

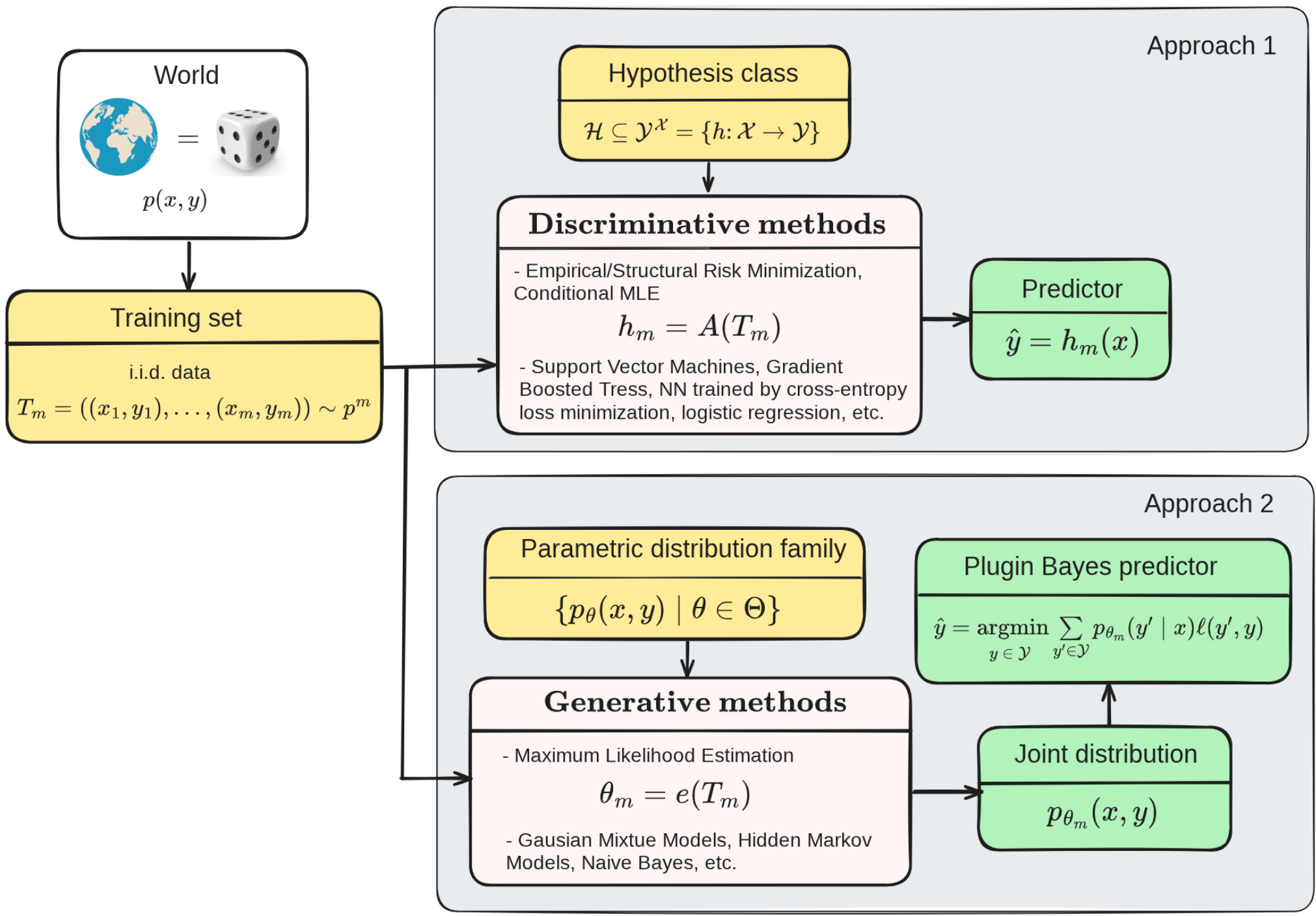
# Statistical Machine Learning (BE4M33SSU)

## Generative learning, Maximum Likelihood Estimator

Czech Technical University in Prague  
B. Flach, V. Franc

- ◆ Discriminative vs. generative learning.
- ◆ When do we need generative learning?
- ◆ Parametric distribution families
- ◆ Maximum Likelihood Estimator and its properties

# Discriminative vs. Generative Learning



# Discriminative learning

**Goal:** train a classifier  $y = h(x)$  for an unknown distribution  $p(x, y)$  over features  $x \in \mathcal{X}$  and classes  $y \in \mathcal{Y}$

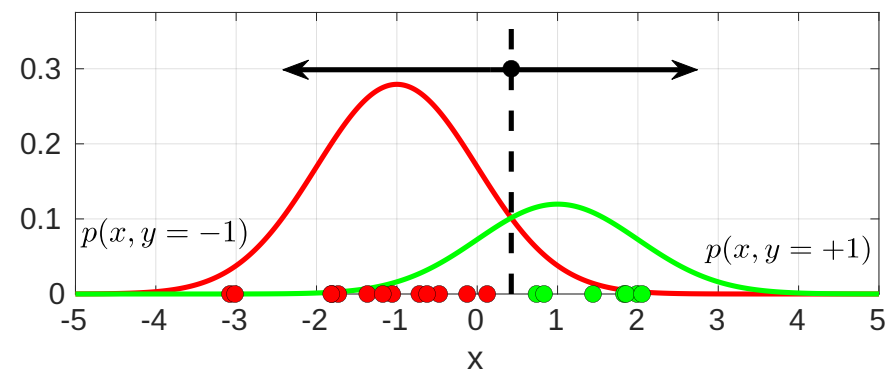
## Discriminative learning:

- ◆ define a hypothesis space  $\mathcal{H}$  of predictors  $h: \mathcal{X} \rightarrow \mathcal{Y}$  and fix a loss  $\ell(y, y')$
- ◆ given a training set  $T_m$ , learn  $h_m: \mathcal{X} \rightarrow \mathcal{Y}$  by empirical (or structural) risk minimization.

