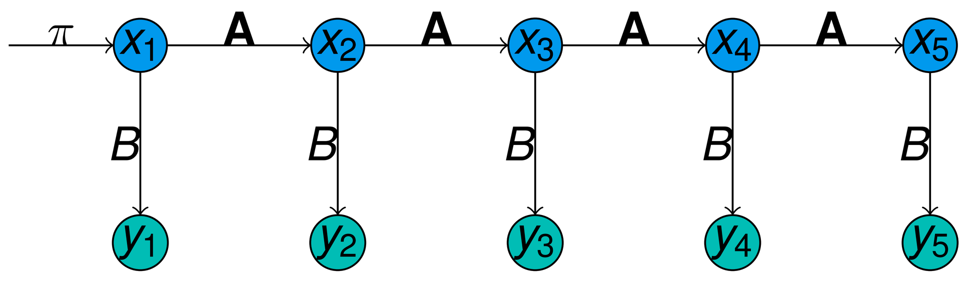
# Co-occurrence-based learning

## How much information do we lose while using approximate probabilties of discrete values?

Uniwersytet Wrocławski

Klaudia Balcer, klaudia.balcer@cs.uni.wroc.pl
Computational Intelligence Research Group,
Institute of Computer Science, University of Wrocław

## Hidden Markov Model

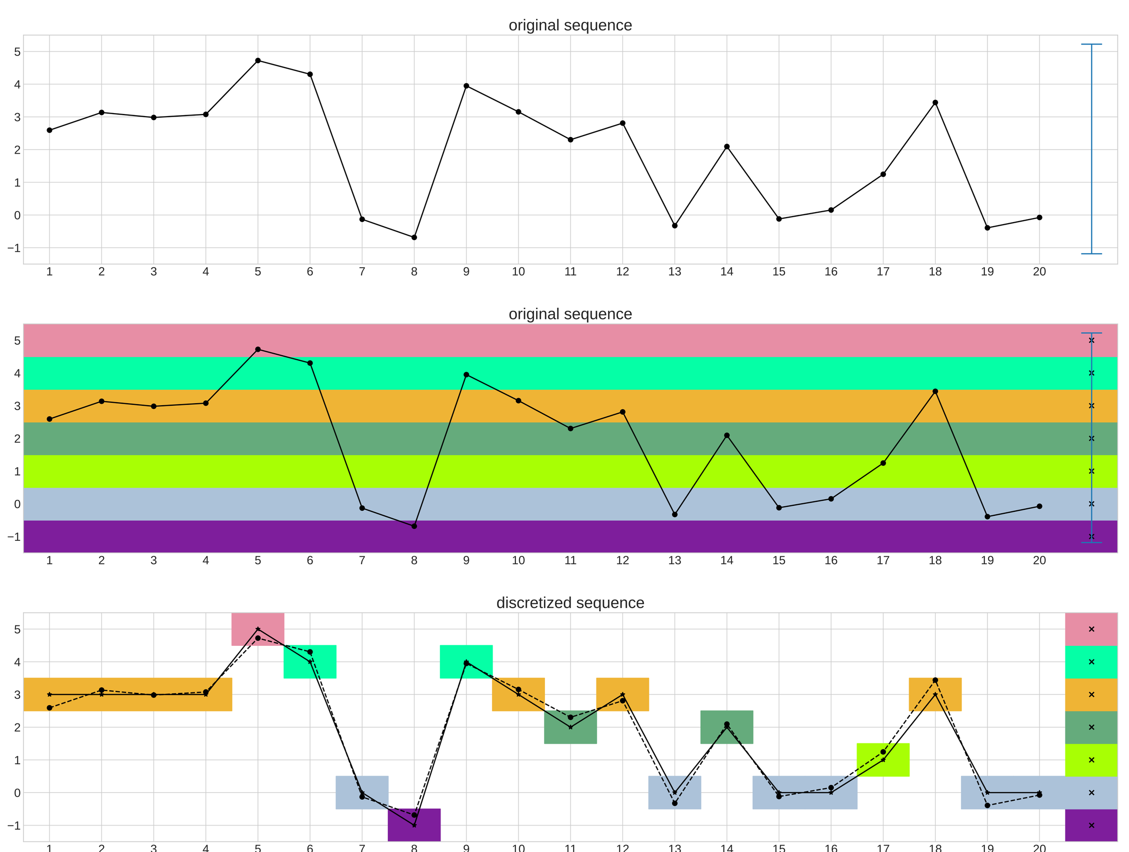


- **HMM**
  $\pi$, **A** - Markov chain, $B$ - emission matrix,
- **GaussianHMM**
  $\pi$, **A** - Markov chain, $B$ - family of Gaussian distributions,
- **DenseHMM** [1] & **GaussianDenseHMM** [2]
  $\pi$, **A** - obtained from embedding, $B$ - emission matrix or a family of Gaussian distributions,
- **FlowHMM** [3]
  $\pi$, **A** - Markov chain, $B$ - family of normalizing flow models,
- ...

