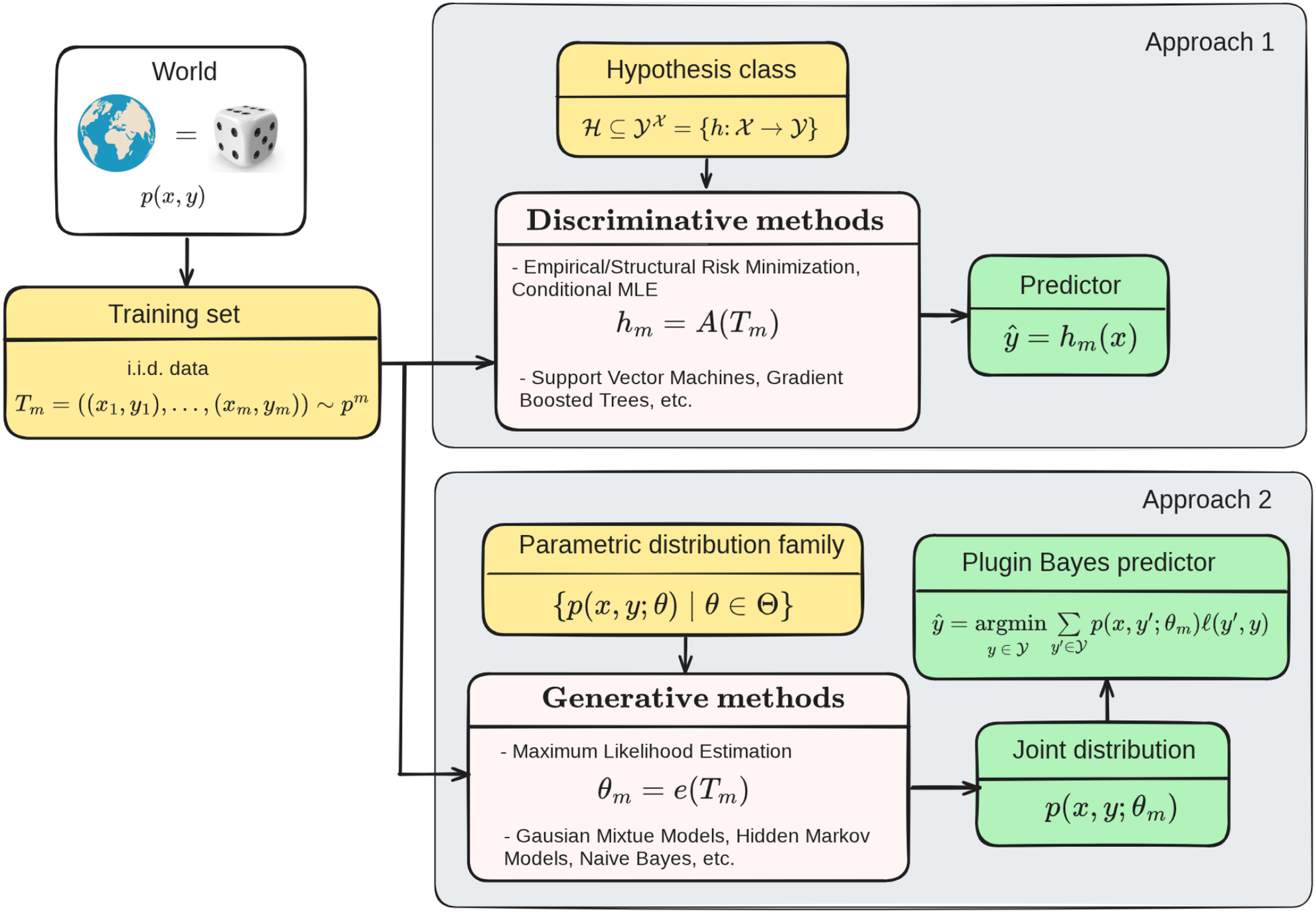
Machine Learning Fundamentals - LS2026
Generative learning

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 $y \in \mathcal{Y}$

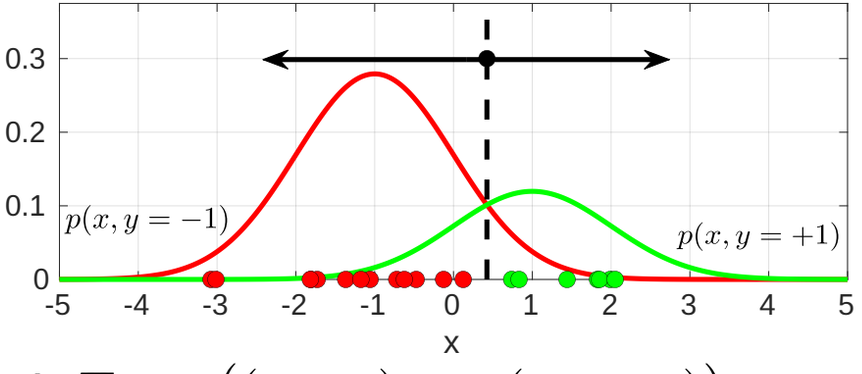
Discriminative learning:

- define a hypothesis space $\mathcal{H}$ of predictors $h\colon \mathcal{X} \to \mathcal{Y}$ and fix a loss $\ell(y, y')$
- given a training set T_m, learn $h_m\colon \mathcal{X} \to \mathcal{Y}$ 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* σ.



The loss is $\ell(y, y') = [\![y' \neq y]\!]$ *and the training set is* $T_m = \big((x_1, y_1), \ldots, (x_m, y_m)\big)$.

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

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

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

Generative learning

Generative learning: Use prior knowledge to restrict the search to a parametric family of distributions $\{p(x,y;\theta) \mid \theta \in \Theta\}$. 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 $p(x,y;\theta_m)$ as the true model. Predict hidden states by the plugin Bayes optimal predictor

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

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

$$\hat{p}(y=1) = \frac{m_+}{m} \quad \hat{\mu}_+ = \frac{1}{m_+} \sum_i x_i [\![y_i = 1]\!] \quad \hat{\mu}_- = \frac{1}{m_-} \sum_i x_i [\![y_i = -1]\!]$$

and

$$\hat{\sigma}^2 = \frac{1}{m} \sum_i \Big(x_i - \mu_+ [\![y_i = 1]\!] - \mu_- [\![y_i = -1]\!] \Big)^2,$$

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

$$\hat{h}(x) = \operatorname{sign}\left(\log \frac{p(x,y=1;\theta_m)}{p(x,y=-1;\theta_m)} \right) = \cdots = \operatorname{sign}(x - \gamma),$$

where γ depends on the estimated $\hat{\mu}_+$, $\hat{\mu}_-$, $\hat{\sigma}$, $\hat{p}(y=1)$ and $\hat{p}(y=-1)$.

Discriminative vs. Generative Learning

Discriminative Models

- Model $\hat{y} = h(x)$ – 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, prediction Neural Networks.

Generative Models

- Model $p(x, y)$ – 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 naturaly handle missing data.
- Examples: Naive Bayes, Gaussian Mixture Models, Generative Adversarial Networks.

Semi-generative Models

- Model $p(y \mid x)$ – learn the class posterior sufficient to design plug-in Bayes predictor.
- Examples: Logistic regression, prediction Neural Networks.

Parametric distribution families

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

Example 2. *The family of Bernoulli distributions for binary r.v.* $x \in \{0,1\}$ *with* $p(x;\beta) = \beta^x(1-\beta)^{1-x}$ *parametrised by a single scalar* $\beta \in (0,1)$.

Example 3. *The family of multivariate normal distributions* $\mathcal{N}(\mu, V)$ *on* $\mathbb{R}^n$

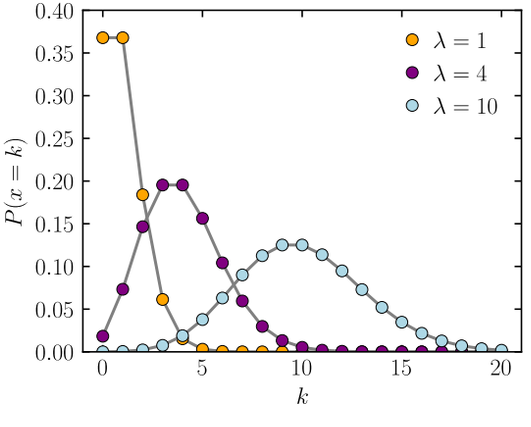
$$p(x;\mu,V) = \frac{1}{(2\pi)^{n/2}|V|^{1/2}} \exp\Big[-\frac{1}{2}(x-\mu)^T V^{-1}(x-\mu)\Big]$$

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

Example 4. *The family of Poisson distributions on* $x \in \mathbb{N}$ *with probability mass*

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

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



Maximum Likelihood Estimator

Given: a parametric family of distributions $\{p(x;\theta) \mid \theta \in \Theta\}$ and an i.i.d. training sample $T_m = (x_i \in \mathcal{X} \mid i = 1,\ldots,m)$ assumed to be i.i.d. generated from $p(x;\theta_*)$ with unknown θ_*.

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

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

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 $p(x_1,\ldots,x_m;\theta) = \prod_{i=1}^{m} p(x_i;\theta)$.