**Example 1** (Gaussian discriminative analysis). Assume we know:  $\mathcal{X} = \mathbb{R}$ ,  $\mathcal{Y} = \{-1, +1\}$

$$p(x, y) = p(y) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x - \mu_y)^2}$$

with unknown  $p(y = 1)$ ,  $\mu_+ > \mu_-$  and  $\sigma$ .



The loss is  $\ell(y, y') = \mathbb{I}[y' \neq y]$  and the training set is  $T_m = ((x_1, y_1), \dots, (x_m, y_m))$ .

- ◆ The Bayes optimal predictor for each such model is in hypothesis space  $\mathcal{H} = \{h(x) = \text{sign}(x - \gamma) \mid \gamma \in \mathbb{R}\}$ , so we apply the empirical risk minimization:

$$\gamma_m = \arg \min_{\gamma \in \mathbb{R}} \left[ \frac{1}{m} \sum_{i=1}^m \mathbb{I}[y_i \neq \text{sign}(x_i - \gamma)] \right]$$

and output the predictor  $h_m(x) = \text{sign}(x - \gamma_m)$ .

# Generative learning

**Generative learning:** Use prior knowledge to restrict the search to a parametric family of distributions  $\{p_\theta(x, y) \mid \theta \in \Theta\}$ . Learning algorithm:

1. Given training data  $T_m$ , estimate the unknown parameter  $\theta_m = e(T_m)$  e.g. using the maximum likelihood estimator.
2. Consider  $p_{\theta_m}(x, y)$  as the true model. Predict hidden states by the plugin Bayes optimal predictor

$$h(x) = \operatorname{arg\,min}_{y \in \mathcal{Y}} \sum_{y' \in \mathcal{Y}} p_{\theta_m}(y' \mid x) \ell(y', y).$$

**Example 1** (cont.). Given  $T_m$ , the estimates of the model parameters are

$$p(y = 1) = \frac{m_+}{m} \quad \mu_+ = \frac{1}{m_+} \sum_i x_i \llbracket y_i = 1 \rrbracket \quad \mu_- = \frac{1}{m_-} \sum_i x_i \llbracket y_i = -1 \rrbracket$$

and

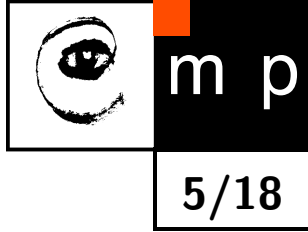
$$\sigma^2 = \frac{1}{m} \sum_i \left( x_i - \mu_+ \llbracket y_i = 1 \rrbracket - \mu_- \llbracket y_i = -1 \rrbracket \right)^2,$$

where  $m_+$  denotes the number of training examples with class  $y_i = 1$ . The predictor is

$$h(x) = \operatorname{sign} \left( \log \frac{p_{\theta_m}(x, y = 1)}{p_{\theta_m}(x, y = -1)} \right) = \dots = \operatorname{sign}(x - \gamma),$$

where  $\gamma$  depends on the estimated  $\mu_+$ ,  $\mu_-$ ,  $\sigma$ ,  $p(y = 1)$  and  $p(y = -1)$ .

# Discriminative vs. Generative Learning



## Discriminative Models

- ◆ Model  $\hat{y} = h(x)$  or  $p(y | x)$  – learn the boundary between classes.
- ◆ Typically require less data and are often more accurate for prediction tasks.
- ◆ Theoretical guarantees – PAC learning.
- ◆ Examples: Support Vector Machines, Gradient Boosted Trees, Logistic Regression, prediction Neural Networks.

## Generative Models

- ◆ Model  $p(x, y)$  or  $p(x | y)$  – learn how the data is generated for each class.
- ◆ Perform tasks beyond prediction, e.g. generate new data samples.
- ◆ Often require more data but provide richer probabilistic understanding.
- ◆ Can naturally handle missing data.
- ◆ Examples: Naive Bayes, Gaussian Mixture Models, Hidden Markov Models, Generative Adversarial Networks.

## Parametric distribution families

A *parametric family of distributions* is a set of distributions  $\{p_\theta(x) \mid \theta \in \Theta\}$  for a r.v.  $X$  which are specified by parameter values.

**Example 2.** *The family of multivariate normal distributions  $\mathcal{N}(\mu, V)$  on  $\mathbb{R}^n$*

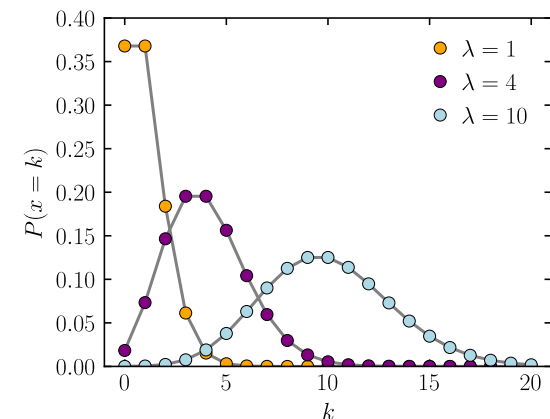
$$p_{\mu, V}(x) = \frac{1}{(2\pi)^{n/2} |V|^{1/2}} \exp \left[ -\frac{1}{2} (x - \mu)^T V^{-1} (x - \mu) \right]$$

*parametrised by the vector  $\mu \in \mathbb{R}^n$  and a positive definite  $n \times n$  matrix  $V$ .*

**Example 3.** *The family of Poisson distributions on  $x \in \mathbb{N}$  with probability mass*

$$p_\lambda(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

*parametrised by  $\lambda \in \mathbb{R}_+$ . Notice that  $\lambda = \mathbb{E}[X] = \mathbb{V}[X]$ .*



Both families are examples of a broad class of distribution families – *exponential families*.

