)$s.

Usually set $q(z|x,\phi)$ as a Gaussian with a diagonal covariance:

$$q(z|x,\phi) = \prod_{j=1}^{M} N\left(z_{(j)}\middle|\mu_{(j)}(x,\phi),\sigma^2_{(j)}(x,\phi)\right), \qquad z = \left(z_{(1)},\dots,z_{(M)}\right).$$

Neural networks

Let's try to perform a **gradient-based optimization** for w and ϕ.

Variational Autoencoder (VAE)

- The reparametrization trick

$$\mathcal{L}_i = \int q(z_i|x_i,\phi)\log p(x_i|z_i,w)dz_i - KL(q(z_i|x_i,\phi)||p(z_i))$$

analytically evaluated: $\frac{1}{2}\sum_{\mathrm{j}=1}^{\mathrm{M}} 1 + \log\sigma_{(j)}^2(x_i) - \mu_{(j)}^2(x_i) - \sigma_{(j)}^2(x_i)$

Consider a Monte Carlo approximation for the first term:

$$\frac{1}{L}\sum_{l=1}^{L}\log p\left(x_i\middle|z_i^{(l)},w\right), \qquad z_i^{(l)} \sim q(z_i|x_i,\phi)$$

Note: When the samples $z_i^{(l)}$ are fixed, changes in ϕ do not alter $z_i^{(l)}$.

→ We do not have a way to obtain the gradient of $z_i^{(l)}$ with respect to ϕ.

Variational Autoencoder (VAE)

- The reparametrization trick

$$z = \sigma\epsilon + \mu, \qquad \epsilon \sim N(0,1) \quad \Rightarrow \quad z \sim N\left(\mu, \sigma^2\right)$$

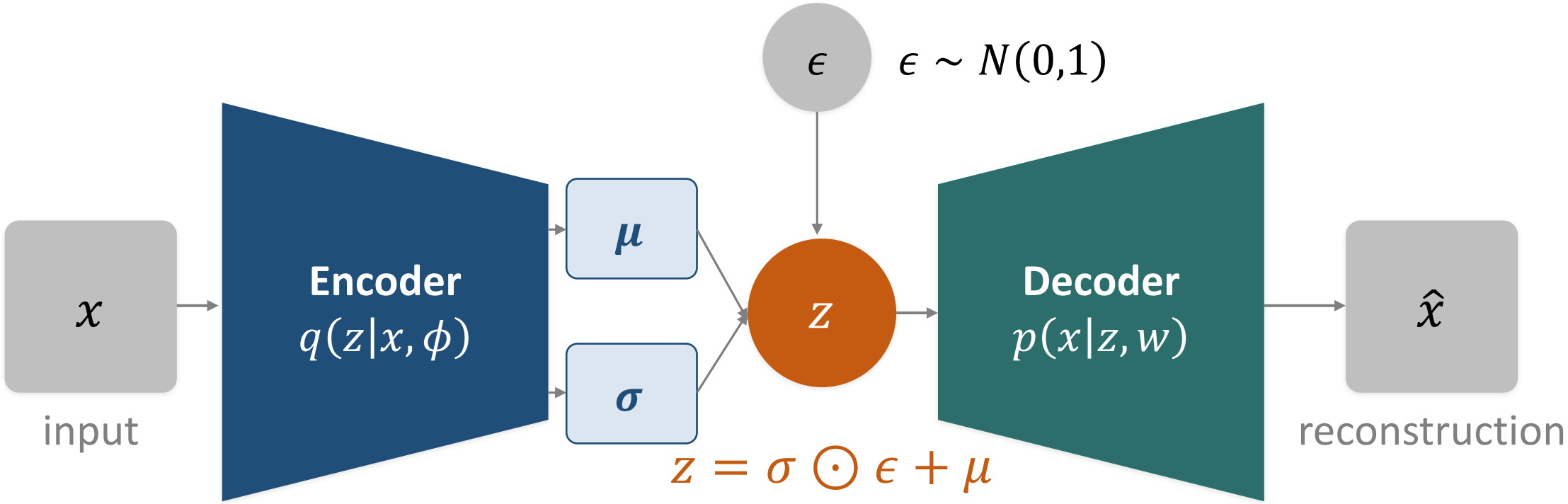
$$z_{i(j)}^{(l)} = \sigma_{(j)}(x_i, \phi)\epsilon_{i(j)}^{(l)} + \mu_{(j)}(x_i, \phi)$$