The **Maximum Likelihood estimator** outputs the parameter θ_m that maximizes the (log-) likelihood of the training sample

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

Maximum Likelihood estimator

Example 5 (MLE for univariate Normal distribution)**.** *Consider the parametric family*

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

parmaterized by the mean value $\mu \in \mathbb{R}$ and the variance $\sigma^2 > 0$. We are given an i.i.d. training set $T_m = (x_1, \ldots, x_m)$ and want to estimate (μ, σ^2) by the MLE. The log-likelihood

$$L(T_m, \mu, \sigma^2) = \frac{1}{m}\sum_{i=1}^{m} \log p(x_i;\mu,\sigma^2) = -\frac{m}{2}\log(2\pi) - \frac{m}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(x_i-\mu)^2$$

is a concave function of the parameters. To find the maximum likelihood estimator, we differentiate with respect to the parameters, set the derivatives to zero, and solve:

$$\frac{\partial L(T_m,\mu,\sigma^2)}{\partial \mu} = -\frac{1}{\sigma^2}\sum_{i=1}^{m}(x_i-\mu) = 0 \quad \Rightarrow \quad \boxed{\mu_m = \frac{1}{m}\sum_{i=1}^{m} x_i}$$

$$\frac{\partial L(T_m,\mu,\sigma^2)}{\partial \sigma^2} = -\frac{m}{2}\cdot\frac{1}{\sigma^2} + \frac{1}{\sigma^4}\sum_{i=1}^{m}(x_i-\mu)^2 = 0 \quad \Rightarrow \quad \boxed{\sigma_m^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i-\mu_m)^2}$$

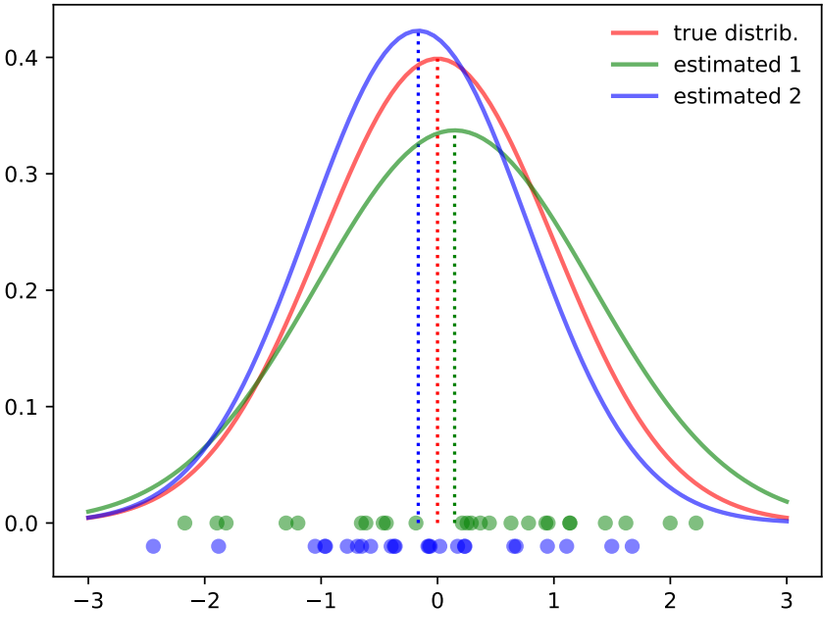
Properties of Parameter Estimators

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

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

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.* $\mu_m = e(T_m) = \frac{1}{m}\sum_{i=1}^{m} x_i$.