## Parametric distribution families

**Definition 1.** A family of distributions for a random variable  $x \in \mathcal{X}$  is an exponential family if its probability density / probability mass has the form

$$p_{\theta}(x) = h(x) \exp[\langle \phi(x), \theta \rangle - A(\theta)],$$

where

$\phi(x) \in \mathbb{R}^n$  is the sufficient statistics,

$\theta \in \mathbb{R}^n$  is the (natural) parameter,

$h(x)$  is the base measure and

$A(\theta)$  is the cumulant function defined by

$$A(\theta) = \log \int_{\mathbb{R}^n} h(x) \exp[\langle \phi(x), \theta \rangle] dx$$

Remarks:

- ◆ The cumulant function is essentially the logarithm of the normalisation constant.
- ◆ The statistic  $\phi(x)$  is called *sufficient* because when estimating the parameter  $\theta$  from a training set  $T$ , all we need to know from it is  $\frac{1}{m} \sum_{i=1}^m \phi(x_i)$ .

## Parametric distribution families

**Example 4.** Consider the family of Bernoulli distributions for  $x \in \{0, 1\}$  with  $p(x) = \beta^x (1 - \beta)^{1-x}$  parametrised by  $\beta \in (0, 1)$ . It can be written as

$$p_\theta(x) = h(x) \exp[\langle \phi(x), \theta \rangle - A(\theta)]$$

with  $h(x) = 1$ ,  $\phi(x) = x$ ,  $\theta = \log \frac{\beta}{1-\beta}$  and  $A(\theta) = \log(1 + e^\theta)$ .

**Example 5.** Consider the family of univariate normal distributions with unit variance and mean  $\mu$  for  $x \in \mathbb{R}$ . Its density is given by

$$p_\mu(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2}$$

and can be written as

$$p_\theta(x) = h(x) \exp[\langle \phi(x), \theta \rangle - A(\theta)]$$

with  $h(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ ,  $\phi(x) = x$ ,  $\theta = \mu$  and  $A(\theta) = \frac{\theta^2}{2}$ .

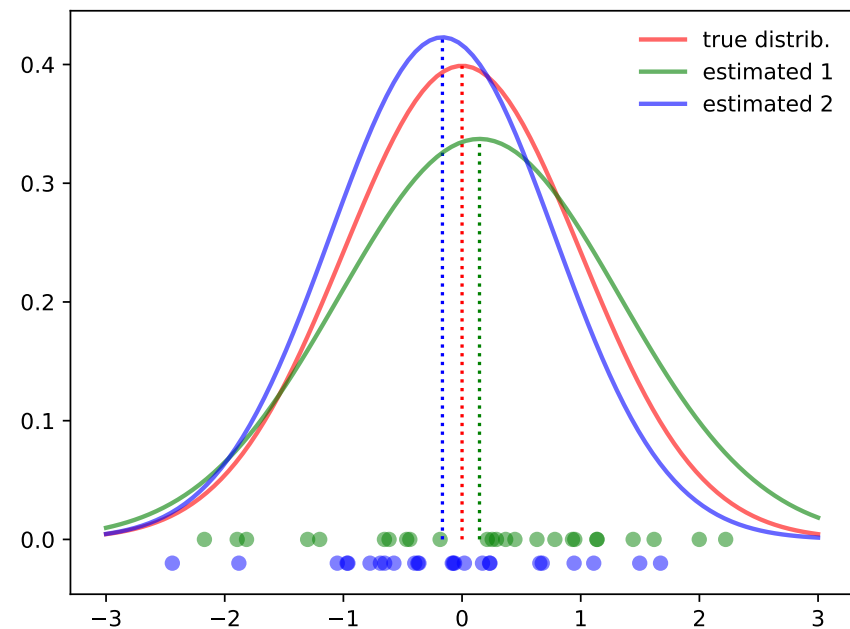
# Parameter estimation

**Given:** a parametric family of distributions  $\{p_\theta(x) \mid \theta \in \Theta\}$  and an i.i.d. training set  $T_m = \{x_i \in \mathcal{X} \mid i = 1, \dots, m\}$  generated from  $p_{\theta^*}(x)$  with unknown  $\theta^*$ .

**Estimator:** a mapping  $\theta_m = e(T_m)$ , which maps training sets to parameters, i.e.  $e: \cup_{m=1}^{\infty} \mathcal{X}^m \rightarrow \Theta$

**Example 6.** Estimating parameters of a normal distribution

- ◆ red: true distribution  $\mathcal{N}(0,1)$
- ◆ blue and green: sample two i.i.d. training sets from it and estimate parameters; e.g.  $\mu_m = e(T_m) = \frac{1}{m} \sum_{i=1}^m x_i$ .

