

December 1, 2023



Inter-University Research Institute Corporation /
Research Organization of Information and Systems

National Institute of Informatics

Bias-Variance Tradeoff

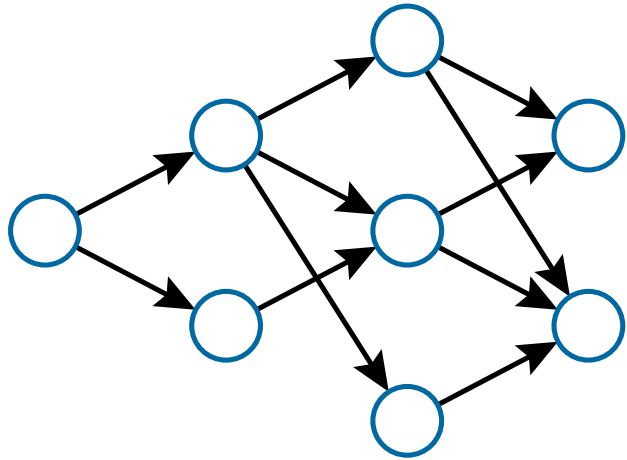
Data Mining 06 (データマイニング)

Mahito Sugiyama (杉山磨人)

Today's Outline

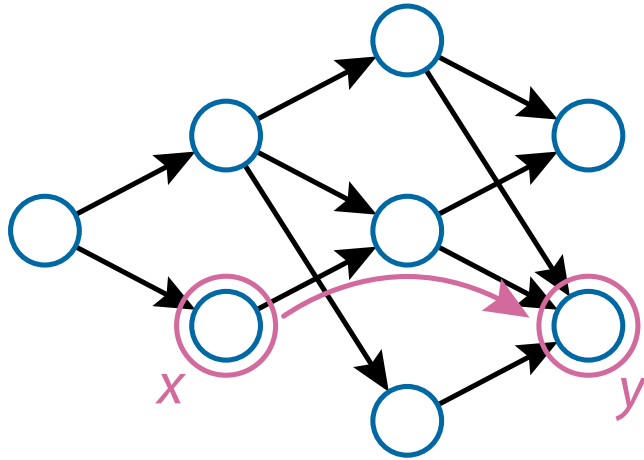
- Log-linear models on posets
 - A generalized formulation of Boltzmann machines
- Bias-variance tradeoff
- Fisher information & Cramér-Rao inequality

Partially Ordered Set (Poset)



- Partially ordered set (**poset**) (S, \leq)
 - $x \leq x$ (reflexivity)
 - $x \leq y, y \leq x \Rightarrow x = y$ (antisymmetry)
 - $x \leq y, y \leq z \Rightarrow x \leq z$ (transitivity)
 - We assume that S is finite and includes the least element (bottom) $\perp \in S$

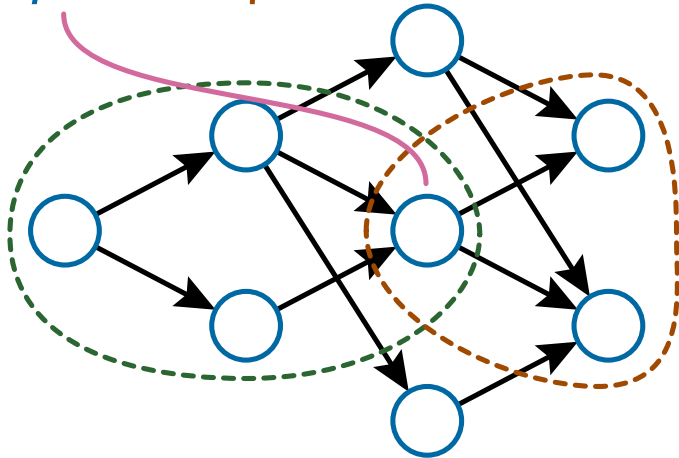
Partially Ordered Set



- Partially ordered set (**poset**) (S, \leq)
 - (i) $x \leq x$ (reflexivity)
 - (ii) $x \leq y, y \leq x \Rightarrow x = y$ (antisymmetry)
 - (iii) $x \leq y, y \leq z \Rightarrow x \leq z$ (transitivity)
 - We assume that S is finite and includes the least element (bottom) $\perp \in S$
- Equivalent to a DAG
 - Each $x \in S$ is a node
 - $x \leq y \iff y$ is reachable from x

Log-Linear Model on Poset

Each $x \in S$ has a triple:
 $(p(x), \theta_x, \eta_x)$



- A probability distribution $p : S \rightarrow (0, 1)$ s.t. $\sum_{x \in S} p(x) = 1$
- We introduce $(\theta_s)_{s \in S}$ and $(\eta_s)_{s \in S}$ as
$$\log p(x) = \sum_{s \leq x} \theta_s$$
$$\eta_x = \sum_{s \geq x} p(s)$$
 - Parameter set $B \subseteq S$
 - $\theta_s = 0$ if $s \notin B$