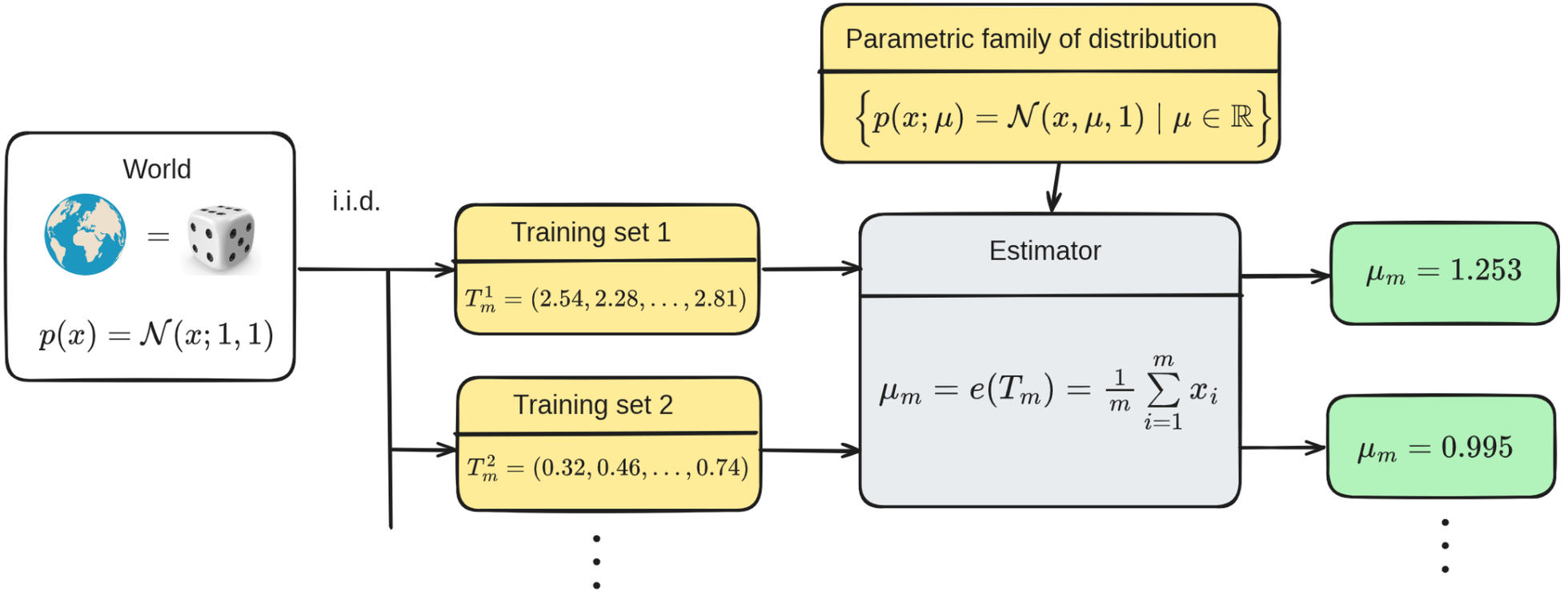


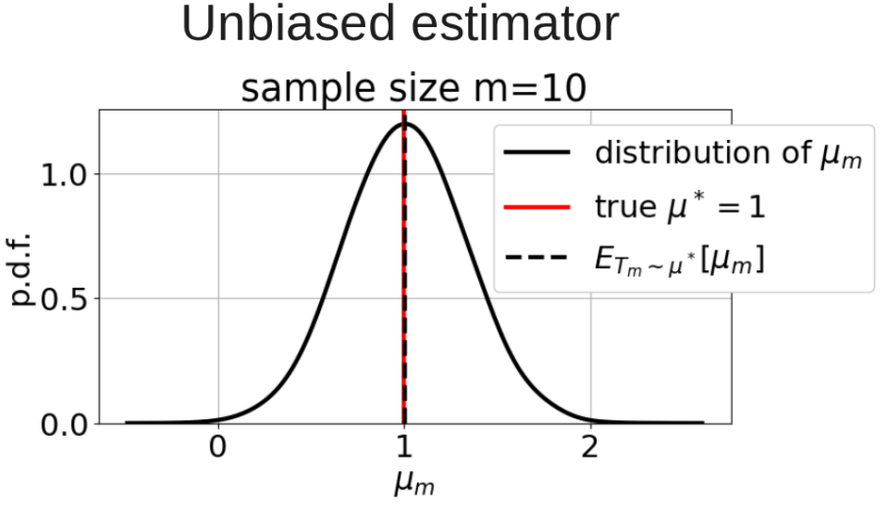
Desired properties of an estimator:

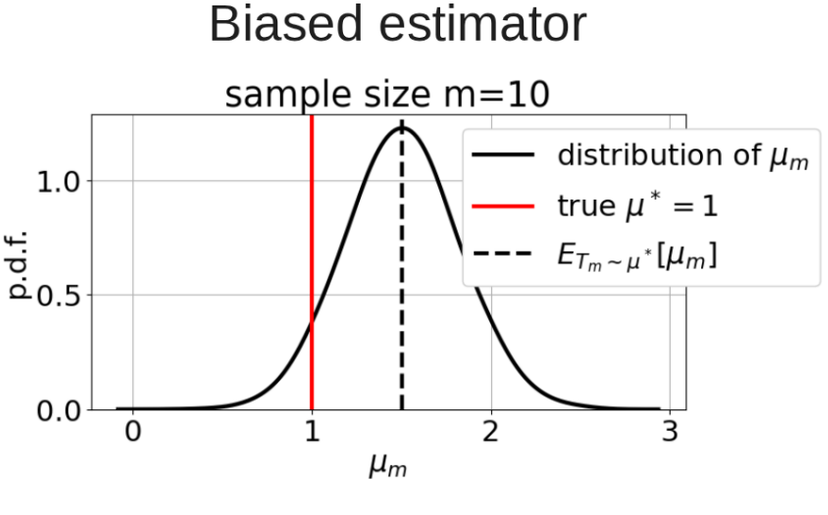
- the estimator is unbiased i.e. $\mathbb{E}_{T_m \sim \theta_*}\big[e(T_m)\big] = \theta_*$
- the estimator has small variance $\mathbb{V}_{T_m \sim \theta_*}\big[e(T_m)\big]$
- the estimator is consistent i.e. $\mathbb{P}_{T_m \sim \theta_*}\Big(|e(T_m) - \theta_*| > \epsilon\Big) \to 0$ for $m \to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