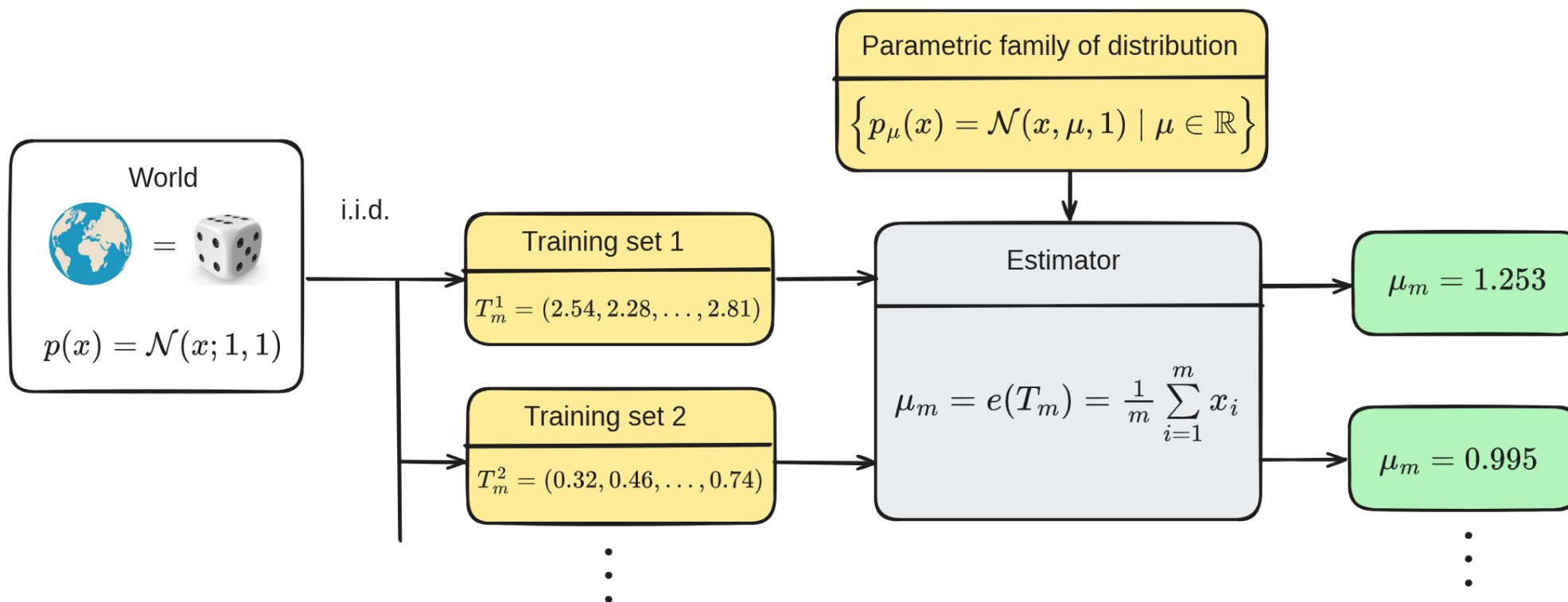


Desired properties of an estimator:

- ◆ the estimator is unbiased i.e.  $\mathbb{E}_{T_m \sim \theta^*} [e(T_m)] = \theta^*$
- ◆ the estimator has small variance  $\mathbb{V}_{T_m \sim \theta^*} [e(T_m)]$
- ◆ the estimator is consistent i.e.  $\mathbb{P}_{T_m \sim \theta^*} (|e(T_m) - \theta^*| > \epsilon) \rightarrow 0$  for  $m \rightarrow \infty$