$$\mathcal{L} \simeq \sum_i \left[\frac{1}{2} \sum_{j=1}^{M} (1 + \log \sigma_{i(j)}^2 - \mu_{i(j)}^2 - \sigma_{i(j)}^2) + \frac{1}{L} \sum_{l=1}^{L} \log p\left(x_i \middle| z_i^{(l)}, w\right) \right]$$

$$\mu_{i(j)} = \mu_{(j)}(x_i, \phi), \qquad \sigma_{i(j)} = \sigma_{(j)}(x_i, \phi)$$

We can perform a **gradient-based optimization** for w and ϕ.

Variational Autoencoder (VAE)



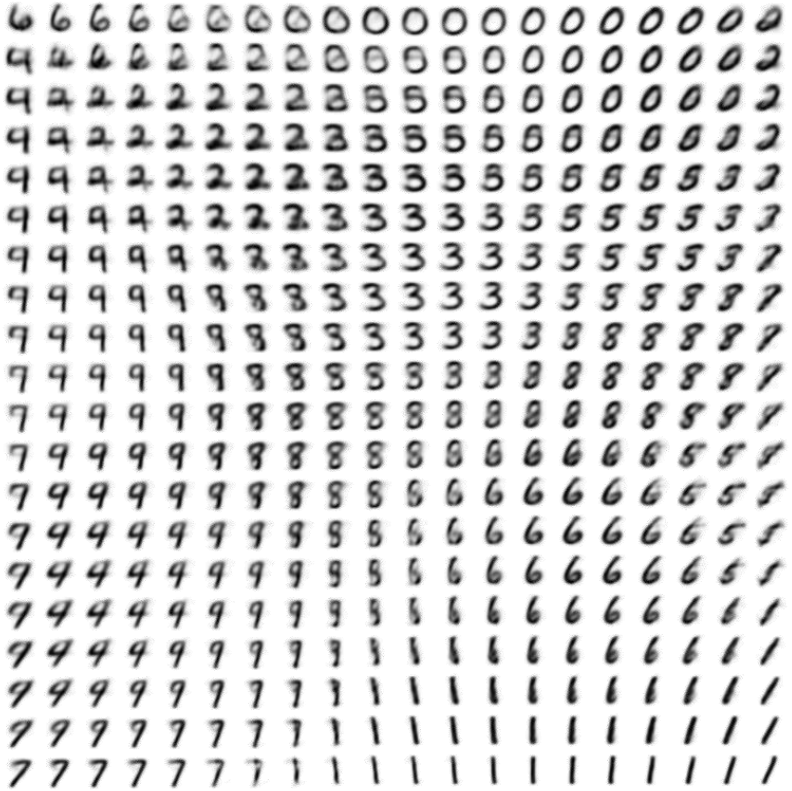
(Negative) ELBO objective:

$$\min_{w,\phi} -\mathcal{L} = \sum_i \underbrace{\left[\mathbb{E}_{q(z_i|x_i,\phi)}[-\log p(x_i|z_i,w)]\right]}_{\text{reconstruction error}} + \underbrace{KL(q(z_i|x_i,\phi)||p(z_i))}_{\text{regularizer}}$$

Variational Autoencoder (VAE)