## Discretization of continuous values

Let us define a (minimal) hypercube $\hat{\mathcal{Y}}$ containing all observed values $y_{1:T}$ and fix $M^{\mathcal{D}} \in N_+$. Let us define a discrete set $\mathcal{Y}^{\mathcal{D}} = \{v_1^{\mathcal{D}}, \dots, v_{M^{\mathcal{D}}}^{\mathcal{D}}\}$, $v_i^{\mathcal{D}} \in \hat{\mathcal{Y}}$. **Discretization** [4] is a function $\mathcal{D} : R^m \longrightarrow \mathcal{Y}^{\mathcal{D}}$:

$$\mathcal{D}(y) = \arg\min_{v \in \mathcal{Y}^{\mathcal{D}}} \|y - v\|.$$




One needs also to discretize the probabilities; see Eq. (3), (4).

## Training schema

- **EM**
- **GD**
  We can use SGD, Adam, etc., for optimizing
  $$\mathbf{S}, B = \arg\min_{\tilde{\mathbf{S}}, \tilde{B}} \left\| \mathbf{Q}^{gt} - \mathbf{Q} \right\|.$$
  We optimize uncontrained matrices $\tilde{\mathbf{S}}, \tilde{B}$ and convert them via softmax:
  $$\mathbf{S}, B = \arg\min_{\tilde{\mathbf{S}}, \tilde{B}} \left\| \mathbf{Q}^{gt} - \left( \frac{\exp(\tilde{B})}{\sum_{w=1}^{M} \exp(\tilde{B}_w)} \right)^T \frac{\exp(\tilde{\mathbf{S}})}{\sum_{i=1}^{N} \sum_{j=1}^{N} \exp(\tilde{\mathbf{S}}_{ij})} \frac{\exp(\tilde{B})}{\sum_{w=1}^{M} \exp(\tilde{B}_w)} \right\|$$
- **NMF** [5]
  We will use the pseudo inverses (from Eq. (1)):
  $$\mathbf{S} = B^{-1}\mathbf{Q}(B^T)^{-1}, \quad (2a)$$
  $$B = ((\mathbf{S}B)^{-1}\mathbf{Q})^T, \quad (2b)$$
  $$B = \mathbf{Q}(B^T\mathbf{S})^{-1}. \quad (2c)$$
  The update procedure is:

  **pseudo inverse > ReLU > softmax.**

  In each iteration, we update the following:
  - **S** with Eq. (2a)
  - $B$ with Eq. (2b)
  - **S** with Eq. (2a)
  - $B$ with Eq. (2c)

## Co-occurrence matrix for discrete emission

**Empirical co-occurrence matrix**:

$$\mathbf{Q}^{gt}_{vw} = \frac{\#\{t : y_t = v, y_{t+1} = w\}}{T-1}$$

*Example*:
Sequence from HMM (letters are observations, colors are hidden states) with underlined co-occurrences of the values $a$ and $b$:

$a, c, a, b, b, c, a, a, b, c, b, c, d, c, b, b.$

Counts the co-occurrences for each pair of values:

| | a | b | c | d |
|---|---|---|---|---|
| a | 1 | 2 | 1 | 0 |
| b | 0 | 2 | 3 | 0 |
| c | 2 | 2 | 0 | 1 |
| d | 0 | 0 | 1 | 0 |

**Co-occurrence matrix derived from model parameters**:

$$\mathbf{Q}_{vw} = P(Y_t = v, Y_{t+1} = w)$$
$$= \sum_{i=1}^{N} \sum_{j=1}^{N} P(X_t = i) B_i(v) \mathbf{A}_{ij} B_j(w).$$