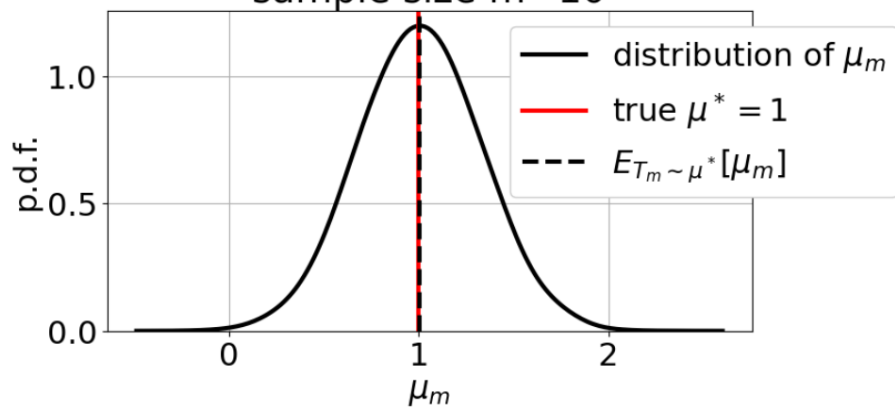
# Parameter estimation

The unbiased estimator:  $\mathbb{E}_{T_m \sim \theta^*} [e(T_m)] = \theta^*$



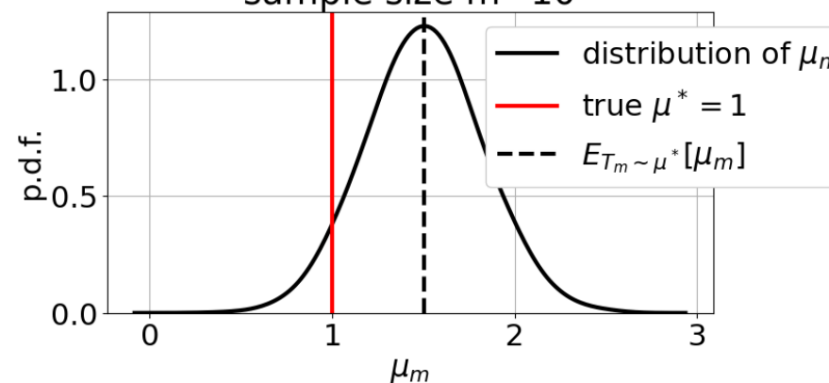
Unbiased estimator

sample size  $m=10$



Biased estimator

sample size  $m=10$

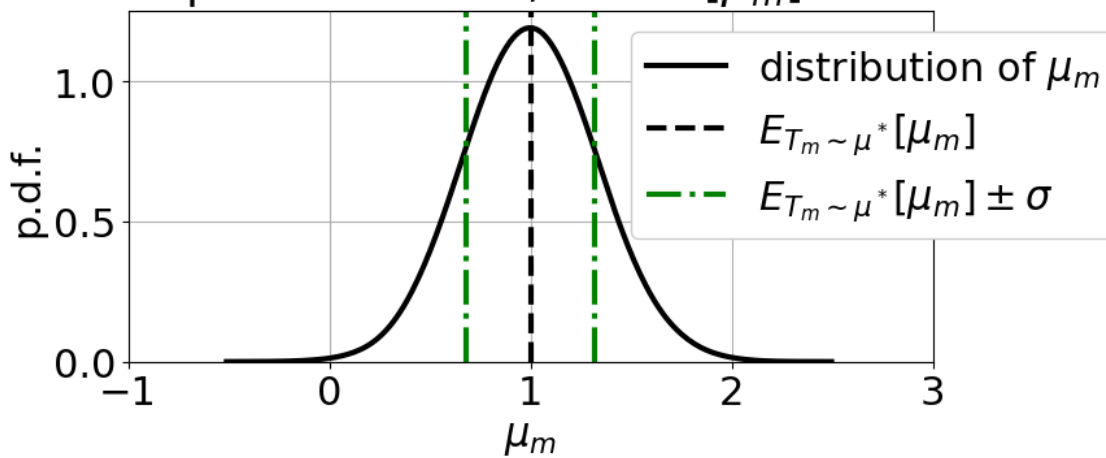


## Parameter estimation

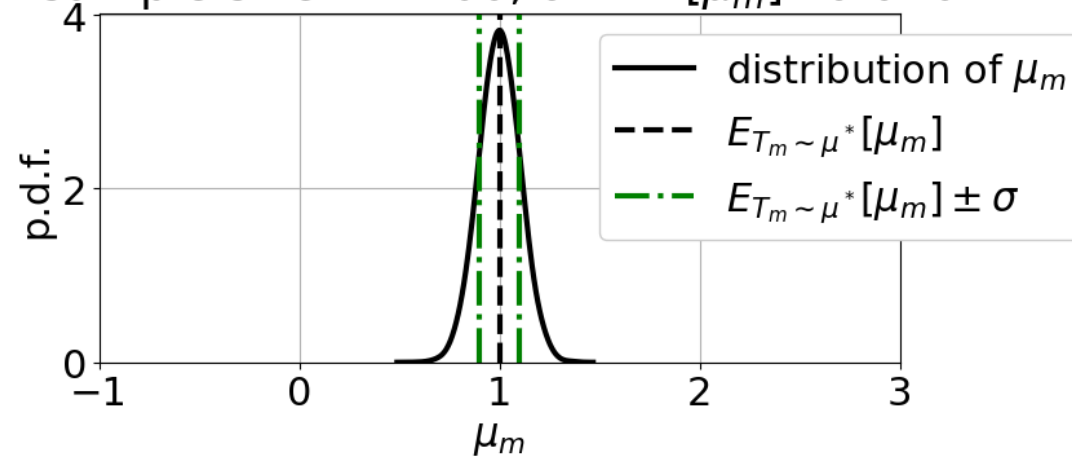
An **efficient estimator** has the smallest possible variance among all unbiased estimators, i.e. it attains the Cramer-Rao lower bound:

$$\text{Var}_{T_m \sim \theta^*} [e(T_m)] \geq \frac{1}{m I(\theta^*)}$$

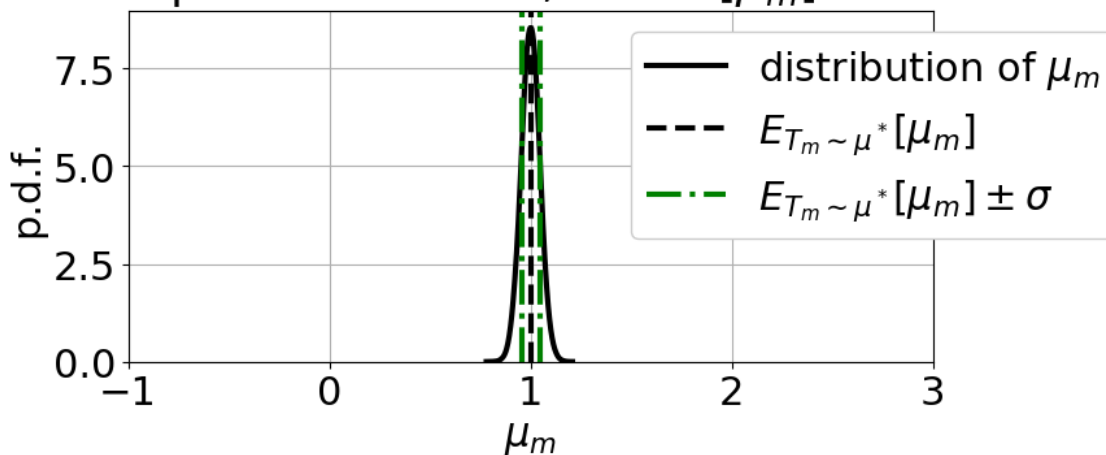
sample size  $m=10$ ,  $\sigma^2 = V[\mu_m]=0.101$