- Sampling

(a) Learned Frey Face manifold



(b) Learned MNIST manifold

Figure 4: Visualisations of learned data manifold for generative models with two-dimensional latent space, learned with AEVB. Since the prior of the latent space is Gaussian, linearly spaced coordinates on the unit square were transformed through the inverse CDF of the Gaussian to produce values of the latent variables $\mathbf{z}$. For each of these values $\mathbf{z}$, we plotted the corresponding generative $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ with the learned parameters $\boldsymbol{\theta}$.

$$z \sim N(0, I)$$

$$x \sim p(x|z, w)$$

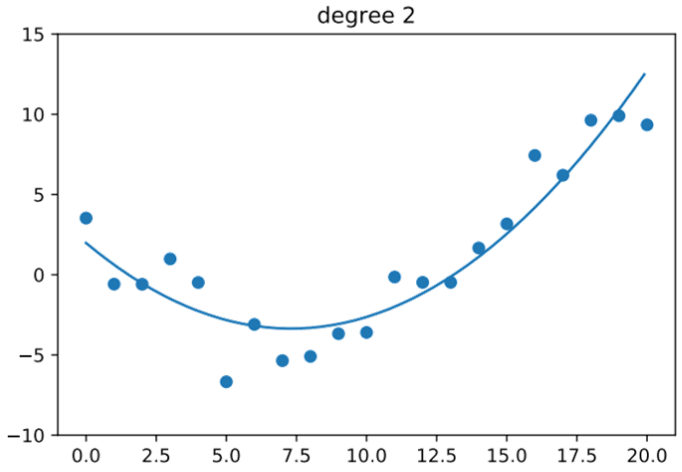
- Many extensions and applications exist.
- Connections to diffusion models can be made.

Figures from Kingma & Welling (2014)

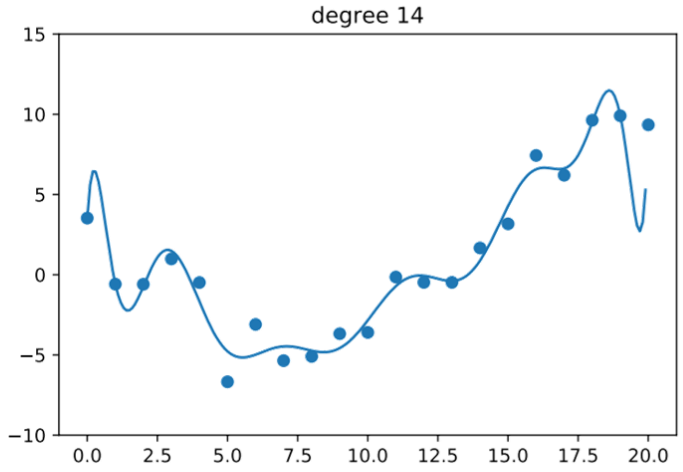
Thank you!