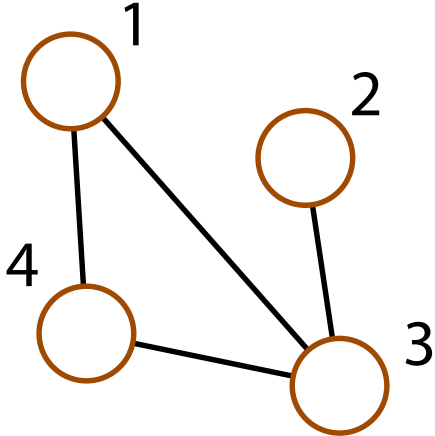
Log-Linear Model on Powerset

- Probability distribution over the power set 2^V with $V = \{1, 2, \dots, n\}$
 - $x \leq y \iff x \subseteq y, S = 2^V$
- Probability $p(x)$ for each $x \in 2^V$ is given as $\log p(x) = \sum_{s \subseteq x} \theta_s$
 - Parameter set $B \subseteq 2^V$, $\theta_s = 0$ if $s \notin B$
- **Maximum Likelihood Estimation (MLE):**
Find $(\theta_s)_{s \in B}$ from a dataset $D \subseteq 2^V$ s.t. $\eta_s = \hat{\eta}_s$

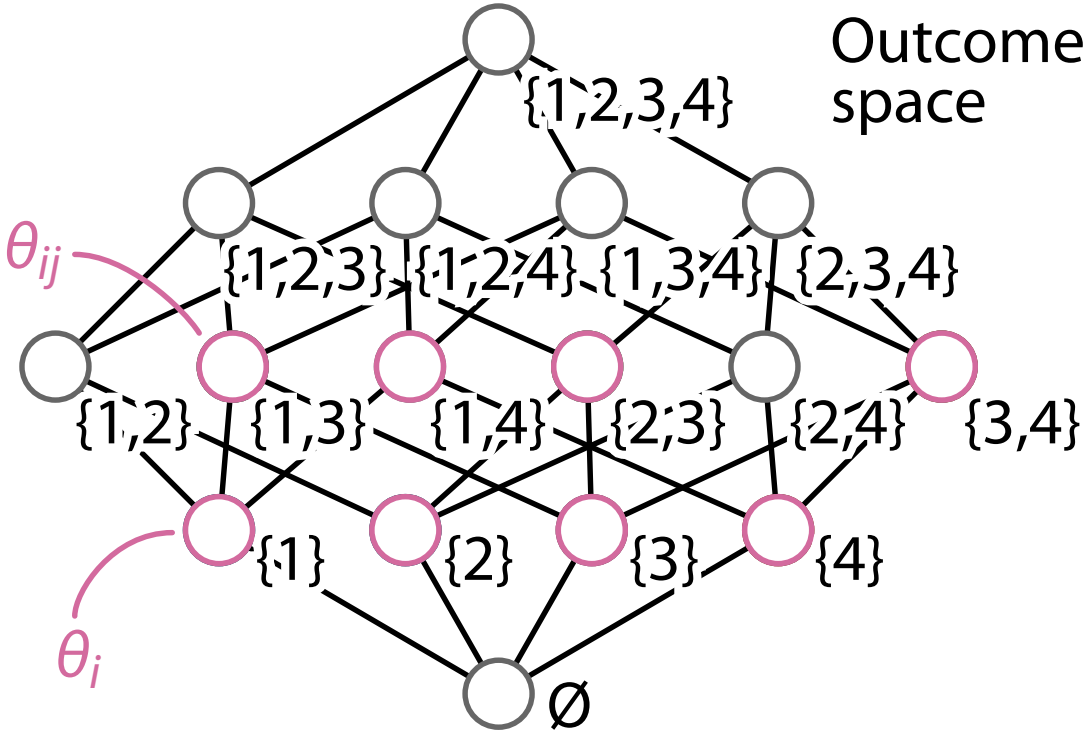
$$\eta_s = \sum_{x \supseteq s} p(x), \quad \hat{\eta}_s = \frac{1}{|D|} \sum_{x \in D} \mathbf{1}[x \supseteq s] = \frac{|\{x \in D \mid x \supseteq s\}|}{|D|}$$

Boltzmann Machines

Boltzmann Machine (BM)