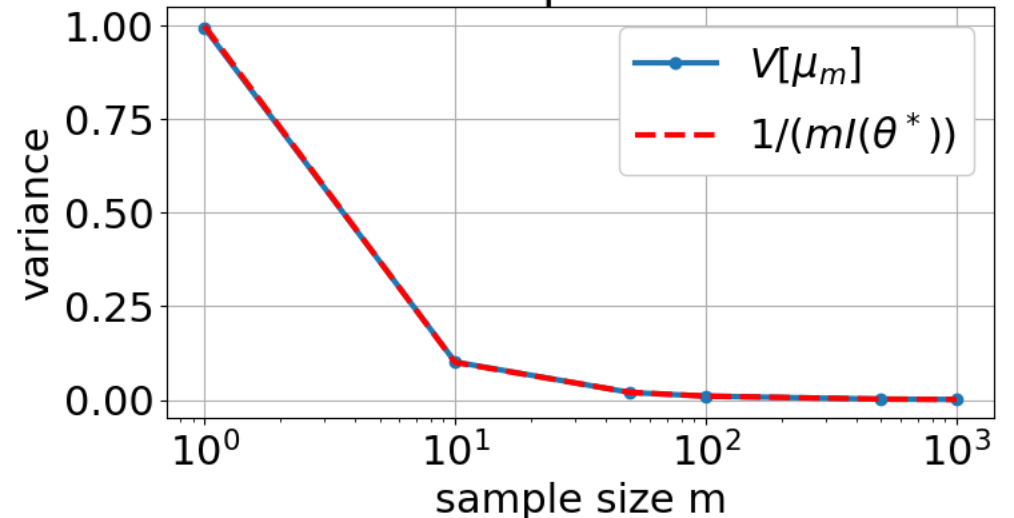
sample size  $m=100$ ,  $\sigma^2 = V[\mu_m]=0.010$



sample size  $m=500$ ,  $\sigma^2 = V[\mu_m]=0.002$



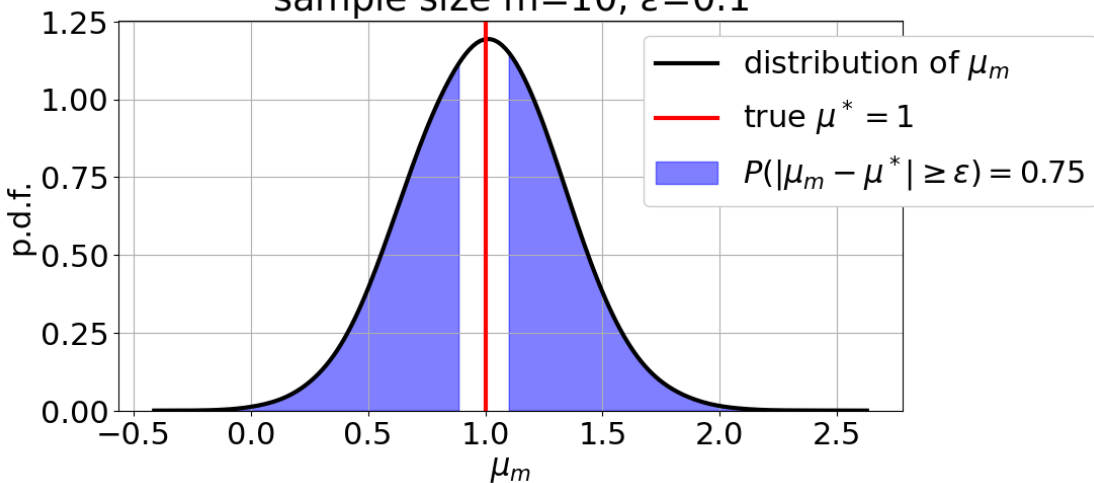
Variance of Sample Mean Estimator



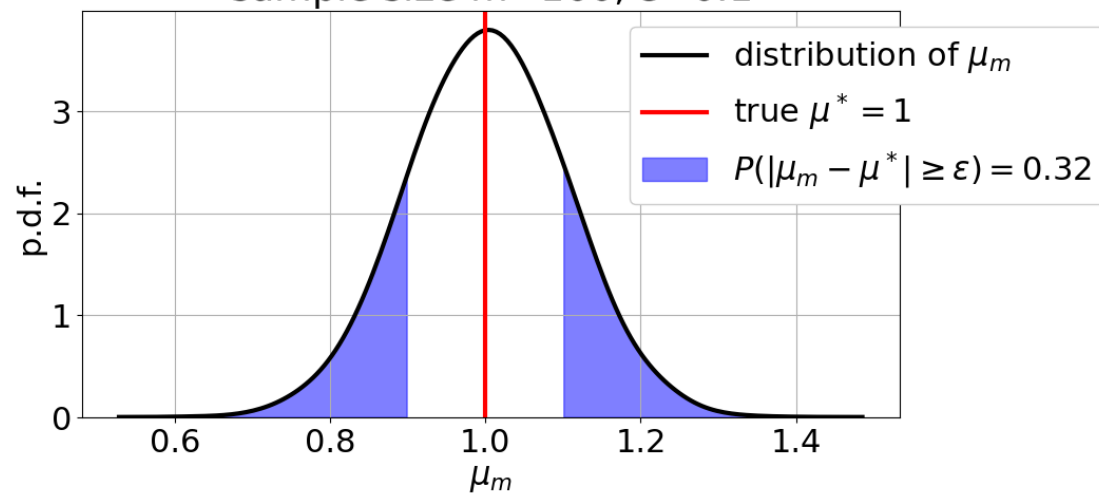
# Parameter estimation

The consistent estimator:  $\mathbb{P}_{T_m \sim \theta^*} (|e(T_m) - \theta^*| > \epsilon) \rightarrow 0$  for  $m \rightarrow \infty$