Properties of Parameter Estimators

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







Properties of Maximum Likelihood Estimator

Is the Maximum Likelihood estimator unbiased ?

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

Example 7. *Consider a normal distribution* $p(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$ *described by mean value* μ *and the variance* σ^2*. The ML estimate of the mean value* $\mu_m = \frac{1}{m}\sum_{i=1}^{m} x_i$ *is unbiased:*

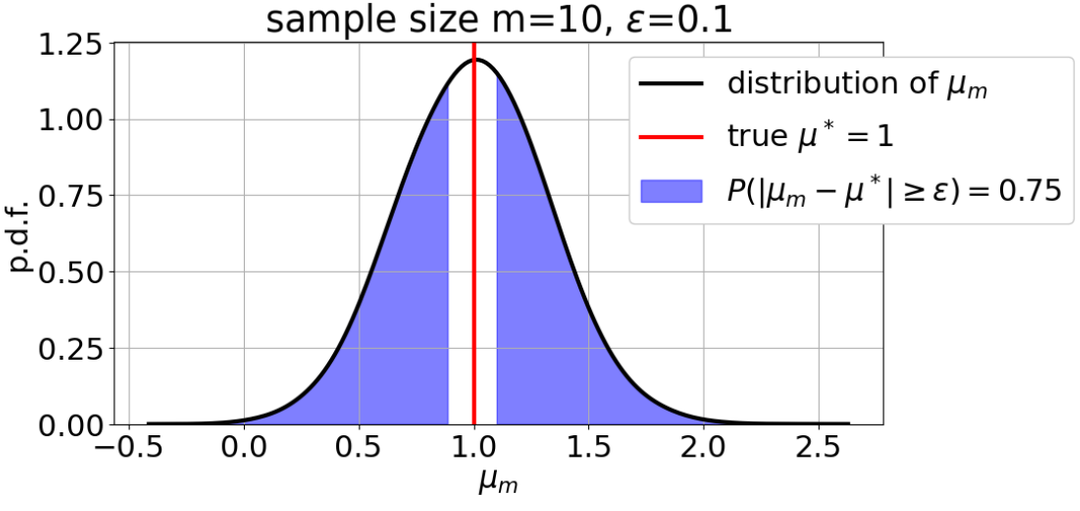
$$\mathbb{E}_{T_m \sim \theta_*}[\mu_m] = \mathbb{E}_{T_m \sim \theta_*}\left[\frac{1}{m}\sum_{i=1}^{m} x_i\right] = \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{T_m \sim \theta_*}[x_i] = \mu$$

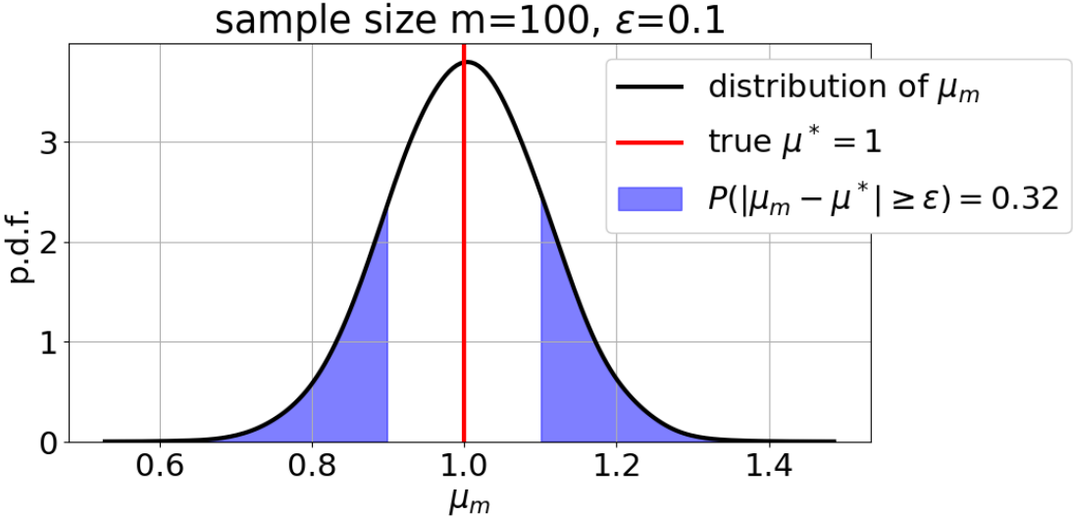
The ML estimate of the $\sigma_m^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_m)^2$ *is biased for* $m < \infty$*, however, asymprotically unbiased:*