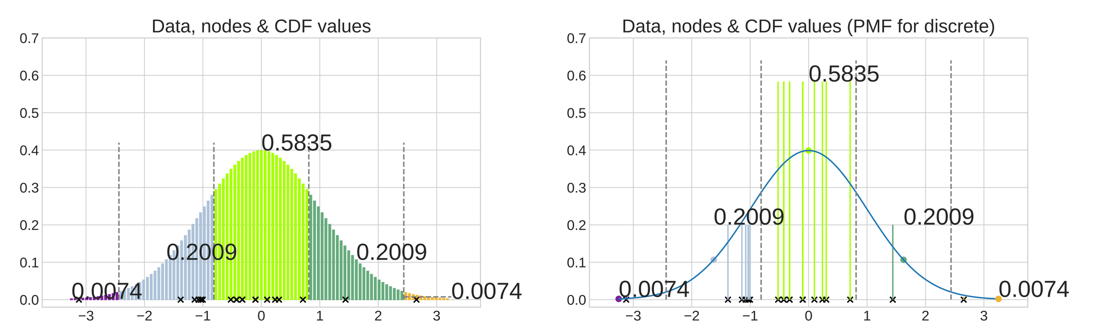
Let us assume that $\pi$ is the stationary distribution of the Markov chain:

$$\mathbf{Q}_{vw} = \sum_{i=1}^{N} \sum_{j=1}^{N} \pi_i B_{iv} \mathbf{A}_{ij} B_{jw},$$

$$\mathbf{Q} = B^T \mathbf{S} B, \text{ where } \mathbf{S}_{ij} = \pi_i A_{ij} (1)$$

## Exact probabilities in discretization

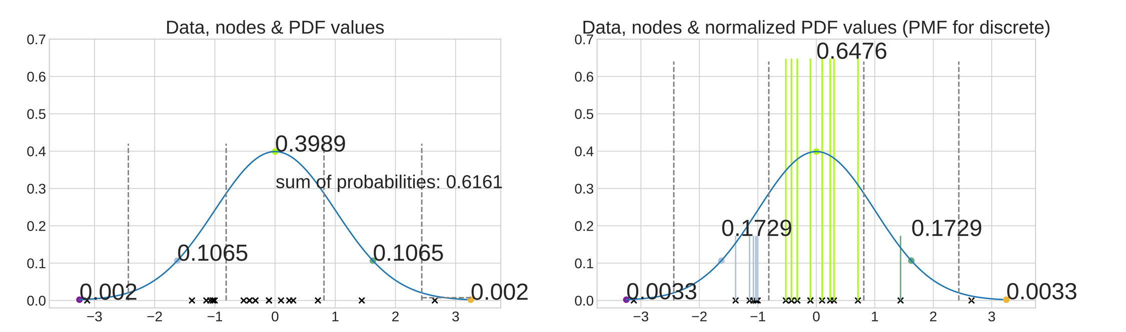
$$P(\mathcal{D}(y) = v^{\mathcal{D}} | x = i) = \int_{\{y : \mathcal{D}(y) = v^{\mathcal{D}}\}} B_i(y) dy \quad (3)$$




- find the region boundaries (the set of observations resulting in a given discrete value)
- integrate the PDF within the region