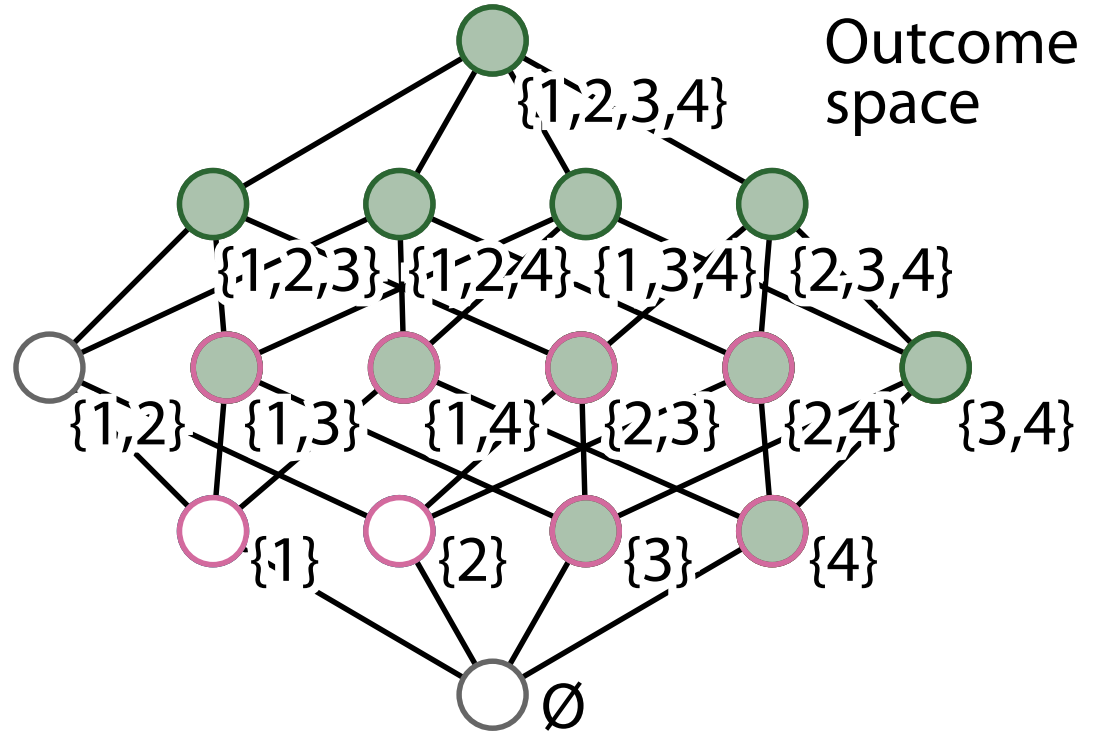
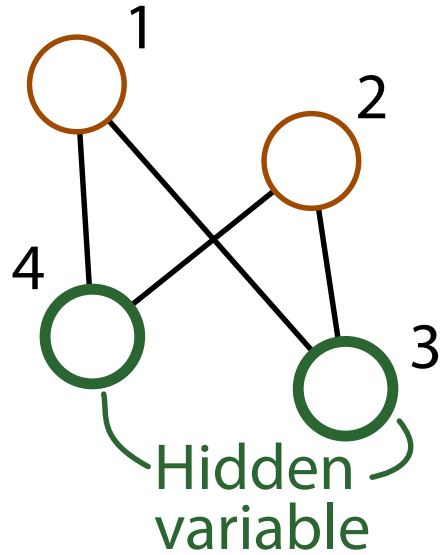