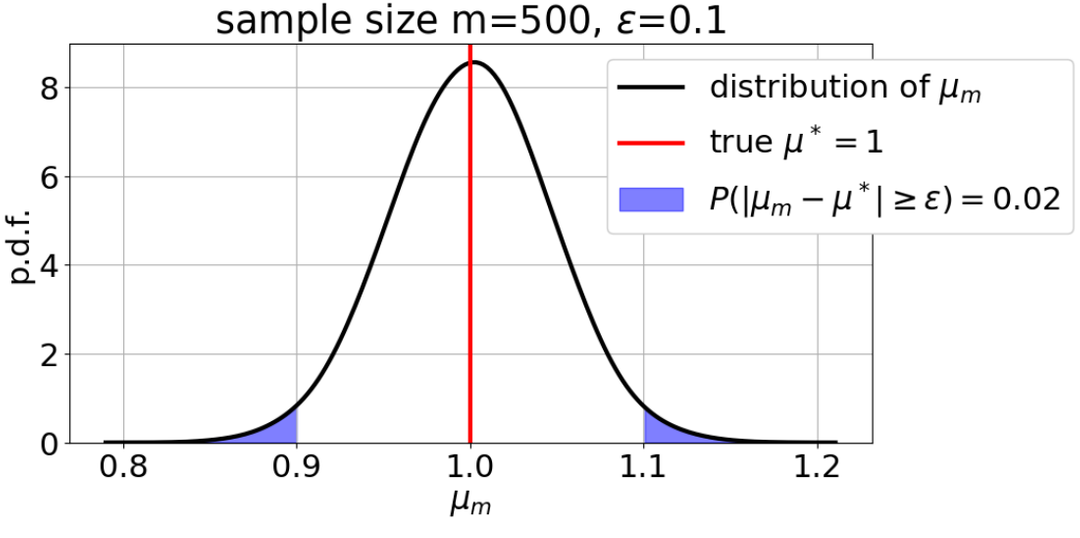
$$\mathbb{E}_{T_m \sim \theta_*}[\sigma_m^2] = \mathbb{E}_{T_m \sim \theta_*}\left[\frac{1}{m}\sum_{i=1}(x_i - \mu_m)^2\right] = \cdots = \frac{m-1}{m}\sigma^2$$

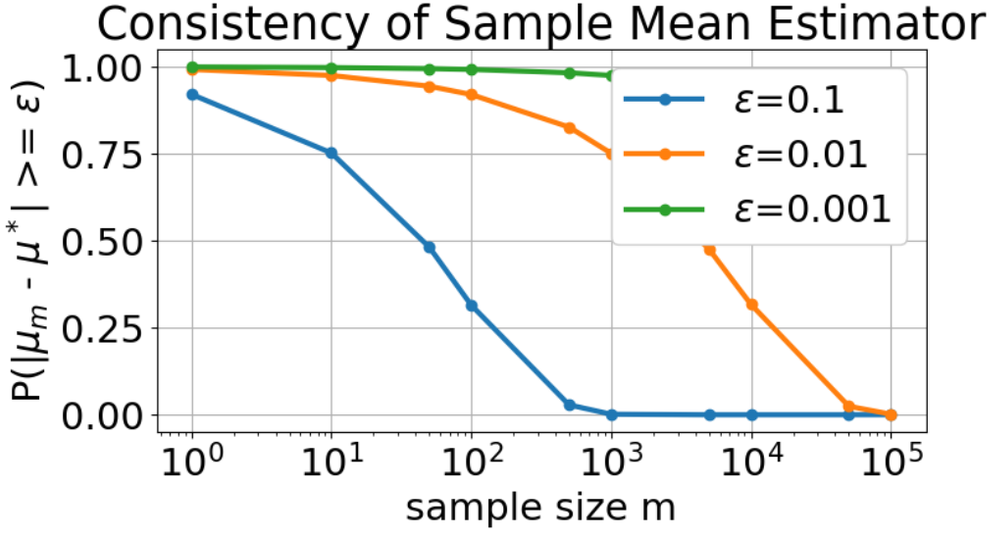
Properties of Parameter Estimators

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









Properties of Maximum Likelihood Estimators

What conditions ensure consistency of the ML estimator, i.e.