Appendix

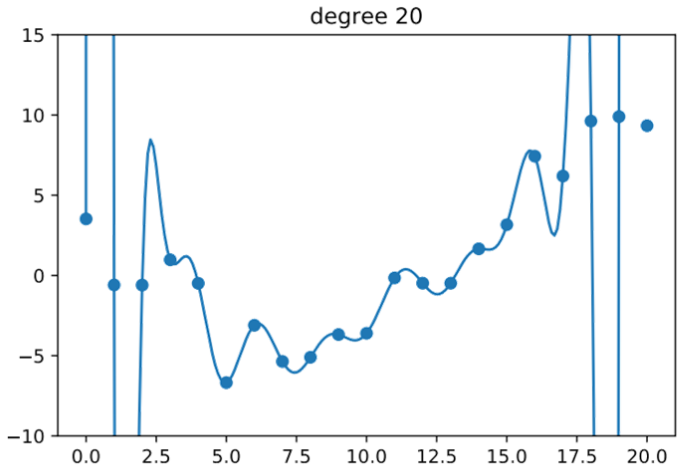
Overfitting and Generalization

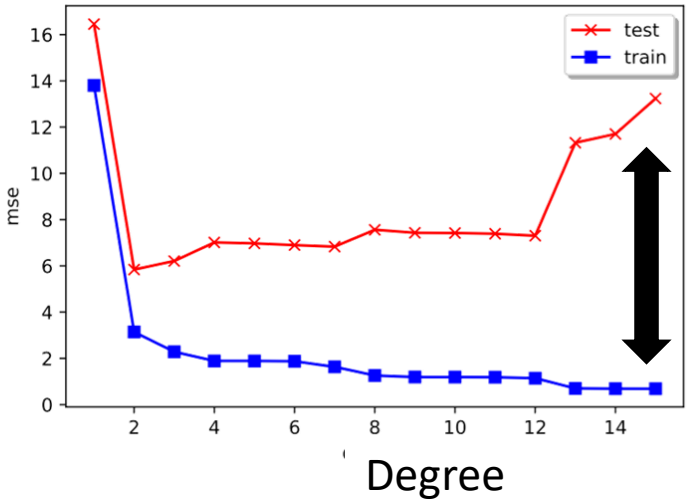


(a)



(b)





A curve fitting example:

Fit data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ using a polynomial model

$f(x;\theta) = a_0 x^0 + a_1 x^1 + \cdots + a_p x^p$, where $\theta = (a_0, \ldots, a_p)$ is the parameter.

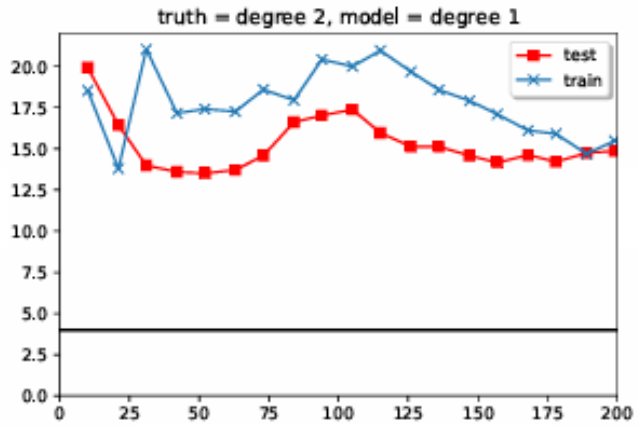
To obtain θ, we minimize the following loss function:

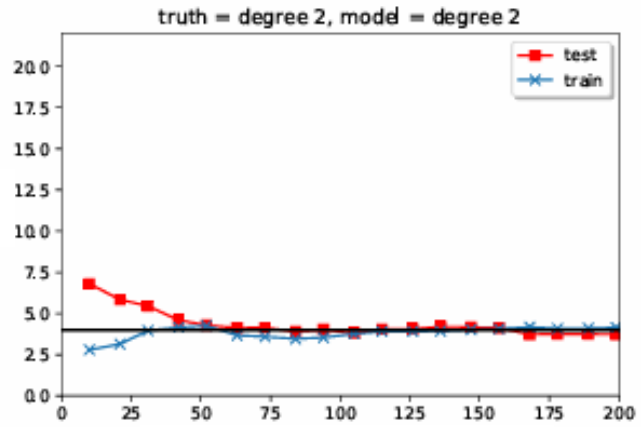
$$\min_\theta \frac{1}{2} \sum_{i=1}^N \|f(x_i;\theta) - y_i\|^2 .$$