Outcome space

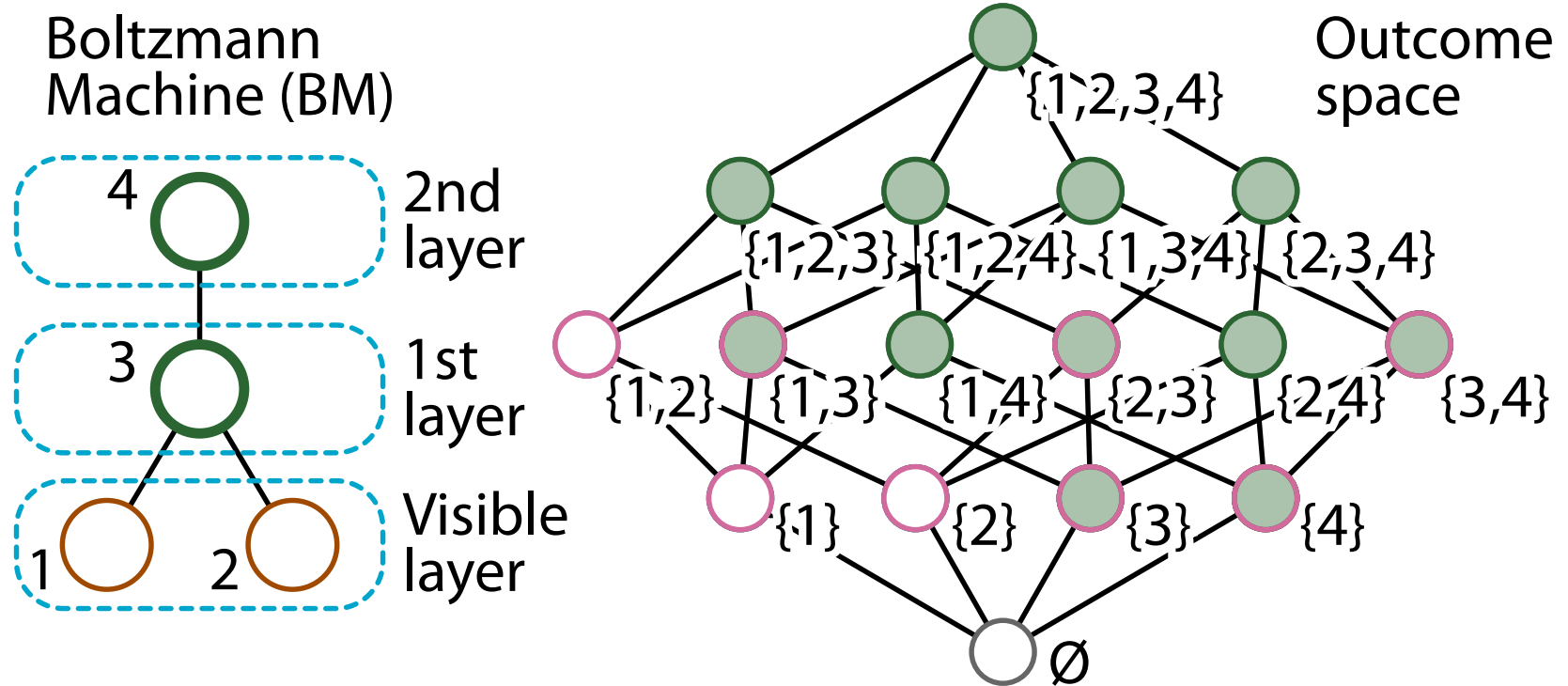


Restricted Boltzmann Machines (RBMs)

Boltzmann Machine (BM)



Deep Boltzmann Machines (DBMs)



Exponential Family

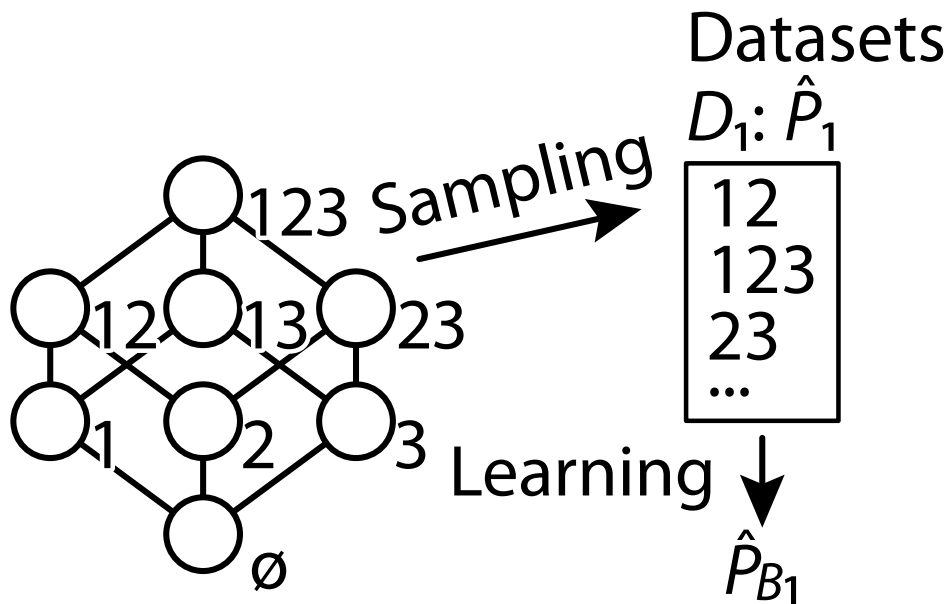
- The general form of the exponential family:

$$p(x; \theta) = \exp \left(\sum_s \theta_s k_s(x) + r(x) - \psi(\theta) \right)$$

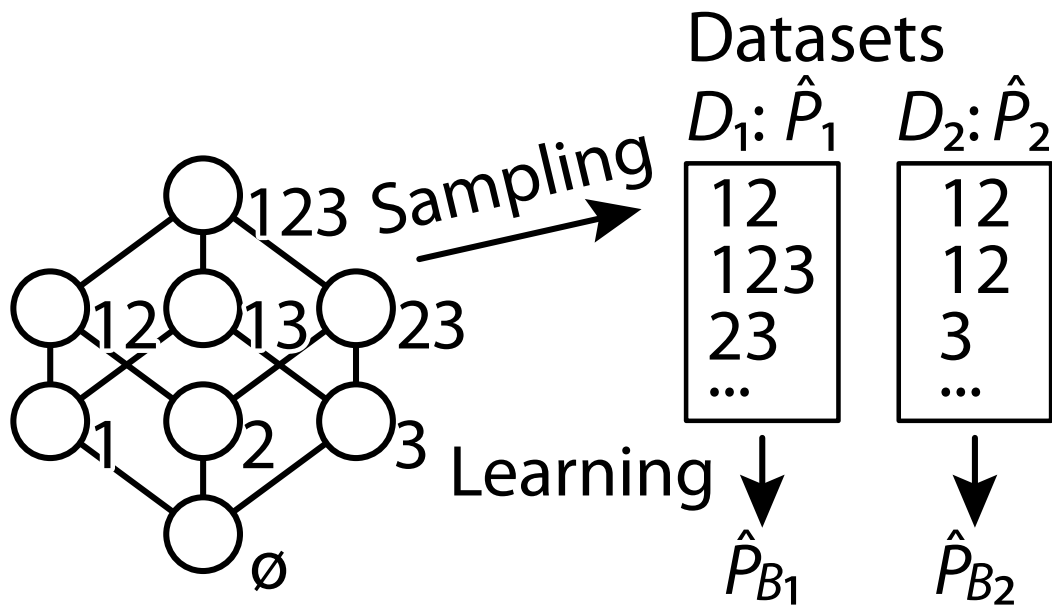
- In the log-linear model on posets,

$$k_s(x) = \begin{cases} 1 & \text{if } s \leq x, \\ 0 & \text{otherwise,} \end{cases} \quad r(x) = 0, \quad \psi(\theta) = -\theta_{\perp}$$

