UNIVERSITY OF AMSTERDAM
Informatics Institute

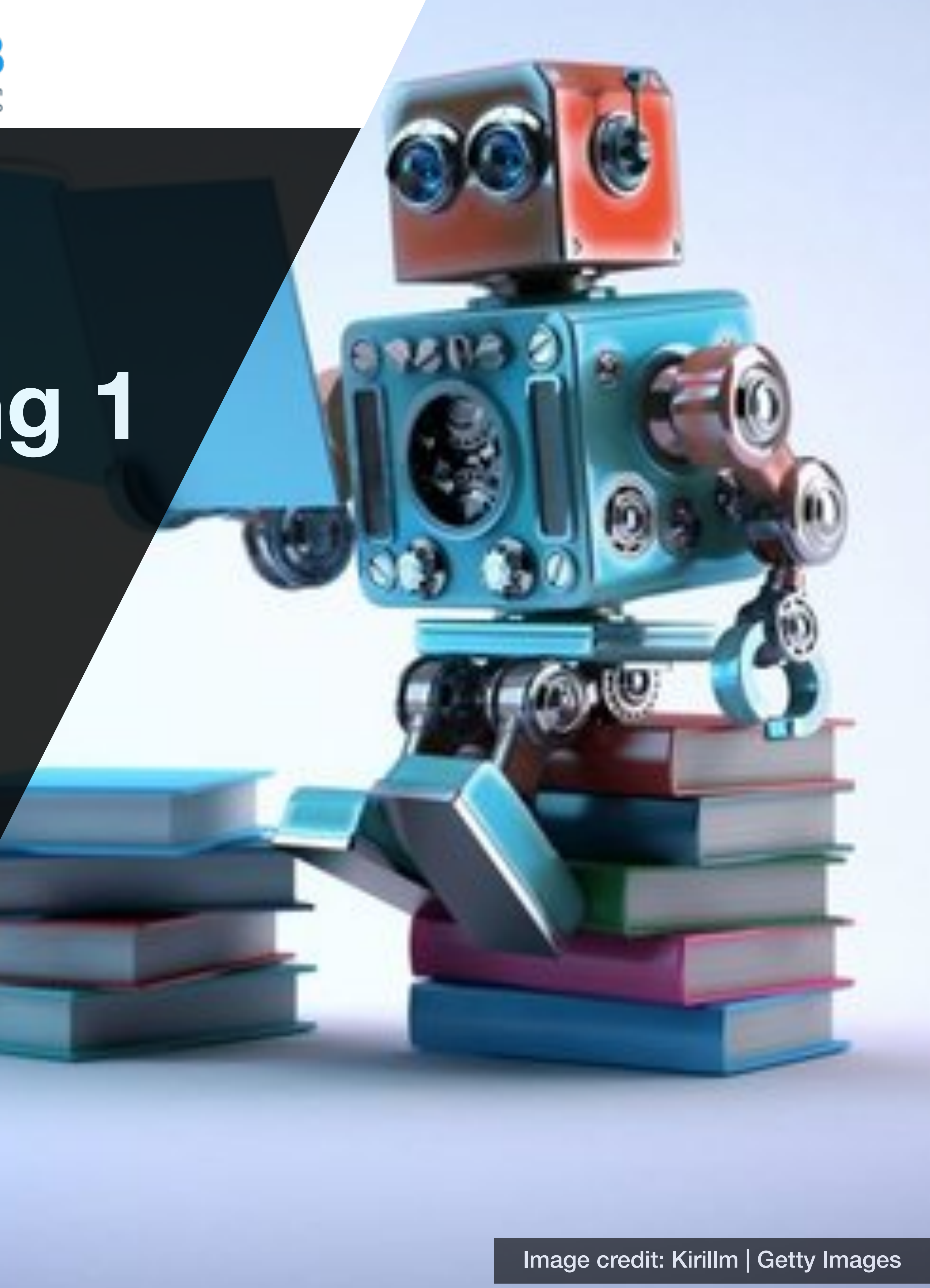
AMLAB Amsterdam Machine Learning Lab

Machine Learning 1

Lecture 2.3 - Maximum Likelihood

Erik Bekkers

(Bishop 1.2.3 - 1.2.5)