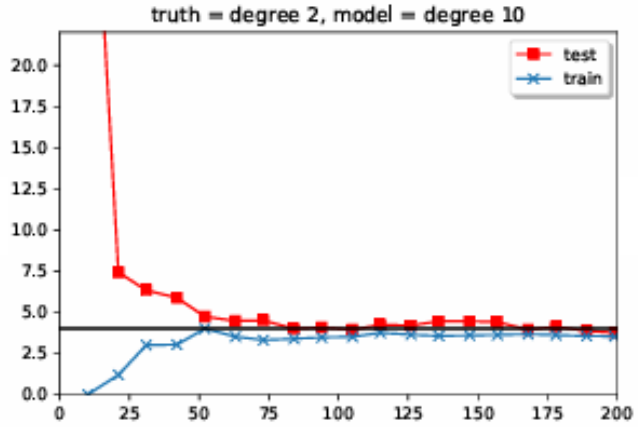
Generalization gap

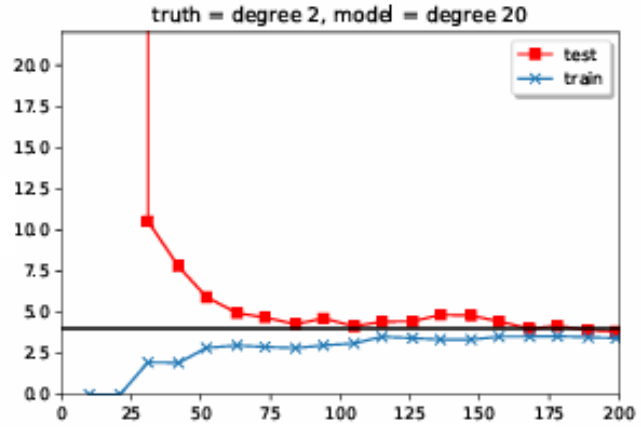
K. Murphy, Probabilistic Machine Learning: An Introduction (2023)

Using More Data









→ Bayes error (the error of the optimal predictor due to inherent noise in the data)

K. Murphy, Probabilistic Machine Learning: An Introduction (2023)

Example: A Bent Coin

Let p_a be the probability of heads, which is unknown.
The probability that F tosses results in a sequence s of F_a heads and F_b tails is

$$P(s|p_a, F) = p_a^{F_a}(1-p_a)^{F_b}. \quad \text{(likelihood)}$$

Assumption: a uniform prior for p_a, i.e., $p(p_a) = 1$ for $p_a \in [0,1]$

Bayes theorem:

$$p(p_a|s, F) = \frac{P(s|p_a, F)p(p_a)}{P(s|F)} = \frac{p_a^{F_a}(1-p_a)^{F_b}}{P(s|F)}$$

The normalizing constant is given by the Beta integral:

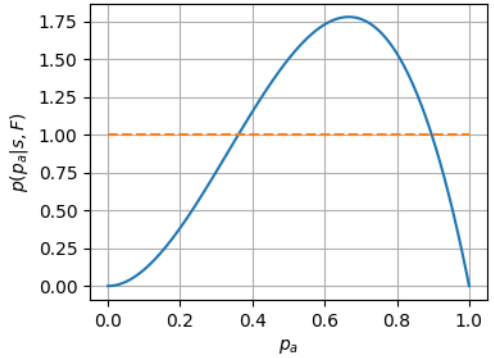
$$P(s|F) = \int_0^1 \underbrace{p_a^{F_a}(1-p_a)^{F_b}}_{P(s,p_a|F)} dp_a = \frac{F_a!\,F_b!}{(F_a+F_b+1)!}$$

Example from D. MacKay (2003), Figure from C. Bishop (2023)

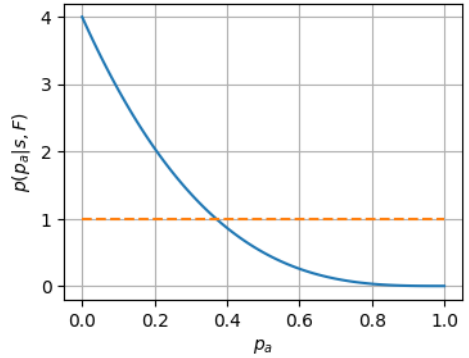
Example: A Bent Coin

Posterior distributions $p(p_a|s,F)$:

Ex) $s = aba, F = 3$



Ex) $s = bbb, F = 3$



Cf) Frequentists' result: $\frac{F_a}{F_a+F_b}$

Prediction for the next toss:

$$P(a|s,F) = \int_0^1 P(a|p_a)p(p_a|s,F)dp_a = \int_0^1 p_a \frac{p_a^{F_a}(1-p_a)^{F_b}}{P(s|F)} dp_a$$

$$= \frac{(F_a+1)!\,F_b!}{(F_a+F_b+2)!} \Big/ \frac{F_a!\,F_b!}{(F_a+F_b+1)!}$$

$$= \frac{F_a+1}{F_a+F_b+2}$$

Example from D. MacKay (2003), Figure from C. Bishop (2023)

Predictive Distribution for Gaussian

$$p(x|\mu)p(\mu|\mathcal{D}) \propto \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)\exp\left(-\frac{1}{2\sigma_N^2}(\mu-\mu_N)^2\right)$$

$$\propto \exp\left(-\frac{1}{2}\left(\frac{x^2}{\sigma^2}+\left(\frac{1}{\sigma^2}+\frac{1}{\sigma_N^2}\right)\mu^2-2\left(\frac{x}{\sigma^2}+\frac{\mu_N}{\sigma_N^2}\right)\mu\right)\right)$$

$$\propto \exp\left(-\frac{1}{2}\left(\frac{x^2}{\sigma^2}+\left(\frac{1}{\sigma^2}+\frac{1}{\sigma_N^2}\right)\left(\mu-\left(\frac{1}{\sigma^2}+\frac{1}{\sigma_N^2}\right)^{-1}\left(\frac{x}{\sigma^2}+\frac{\mu_N}{\sigma_N^2}\right)\right)^2-\left(\frac{1}{\sigma^2}+\frac{1}{\sigma_N^2}\right)^{-1}\left(\frac{x}{\sigma^2}+\frac{\mu_N}{\sigma_N^2}\right)^2\right)\right)$$

$$p(x|\mathcal{D}) = \int p(x|\mu)p(\mu|\mathcal{D})\,d\mu$$

Marginalize out μ.

$$p(x|\mathcal{D}) \propto \exp\left(-\frac{1}{2}\left(\left(\frac{1}{\sigma^2}-\frac{1}{\sigma^4}\left(\frac{1}{\sigma^2}+\frac{1}{\sigma_N^2}\right)^{-1}\right)x^2-2\frac{\mu_N}{\sigma^2\sigma_N^2}\left(\frac{1}{\sigma^2}+\frac{1}{\sigma_N^2}\right)^{-1}x\right)\right)$$

$$\propto \exp\left(-\frac{(x-\mu_N)^2}{2(\sigma^2+\sigma_N^2)}\right) \quad \Rightarrow \quad p(x|\mathcal{D}) = N(x;\mu_N,\sigma^2+\sigma_N^2)$$

Uncertainty from model

Predictive Distribution for Bayesian Linear Regression

- Changing the expressions:

$$\mu_{y|x,\mathcal{D}} = x^\top \left(XX^\top + \frac{\sigma^2}{\sigma_0^2} I \right)^{-1} Xy = x^\top X \left(X^\top X + \frac{\sigma^2}{\sigma_0^2} I \right)^{-1} y$$

$$(XX^\top + \alpha I)X = X(X^\top X + \alpha I)$$

$$\sigma^2_{y|x,\mathcal{D}} = \sigma^2 + \sigma^2 \cdot x^\top \left(XX^\top + \frac{\sigma^2}{\sigma_0^2} I \right)^{-1} x = \sigma^2 + \sigma_0^2 x^\top x - \sigma_0^2 x^\top X \left(X^\top X + \frac{\sigma^2}{\sigma_0^2} I \right)^{-1} X^\top x$$

$$(XX^\top + \alpha I)^{-1} = \alpha^{-1}(I - X(X^\top X + \alpha I)^{-1} X)$$

from the Woodbury matrix inversion lemma

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$$

➡ All the results are based on inner products, e.g., $x_i^\top x_j$, $x_i^\top x$, and $x^\top x$.