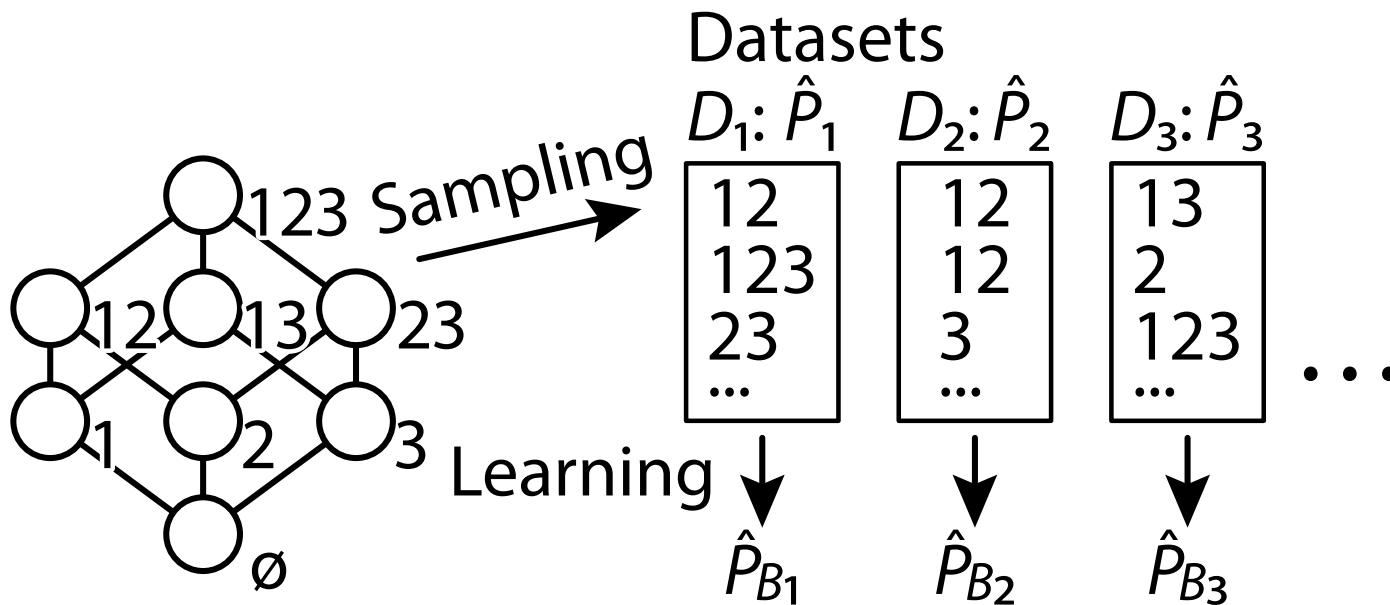
Learning from Data



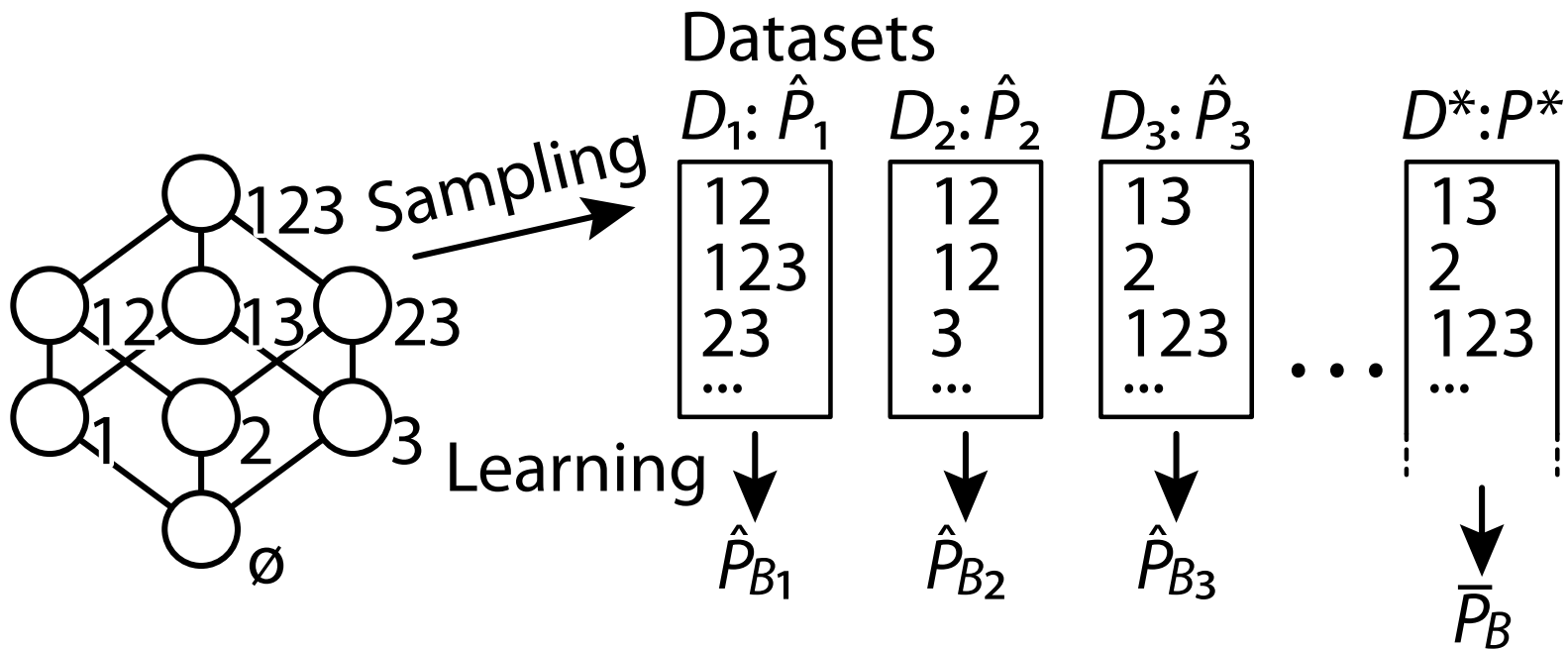
Learning from Data



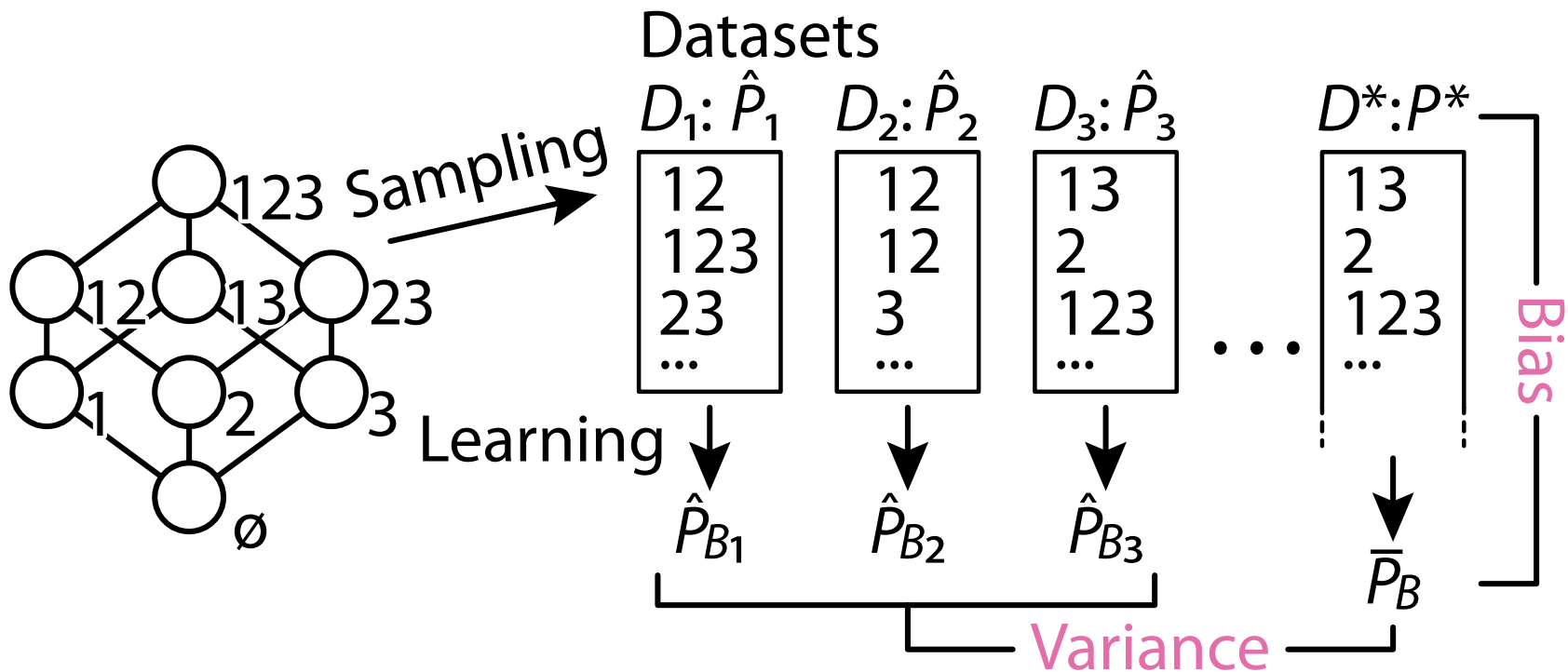
Learning from Data



Learning from Data



Learning from Data



Bias-Variance Tradeoff

- Bias = $D_{\text{KL}}(P^*, \bar{P}_B)$
- Variance = $\mathbf{E}[D_{\text{KL}}(\bar{P}_B, \hat{P}_B)]$
- If we include more parameters in B :
 - Bias will **decrease**
 - Variance will **increase**
- Two extreme cases:
 - If $B = 2^V$, then $\hat{P}_B = \hat{P}$, thus bias = 0 but variance will be large
 - If $B = \emptyset$, \hat{P}_B is always the uniform distribution U , thus bias = $D_{\text{KL}}(U, P^*)$ and variance = 0