Slide credits: Patrick Forré and Rianne van den Berg

Maximum Likelihood Principle

- Dataset $D = (x_1, x_2, \ldots, x_N)$ of N independent observations.
- Likelihood of the dataset: $p(D|w)$
- Maximum likelihood principle: the most likely "explanation" of D is given by $\boldsymbol{w}_{\text{ML}}$ which maximizes the likelihood function

$$\mathbf{w}_{\text{ML}} = \underset{w}{\operatorname{argmax}}\ p(D|w)$$

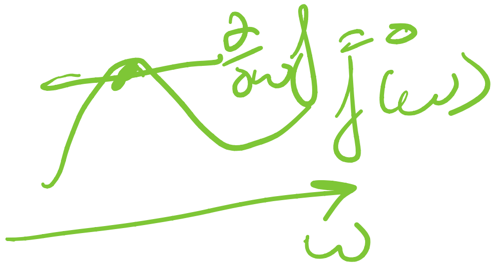
- i.i.d. assumption: each $x_i \in D$ is independently distributed according to the same identically distribution, conditioned on $\mathbf{w}$.

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

- If i.i.d., joint distribution

$$p(D|\mathbf{w}) = p(x_1, x_2, \ldots, x_N|\mathbf{w}) = \prod_{i=1}^{N} p(x_i|w)$$

Maximum Likelihood Estimation



- Maximum likelihood estimation $\boldsymbol{w}_{\mathrm{ML}}$

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

$$\Rightarrow \boxed{\frac{\partial}{\partial w} p(D|w) = 0}$$

numerical underflow/overflow

$$\frac{\partial}{\partial w} \log p(D|w) = 0$$

$$\cancel{\frac{1}{p(D|w)}} \boxed{\frac{\partial}{\partial w} p(D|w) = 0}$$

- How do we maximize?

- Maximize log-likelihood instead:

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

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

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

ML Estimator for Gaussian Distributions (I)

- i.i.d. Gaussian distributed real variables $D = (x_1, x_2, ..., x_N)$

$$p(x|\mathbf{w}) = \mathcal{N}(x|\mu, \sigma^2) \Rightarrow p(D|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \prod_{i=1}^{N} \exp\left[-\frac{1}{2\sigma^2}(x_i - \mu)^2\right]$$

$$p(x_1, x_2, ..., x_N | \mu, \sigma^2) = p(x_1|\mu, \sigma^2)\, p(x_2|\mu, \sigma^2) \dots$$

- Log likelihood

$\log x^a = a \log x$

$$\log p(D|\mu, \sigma^2) = \log (2\pi\sigma^2)^{-N/2} + \sum_{i=1}^{N} \log \exp\left[-\frac{1}{2\sigma^2}(x_i - \mu)^2\right]$$

$$= -\frac{N}{2} \log(2\pi\sigma^2) + \sum_{i=1}^{N} -\frac{1}{2\sigma^2}(x_i - \mu)^2$$

- Estimate model parameters: $\mu_{ML}, \sigma^2_{ML} = \underset{\mu,\sigma^2}{\operatorname{argmax}} \log p(D\,|\,\mu, \sigma^2)$

$$\frac{\partial}{\partial \mu} \log p(D|\mu, \sigma^2) = 0$$

solve for $\mu \rightarrow \mu_{ML}$

ML Estimator for Gaussian Distributions (II)

- log likelihood:

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

$$\frac{\partial}{\partial\mu}(x_i-\mu)^2 = 2(x_i-\mu)$$

- Maximum Likelihood solution for μ

$$\frac{\partial}{\partial\mu}\log p(D|\mu,\sigma^2) = \frac{1}{\cancel{2}\sigma^2}\sum_{i=1}^{N}\cancel{2}(x_i-\mu) = 0$$

$$\Rightarrow \sum_{i=1}^{N}(x_i-\mu) = 0$$

$$\Rightarrow \sum_{i=1}^{N}\mu = \sum_{i=1}^{N}x_i$$

$$\Rightarrow N\cdot\mu = \sum_{i=1}^{N}x_i \Longrightarrow \boxed{\mu = \frac{1}{N}\sum_{i=1}^{N}x_i}$$

sample mean

ML Estimator for Gaussian Distributions (II)

- log likelihood:

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

$$\frac{\partial}{\partial \sigma^2} \cdot \frac{1}{2\sigma^2} = -\frac{1}{2\sigma^4}$$

- Maximum Likelihood solution for σ^2

$$\frac{\partial}{\partial\sigma^2}\log p(D|\mu,\sigma^2) = -\frac{N}{2}\frac{1}{\cancel{2\pi}\sigma^2}\cdot\cancel{2\pi} + \frac{1}{2\sigma^4}\sum_{i=1}^{N}(x_i-\mu)^2 = 0$$

$$\cdot\, 2\sigma^4$$

$$\Rightarrow -N\sigma^2 + \sum_{i=1}^{N}(x_i-\mu)^2 = 0$$

$$-\frac{1}{N}$$

$$\Rightarrow \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$$ sample variance

ML Estimator for Gaussian Distributions (IV)

- How well do the ML estimators represent the true parameters? $$p(D|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \prod_{i=1}^{N} \exp\left[-\frac{1}{2\sigma^2}(x_i-\mu)^2\right]$$

- If I draw multiple datasets, what is the expected value of μ_{ML}?