$$\lim_{m\to\infty} \mathbb{P}_{T_m\sim\theta_*}\big(|\theta_* - e_{ML}(T_m)| > \epsilon\big) = 0$$

The maximum likelihood estimator is statistically consistent provided that:

- **Correct specification.** The true data-generating distribution $p(x)$ belong to the family:

$$p(x) = p(x;\theta_*) \quad \text{for some} \quad \theta_* \in \Theta$$

- **Identifiability.** Different parameter values correspond to different distribution:

$$\theta \neq \theta' \Rightarrow p(x;\theta) \neq p(x;\theta')$$

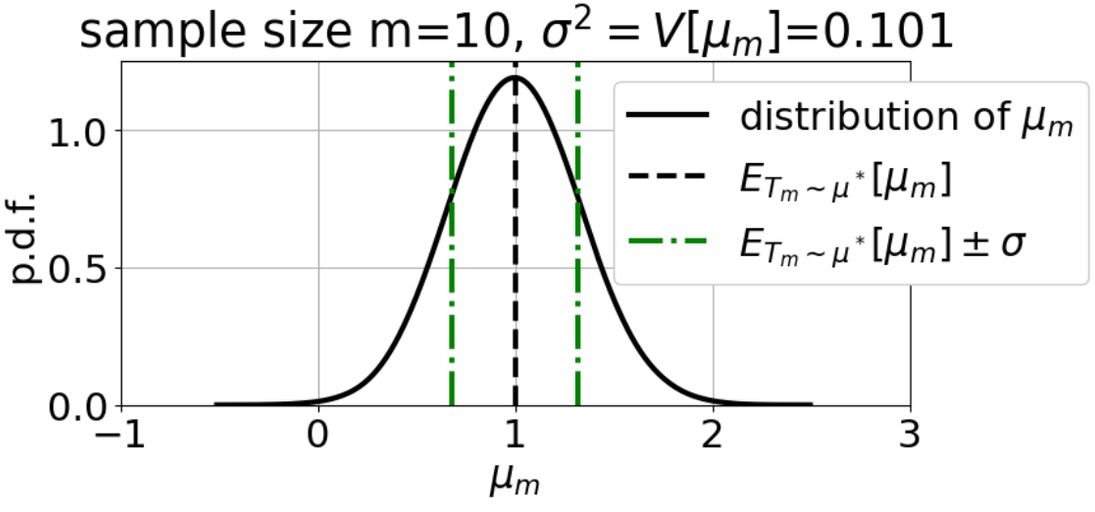
- **Regularity conditions.** The parameter space Θ and the family $p(x;\theta)$ are "well behaved" (e.g. smoothness, existence of MLE, compactness of Θ).

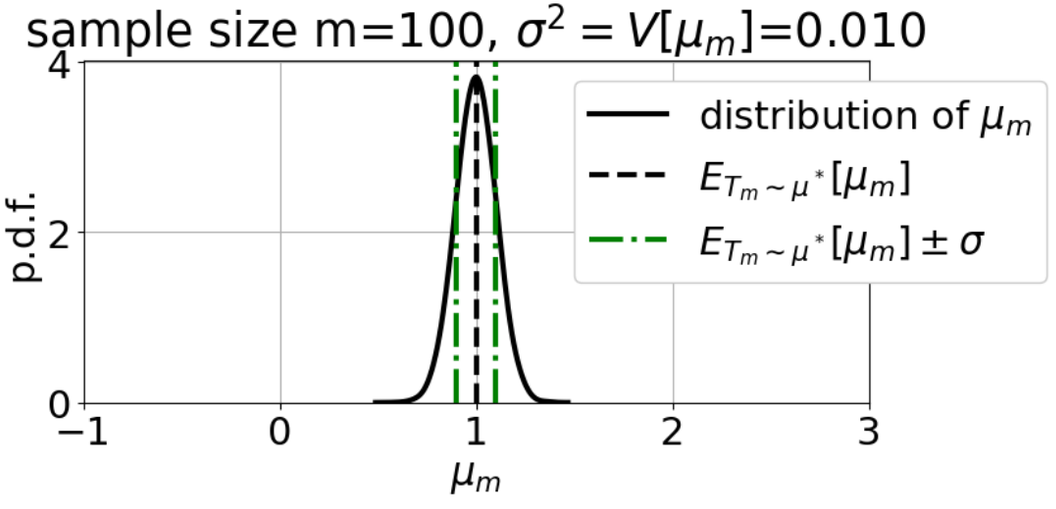
Examples of distributions for which the MLE is statisticlly consistent include Normal, Bernoulli, Poisson, Binomial, Exponential, Gamma, etc.