→ Extend to **kernel methods!**

Selecting Priors: Empirical Bayes Method

- Bayesian inference utilizes the **prior distribution** to make predictions. In many cases, however, we may **not** have any useful assumptions to make for the prior distribution.

  An idea: Use data $\mathcal{D} = \{x_i\}_{i=1}^N$ to objectively set the prior.

  Let the prior distribution be $p(\theta; \beta)$, where β is a parameter to control the prior, often called the **hyperparameter**.

  Marginal likelihood: $p(\mathcal{D}; \beta) = \int \prod_{i=1}^N p(x_i|\theta)\, p(\theta; \beta)\, d\theta$

- Empirical Bayes method:

  Set the prior as $p(\theta; \beta_{EB})$, where $\beta_{EB} = \underset{\beta}{\operatorname{argmax}}\, p(\mathcal{D}; \beta)$.

Selecting Priors: Empirical Bayes Method

- Example: log marginal likelihood of Gaussian

$$p(x|\mu) = N(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$p(\mu;\beta) = N(\mu;0,\beta^2) = \frac{1}{\sqrt{2\pi\beta^2}}\exp\left(-\frac{\mu^2}{2\beta^2}\right)$$

$$p(\mathcal{D};\beta) = \int \prod_{i=1}^{N} p(x_i|\mu)\ p(\mu;\beta)\ d\mu$$

$$\Rightarrow \quad \log p(\mathcal{D};\beta) = -N\log\left(\sqrt{2\pi}\sigma\right) - \frac{1}{2}\log\left(1+\frac{N\beta^2}{\sigma^2}\right) + \frac{1}{2}\left(\frac{1}{\beta^2}+\frac{N}{\sigma^2}\right)^{-1}\left(\sum_{i=1}^{N}\frac{x_i}{\sigma^2}\right)^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{N}x_i^2$$

- Empirical Bayes method:

$$\beta_{EB} = \underset{\beta\geq 0}{\operatorname{argmax}}\, p(\mathcal{D};\beta) \quad \Rightarrow \quad \beta_{EB} = \begin{cases} \sqrt{\bar{x} - \frac{\sigma^2}{N}} & \text{if } |\bar{x}| \geq \sqrt{\frac{\sigma^2}{N}} \\ 0 & \text{otherwise} \end{cases}$$

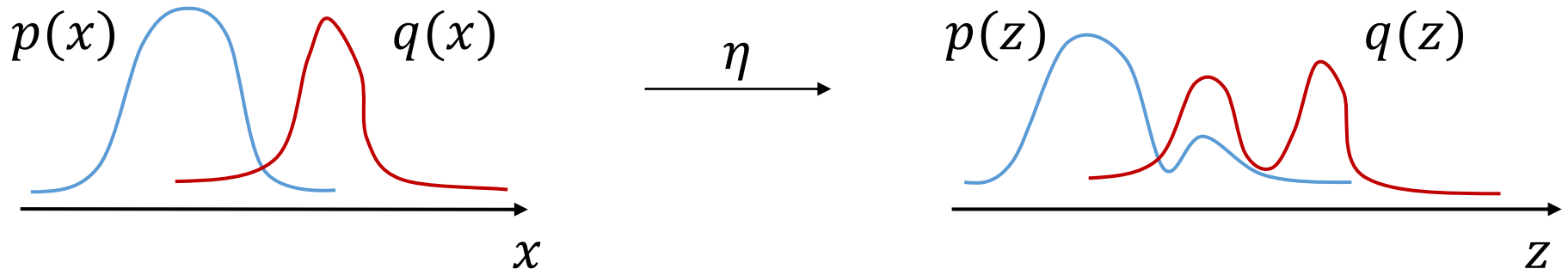
where $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$.