Bias-Variance Decomposition

- Decomposition of **MSE** (Mean Squared Error)

$$\begin{aligned}\mathbf{E} \left[(\hat{\theta} - \theta^*)^2 \right] &= (\bar{\theta} - \theta^*)^2 + \mathbf{E} \left[(\hat{\theta} - \bar{\theta})^2 \right] \\ &= \text{bias}^2(\hat{\theta}) + \text{var} \left[\hat{\theta} \right]\end{aligned}$$

$$\text{MSE} = \text{bias}^2 + \text{variance}$$

- θ^* : the true parameter
- $\hat{\theta}$: the estimate
- $\bar{\theta}$: the expected value of the estimate, $\bar{\theta} = \mathbf{E}[\hat{\theta}]$
(the estimate obtained from infinitely many data points)
- The expectation \mathbf{E} is about the true distribution $p(D; \theta^*)$

Example: Gaussian Mean Estimation

- Estimate the **mean** from N data points x_1, x_2, \dots, x_N sampled from a Gaussian distribution $N(\theta^* = 1, \sigma^2)$
- Strategy 1: MLE

$$\text{bias} = \mathbf{E}[\hat{\theta}] - \theta^* = \mathbf{E}\left[\frac{1}{N} \sum_{i=1}^N x_i\right] - \theta^* = N\theta^*/N - \theta^* = 0$$

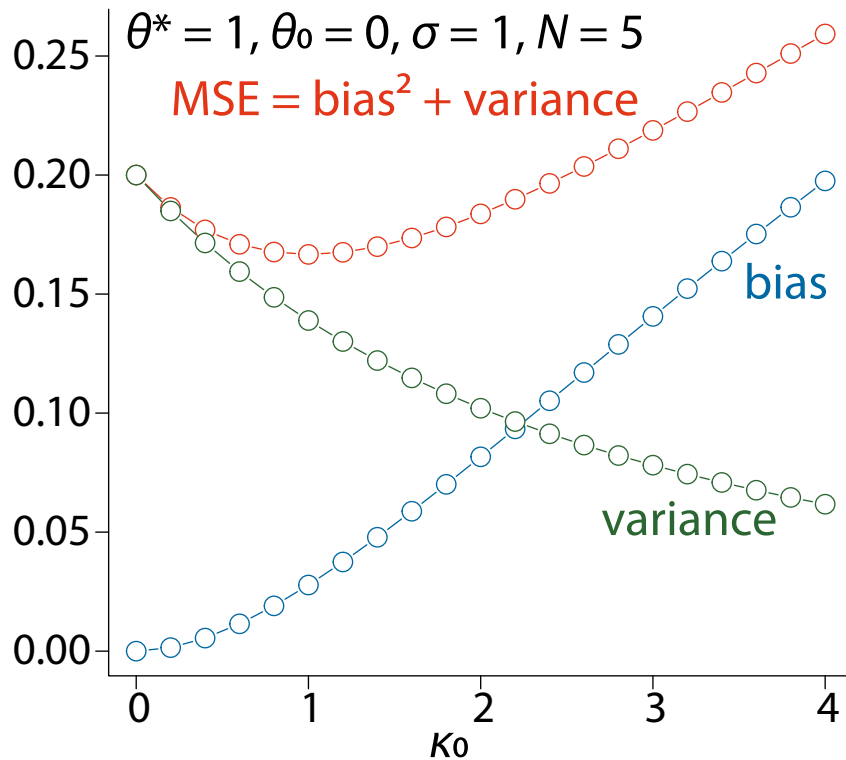
$$\text{variance} = \sigma^2/N$$

- Strategy 2: MAP estimate with a prior $N(\theta_0, \sigma^2/\kappa_0)$

$$\text{bias} = \left(w\mathbf{E}[\hat{\theta}] + (1-w)\theta_0\right) - \theta^* = (1-w)(\theta_0 - \theta^*), \quad w = N/(N + \kappa_0)$$