$D_1 = \{x_1, \ldots, x_n\}$, $D_2 = \{x_1, \ldots, x_n\}$, ...

- ML estimate of the mean:

$$\mathbb{E}_{D\sim p(D|\mu,\sigma^2)}[\mu_{\mathrm{ML}}] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{D\sim p(D|\mu,\sigma^2)}[x_i]$$

$$= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{x_i\sim p(x_i|\mu,\sigma^2)}[x_i] = \frac{1}{N}\sum_{i=1}^{N} \mu = \mu$$

- Bias of estimator:

$$\mathbb{E}[\mu_{\mathrm{ML}}] - \mu = 0$$

ML Estimator for Gaussian Distributions (V)

- ML estimate of the variance:

$$\mathbb{E}_{D \sim p(D|\mu,\sigma^2)}[\sigma^2_{ML}] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_{n=1}^{N} x_n\right)^2\right]$$

ML Estimator for Gaussian Distributions (V)

- ML estimate of the variance:

$$\mathbb{E}_{D\sim p(D|\mu,\sigma^2)}[\sigma^2_{ML}] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_{n=1}^{N}x_n\right)^2\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[(xi - \frac{1}{N}\sum_{n=1}^{N}x_n)^2\right]$$

ML Estimator for Gaussian Distributions (V)

- ML estimate of the variance:

$$\mathbb{E}_{D\sim p(D|\mu,\sigma^2)}[\sigma^2_{ML}] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_{n=1}^{N}x_n\right)^2\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[(xi - \frac{1}{N}\sum_{n=1}^{N}x_n)^2\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[x_i^2 - \frac{2x_i}{N}\sum_{n=1}^{N}x_n + \frac{1}{N^2}\sum_{m=1}^{N}\sum_{n=1}^{N}x_m x_n\right]$$

ML Estimator for Gaussian Distributions (V)

- ML estimate of the variance:

$$\mathbb{E}_{D\sim p(D|\mu,\sigma^2)}[\sigma^2_{ML}] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \frac{1}{N}\sum_{n=1}^{N}x_n\right)^2\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[(xi - \frac{1}{N}\sum_{n=1}^{N}x_n)^2\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[x_i^2 - \frac{2x_i}{N}\sum_{n=1}^{N}x_n + \frac{1}{N^2}\sum_{m=1}^{N}\sum_{n=1}^{N}x_m x_n\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left\{ \mathbb{E}\left[x_i^2\right] - \frac{2}{N}\sum_{n=1}^{N}\mathbb{E}\left[x_i x_n\right] + \frac{1}{N^2}\sum_{m=1}^{N}\sum_{n=1}^{N}\mathbb{E}\left[x_m x_n\right] \right\}$$

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

$$= \cancel{\mu^2} + \sigma^2 - 2\cancel{\mu^2} - \frac{2}{N}\sigma^2 + \cancel{\mu^2} + \frac{1}{N}\sigma^2 = \sigma^2 - \frac{1}{\sigma^2} = \frac{N-1}{N}\sigma^2$$

$$\mathbb{E}[x_i x_j] = \begin{cases} \mu^2 + \sigma^2 & \text{if } \underline{i = j} \\ \mu^2 & \text{if } i \neq j \end{cases}$$

$$\mathrm{Cov}[x_i, x_i] = \mathbb{E}[x_i^2] - \mathbb{E}[x_i]^2 = \sigma^2$$

$$\mathrm{Cov}[x_i, x_j] = \mathbb{E}[x_i x_j] - \mathbb{E}[x_i]\mathbb{E}[x_j] = 0$$

ML Estimator for Gaussian Distributions (VI)

- For data generated from

$$p(D|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \prod_{i=1}^{N} \exp\left[-\frac{1}{2\sigma^2}(x_i - \mu)^2\right]$$