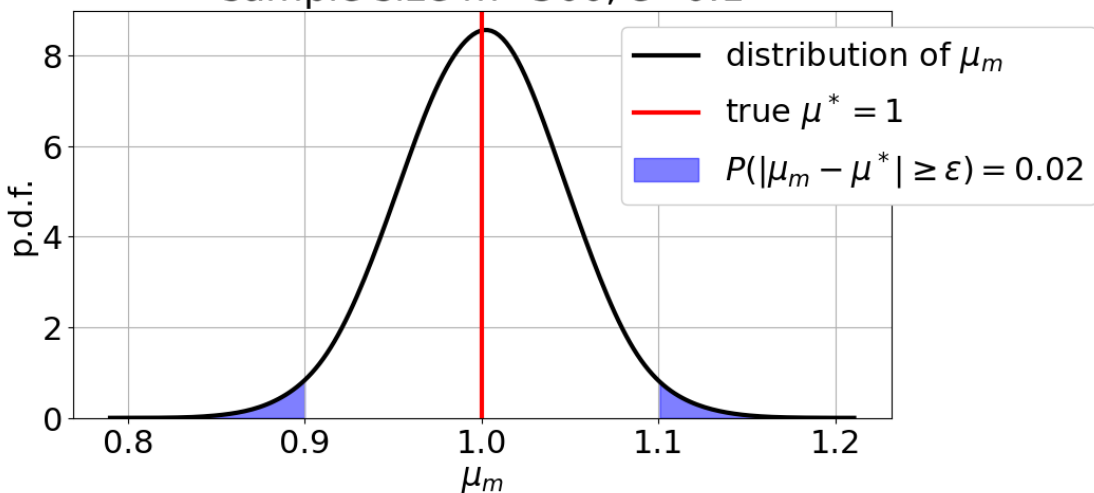
sample size  $m=10$ ,  $\epsilon=0.1$



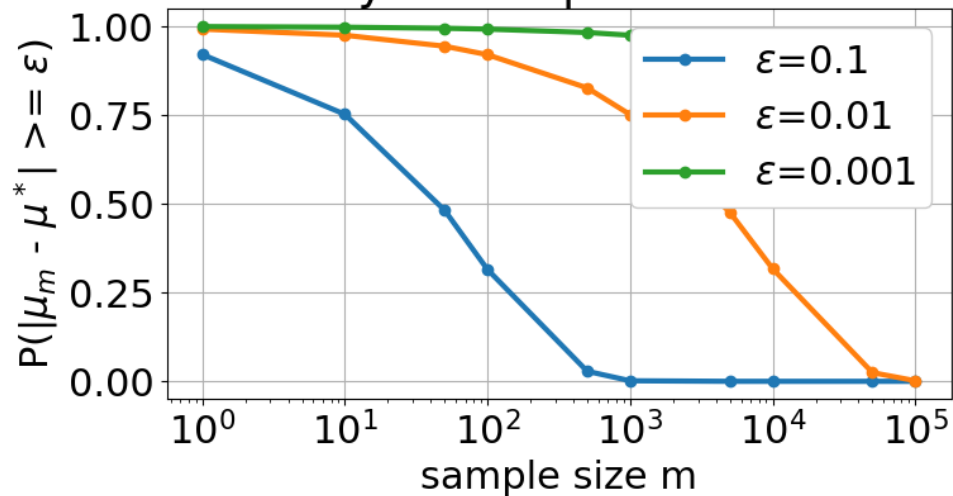
sample size  $m=100$ ,  $\epsilon=0.1$



sample size  $m=500$ ,  $\epsilon=0.1$



Consistency of Sample Mean Estimator



## Maximum Likelihood estimator

**Given:** a parametric family of distributions  $\{p_\theta(x) \mid \theta \in \Theta\}$  and an i.i.d. training set  $T_m = \{x_i \in \mathcal{X} \mid i = 1, \dots, m\}$  generated from  $p_{\theta^*}(x)$  with unknown  $\theta^*$ .

Define the log-likelihood to obtain the given i.i.d. training data  $T_m$  from the distribution with parameter  $\theta \in \Theta$

$$L(T_m, \theta) = \frac{1}{m} \log \left( \prod_{i=1}^m p_\theta(x_i) \right) = \frac{1}{m} \sum_{i=1}^m \log p_\theta(x_i)$$

Remarks:

- ◆ We normalize the log-likelihood by the sample size for notational convenience.
- ◆ If  $\mathcal{X}$  is finite,  $L(T_m, \theta)$  is proportional to the logarithm of the probability that  $T_m$  was generated from  $p_\theta(x_1, \dots, x_m) = \prod_{i=1}^m p_\theta(x_i)$ .