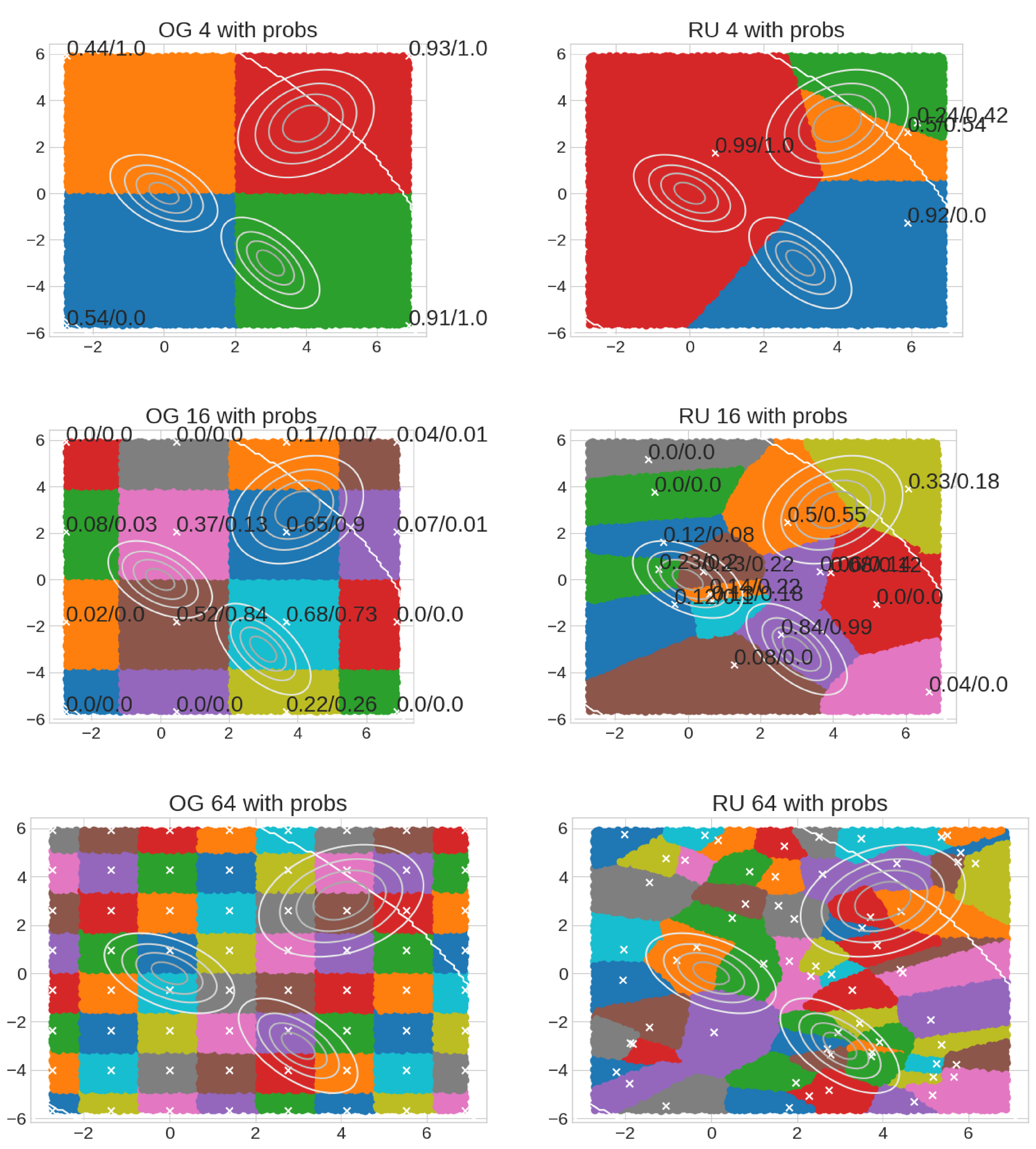
## Approximate probabilities in discreti