- ML gives biased estimator

$$\mathbb{E}[\sigma^2_{\mathrm{ML}}] = \underline{\underline{\frac{N-1}{N}}} \sigma^2$$

- Unbiased variance estimator:

$$\tilde{\sigma}^2 = \frac{N}{N-1} \sigma^2_{ML} = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu)^2$$

$$\mathbb{E}[\tilde{\sigma}^2] = \sigma^2$$

Biased Maximum Likelihood Estimator

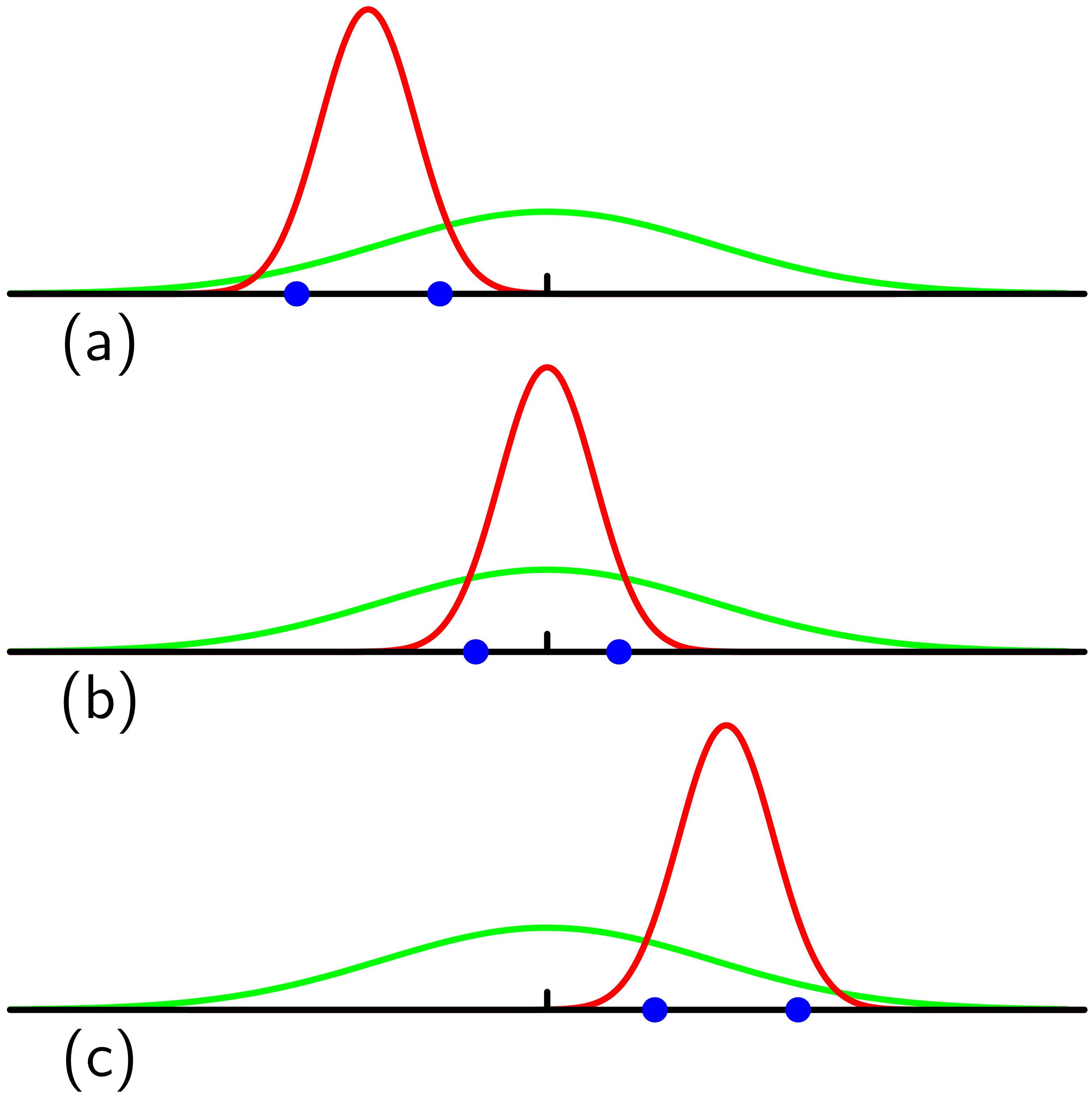


Figure: Bias in ML estimator for variance (Bishop 1.15)