The **Maximum Likelihood estimator** predicts the parameter  $\theta_m$  that maximizes the (log-)likelihood of the training data

$$\theta_m = e_{ML}(T_m) \in \arg \max_{\theta \in \Theta} L(T_m, \theta) = \arg \max_{\theta \in \Theta} \frac{1}{m} \sum_{i=1}^m \log p_\theta(x_i)$$

## Maximum Likelihood estimator

**Example 7** (MLE for exponential families). Consider the parametric family

$$p_{\theta}(x) = h(x) \exp[\langle \phi(x), \theta \rangle - A(\theta)]$$

with sufficient statistic  $\phi(x) \in \mathbb{R}^n$  and natural parameter  $\theta \in \mathbb{R}^n$ . We are given an i.i.d. training set  $T_m$  and want to estimate  $\theta$  by the MLE. The log-likelihood

$$L(T_m, \theta) = \frac{1}{m} \sum_{i=1}^m \log p_{\theta}(x_i) = \frac{1}{m} \sum_{i=1}^m \log \left( h(x_i) \exp[\langle \phi(x_i), \theta \rangle - A(\theta)] \right)$$

is a concave function of  $\theta$  and has only global maxima (see seminar). We compute its derivative and set it to zero.

$$\begin{aligned} \nabla L(T_m, \theta) &= \nabla \frac{1}{m} \sum_{i=1}^m [\log h(x_i) + \langle \phi(x_i), \theta \rangle - A(\theta)] = \frac{1}{m} \sum_{i=1}^m \phi(x_i) - \nabla \log A(\theta) \\ &= \mathbb{E}_{x \sim T_m}[\phi(x)] - \mathbb{E}_{x \sim \theta}[\phi(x)] = 0 \end{aligned}$$

The maximum likelihood estimator picks  $\theta$  so that the expectation of  $\phi(x)$  under the model coincides with its empirical expectation on the training data.

We cannot always compute this estimator in closed form, but we can use e.g. gradient ascent to find the maximum.

## Properties of the ML estimator

Is the Maximum Likelihood estimator unbiased ?

- ◆ On a finite sample when  $m < \infty$ , the ML estimator can be biased.
- ◆ Asymptotically when  $m \rightarrow \infty$ , the ML estimator is unbiased.

**Example 8.** Consider a normal distribution  $p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$  described by mean value  $\mu$  and the variance  $\sigma^2$ . The ML estimate of the mean value  $\mu_m = \frac{1}{m} \sum_{i=1}^m x_i$  is unbiased:

$$\mathbb{E}_{T_m \sim \theta}[\mu_m] = \mathbb{E}_{T_m \sim \theta} \left[ \frac{1}{m} \sum_{i=1}^m x_i \right] = \frac{1}{m} \sum_{i=1}^m \mathbb{E}_{T_m \sim \theta}[x_i] = \mu$$

The ML estimate of the  $\sigma_m^2 = \frac{1}{m} \sum_{i=1}^m (x_i - \mu_m)^2$  is biased for  $m < \infty$ , however, asymptotically unbiased:

$$\mathbb{E}_{T_m \sim \theta}[\sigma_m^2] = \mathbb{E}_{T_m \sim \theta} \left[ \frac{1}{m} \sum_{i=1}^m (x_i - \mu_m)^2 \right] = \dots = \frac{m-1}{m} \sigma^2$$

## Properties of the ML estimator

What conditions ensure MLE consistency, i.e.

$$\mathbb{P}_{T_m \sim \theta^*} (|\theta^* - e_{ML}(T_m)| > \epsilon) \rightarrow 0 \quad \text{for } m \rightarrow \infty$$

For exponential family

$$p_\theta(x) = h(x) \exp(\langle \phi(x), \theta \rangle - A(\theta)),$$

the MLE is statistically consistent for  $\theta^*$  if:

- ◆  $T_m$  is i.i.d. drawn from  $p_{\theta^*}$  in the family,
- ◆ the parameter space  $\Theta = \{\theta \in \mathbb{R}^n \mid A(\theta) < \infty\}$  is open,
- ◆  $A(\theta) = \log \int h(x) \exp(\langle \phi(x), \theta \rangle) dx$  is differentiable, strictly convex, and invertible,
- ◆  $\theta^*$  is in the interior of  $\Theta$ ,
- ◆  $\mathbb{E}_{x \sim \theta^*} \|\phi(x)\| < \infty$

The conditions are satisfied e.g. for: Bernoulli, Poisson, Binomial, Normal, Exponential, and Gamma distributions.

## Properties of the ML estimator

What can we say about the variance of the ML estimator, i.e.  $\mathbb{V}_{T_m \sim \theta^*} [e_{ML}(T_m)]$ ?

The asymptotic variance of the ML estimator is the smallest possible, i.e. MLE is asymptotically efficient!

To make this precise, we need the notion of *Fisher information*

$$I(\theta) = \int \left[ \frac{d}{d\theta} \log p_\theta(x) \right]^2 p_\theta(x) dx = \mathbb{V}_{x \sim \theta} \left[ \frac{d}{d\theta} \log p_\theta(x) \right]$$

Now, we have the following two statements about the variance of estimators

- ◆ The asymptotic distribution of the ML estimator is:

$$e_{ML}(T_m) \sim \mathcal{N}\left(\theta^*, \frac{1}{mI(\theta^*)}\right) \quad \text{for } m \rightarrow \infty$$

- ◆ If  $e$  is an arbitrary unbiased estimator, then its variance can not be smaller (Cram r-Rao bound), i.e.

$$\mathbb{V}_{T_m \sim \theta^*} [e(T_m)] \geq \frac{1}{mI(\theta^*)}$$

## Summary

- ◆ Discriminative vs. generative learning.
- ◆ The exponential family encompasses many common distributions (e.g., Bernouli, Normal, Poisson, Binomial, Gamma).
- ◆ Maximum-Likelihood estimator (MLE).
- ◆ MLE can be biased.
- ◆ MLE is consistent under mild conditions.
- ◆ MLE achieves asymptotically optimal variance.