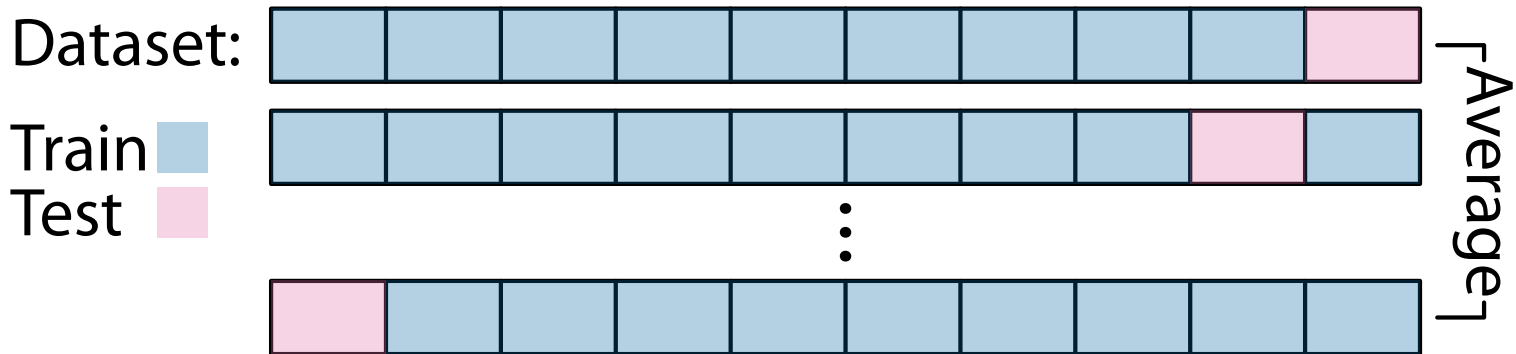
$$\text{variance} = w^2\sigma^2/N$$

Plot of Bias and Variance



Practical Solution: Cross-Validation

- CV (Cross Validation) is the most convenient way to find the best parameter from data without seeing the true parameter
- K -fold cross-validation is typically used



Fisher Information

- Let $p(x; \xi)$ be a distribution with a parameter ξ
- The **Fisher information** $g(\xi)$ of ξ is

$$g(\xi) = \mathbf{E} \left[\left(\frac{\partial}{\partial \xi} \log p(x; \xi) \right)^2 \right] = \sum_{x \in \mathcal{S}} p(x; \xi) \left(\frac{\partial}{\partial \xi} \log p(x; \xi) \right)^2$$

- If there are multiple parameters $\xi = (\xi_1, \xi_2, \dots, \xi_m)$, the **Fisher information matrix** is an $m \times m$ matrix $G(\xi)$ given as

$$g(\xi)_{ij} = \mathbf{E} \left[\frac{\partial}{\partial \xi_i} \log p(x; \xi) \frac{\partial}{\partial \xi_j} \log p(x; \xi) \right]$$

Cramér-Rao Lower Bound

- Let ξ be unbiased: $\mathbf{E}[\hat{\xi}] = \xi^*$
- Cramér-Rao inequality:

$$E \geq \frac{1}{N} G(\xi)^{-1}$$

where $E = (e_{ij})$, each $e_{ij} = \mathbf{E} \left[\left(\hat{\xi}_i - \xi_i^* \right) \left(\hat{\xi}_j - \xi_j^* \right) \right]$

- E coincides with the covariance matrix, $e_{ii} = \mathbf{E} \left[\left(\hat{\xi}_i - \xi_i^* \right)^2 \right] = \text{var}(\hat{\xi}_i)$
 - $A > B$ if $A - B$ is positive definite
 - C is positive definite if $\mathbf{x}^T C \mathbf{x} > 0$ for any non-zero $\mathbf{x} \in \mathbb{R}^n$
- In MLE, $E \rightarrow (1/N)G(\xi)^{-1}$ when $N \rightarrow \infty$

Example in Gaussian Mean Estimation

- Estimate the **mean** from N data points x_1, x_2, \dots, x_N sampled from a Gaussian distribution $N(\theta^*, \sigma^2)$

- Fisher information:

$$g(\theta) = \frac{1}{\sigma^2}$$

- Cramér-Rao bound:

$$\text{var}[\hat{\theta}] \geq \frac{\sigma^2}{N}$$

- In this case, $\text{var}[\hat{\theta}] = \sigma^2/N$ always holds

Example in Log-linear Model

- Fisher information:

$$g_{xy} = \sum_{s \in S} \zeta(x, s) \zeta(y, s) p(s) - \eta_x \eta_y$$

- $\zeta(x, s) = 1$ if $x \leq s$ and $\zeta(x, s) = 0$ otherwise

- Cramér-Rao bound:

$$\text{var}(B) \geq \frac{|B|}{2N} + O(N^{-1.5})$$

for a parameter set $B \subseteq S$

Model Selection by AIC

- The **AIC** (Akaike information criterion) is one of the most famous measure of the quality of statistical models

$$\text{AIC} = -2\ell(D) + 2k$$

- $\ell(D)$ is the maximized log-likelihood
 - k is the number of parameters
- Other criteria:
BIC, MDL, GIC, ...