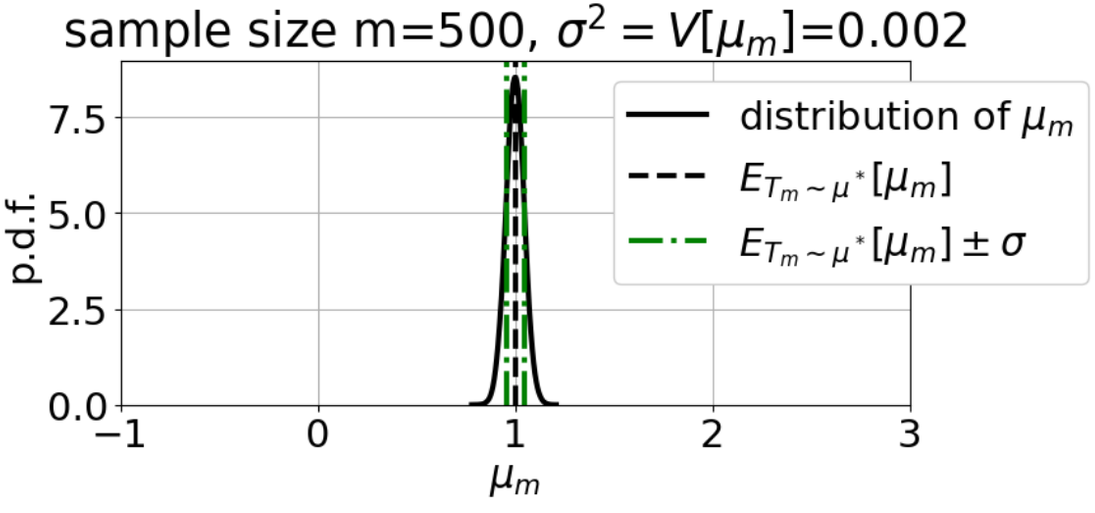
Properties of Parameter Estimatior

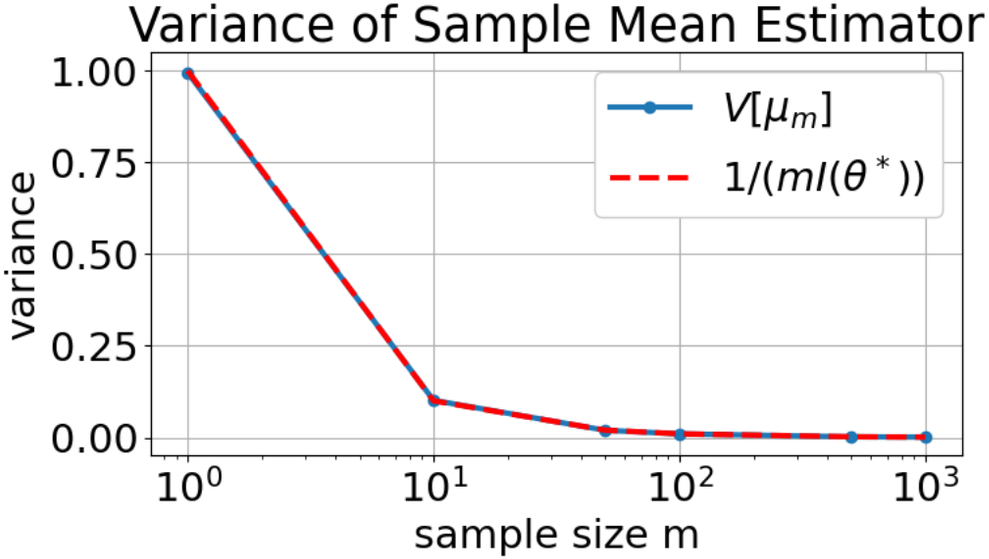
An **efficient estimator** has the smallest possible variance among all unbiased estimators, i.e. it attains the Cramer-Rao lower bound:

$$\mathbb{V}_{T_m \sim \theta_*}\left[e(T_m)\right] \geqslant \frac{1}{m\, I(\theta_*)}$$









Properties of the ML estimator

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

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

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

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

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

- The asymptotic distribution of the ML estimator is:

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

- If e is an arbitrary unbiased estimator, then its variance can not be smaller (Crame-Rao bound), i.e.

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

Summary

- Discriminative vs. generative learning.
- Parametric distribution families (e.g., Bernouli, Normal, Poisson, Binomial, Gamma).
- Maximum-Likelihood estimator (MLE).
- MLE can be biased.
- MLE is statistically consistent under regularity conditions (e.g. the model has to well specified and identifiable).
- MLE achieves asymptotically optimal variance.