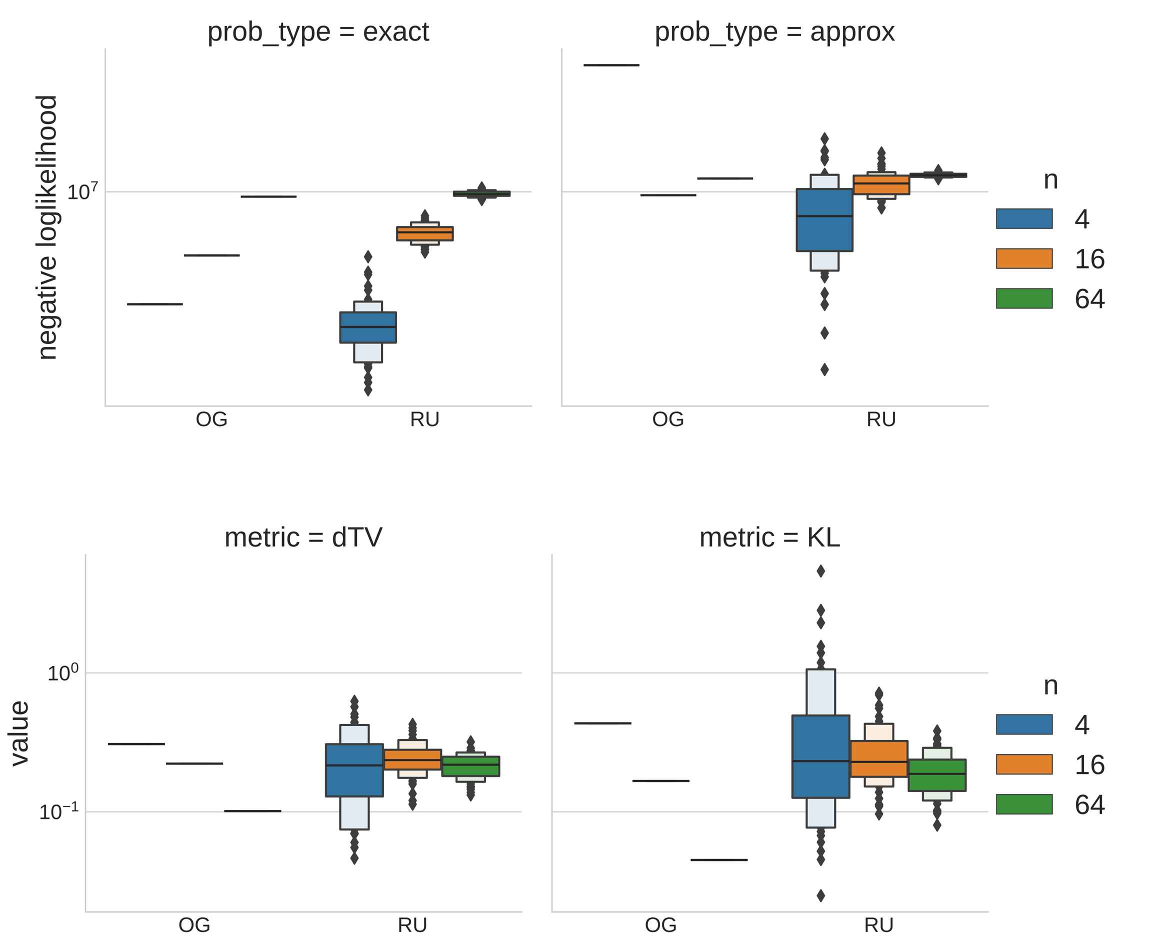
$$P(\mathcal{D}(y) = v^{\mathcal{D}} | x = i) \approx \frac{B_i(v^{\mathcal{D}})}{\sum_{w^{\mathcal{D}} \in \mathcal{Y}^{\mathcal{D}}} B_i(w^{\mathcal{D}})} \quad (4)$$




- calculate the PDF function in the discrete values
- normalize the obtained PDF values (to assure obtaining a discrete probability distribution)

## Comparison of exact and approximate probabilities




$$d_{TV}(\mu, \nu) = \tfrac{1}{2} \sum_i |\mu(v_i^{\mathcal{D}}) - \nu(v_i^{\mathcal{D}})|, \; KL(\mu \| \nu) = \sum_i \mu(v_i^{\mathcal{D}}) \log \frac{\mu(v_i^{\mathcal{D}})}{\nu(v_i^{\mathcal{D}})}$$

## Conclusions

| | exact | approx |
|---|---|---|
| fast to compute | ✗ | ✓ |
| available for all distributions | ✗ | ✓ |
| domain aware | ✓ | ✗ |
| accurate | ✓ | ✗ |

- Many multivariate distributions don't have analytical formulas for CDFs (thus, we estimate the exact probabilities with Monte Carlo methods, which is time-consuming to perform in each training step)
- The approximation of probabilities disrupts the distribution of discrete values visibly (over 10%) and may affect the learning process.
- Exact values result in better likelihood than approximate.
- The random uniform grid is expected to work better than the ordinary grid.
- Discrete likelihood is incomparable to continuous likelihood.

## References

[1] Joachim Sicking, Maximilian Pintz, Maram Akila, and Tim Wirtz. Densehmm: Learning hidden markov models by learning dense representations, 2020.

[2] Klaudia Balcer and Piotr Lipinski. Extending densehmm with continuous emission. *ICONIP*, 2023.

[3] Pawel Lorek, Rafał Nowak, Tomasz Trzcinski, and Maciej Zieba. FlowHMM: Flow-based continuous hidden markov models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[4] Klaudia Balcer. Flowhmm. discretisation in the co-occurence-based learning algorithm. Master's thesis, University of Wrocław, 2023.

[5] Balaji Lakshminarayanan and Raviv Raich. Non-negative matrix factorization for parameter estimation in hidden markov models. In *2010 IEEE International Workshop on Machine Learning for Signal Processing*. IEEE, aug 2010.

## Acknowledgement

This work is a part of a national research project on discovering hidden structure in large datasets (supported by the **Polish National Science Centre** (NCN) under grant OPUS 18 no. 2019/35/B/ST6/04379).

I acknowledge the support of **G-Research** in the form of a Travel Grant, which enabled me to attend this conference.