KL-Divergence

- Reparametrization invariance

  Under a reparametrization $x \mapsto z = \eta(x)$, the probability density function transforms as follows:

$$p(x) = p(z) \cdot \left| \frac{dz}{dx} \right|.$$



$$KL(p(x)\|q(x)) = \int p(x) \log \frac{p(x)}{q(x)}\, dx = \int p(z) \log \frac{p(z) \cdot \left| \frac{dz}{dx} \right|}{q(z) \cdot \left| \frac{dz}{dx} \right|}\, dz = KL(p(z)\|q(z))$$

Q. How about the differential entropy $H(X)$?

Connections to Information Geometry

- The **Fisher information matrix** (FIM):

$$F_{ij}(\theta) = \mathbb{E}_{(x)\sim p(x;\theta)}\left[\frac{\partial \log p(x;\theta)}{\partial \theta_i}\frac{\partial \log p(x;\theta)}{\partial \theta_j}\right]$$

- Connection between **KL-divergence** and the FIM:

$$KL(p(x;\theta+d\theta)\|p(x;\theta)) = \frac{1}{2}d\theta^{\top}F(\theta)d\theta + o\left(\|d\theta\|^2\right)$$

f-Divergence

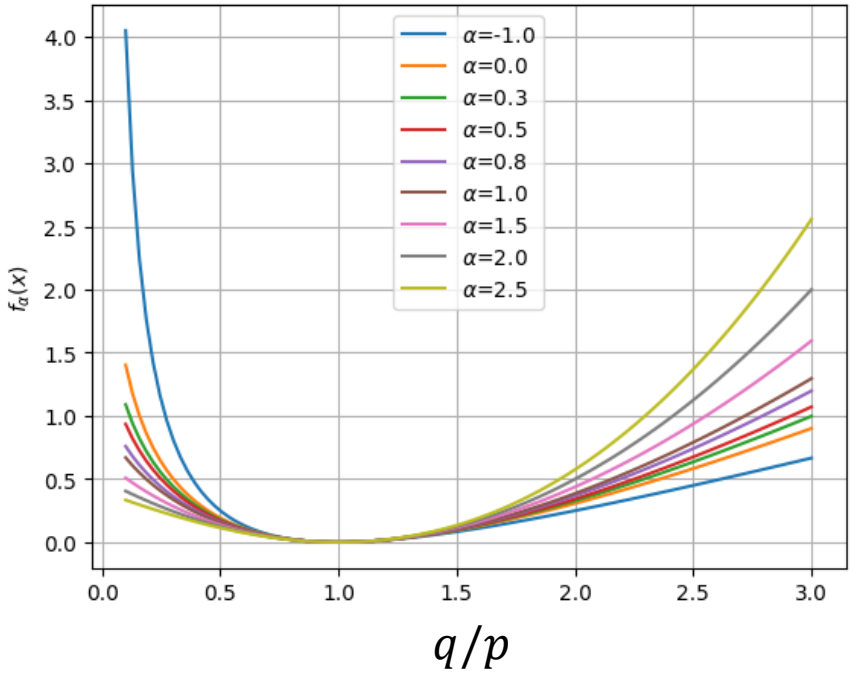
- The f-divergence between two PDFs $p(x)$ and $q(x)$:

$$D_f(p\|q) = \int p(x) f\left(\frac{q(x)}{p(x)}\right) dx$$

where f is a convex function satisfying $f(1) = 0$ and strictly convex at $x = 1$.

Ex) α-divergence

$$f(x) = \frac{1}{\alpha(1-\alpha)}(1 - x^{\alpha} - \alpha(1-x))$$



Some properties of f-divergence:

$D_f(p||q) \geq 0$ with equality if and only if $p(x) = q(x)$ for all x.

Asymmetric.

Reparametrization invariance.

Connections to FIM: $D_f(p(x;\theta + d\theta)\|p(x;\theta)) = \frac{f''(1)}{2} d\theta^{\top} F(\theta) d\theta + o\left(\|d\theta\|^2\right)$