Week 8: Mixture Models & Expectation-Maximisation

Matthew Caldwell

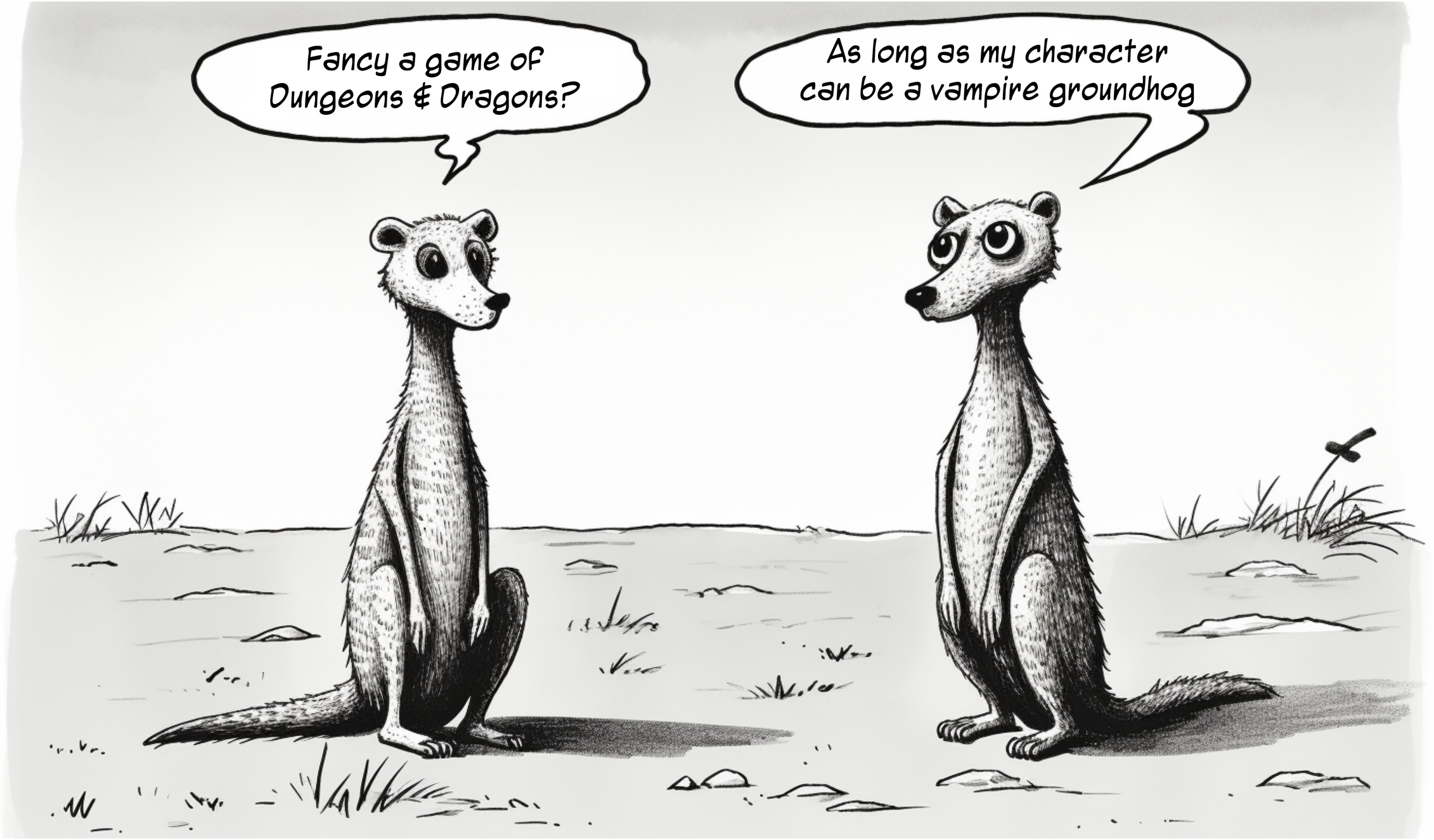
COMP0088 Introduction to Machine Learning • UCL Computer Science • Autumn 2023

Admin

- Good news: no more lab exercises!
- Bad news: exam practice
- Group 1 will be shuffled into Groups 2 & 3
- All sessions will be in 66GS G01

Week 8 Recap

The Invisible Enemy


Fancy a game of Dungeons & Dragons?
As long as my character can be a vampire groundhog

d6

d20

Dice Model 1

- **Repeat**
  - → **Draw a die**
  - → **Roll it and record result**
  - → **Return it to the bag**

Dice Model 2

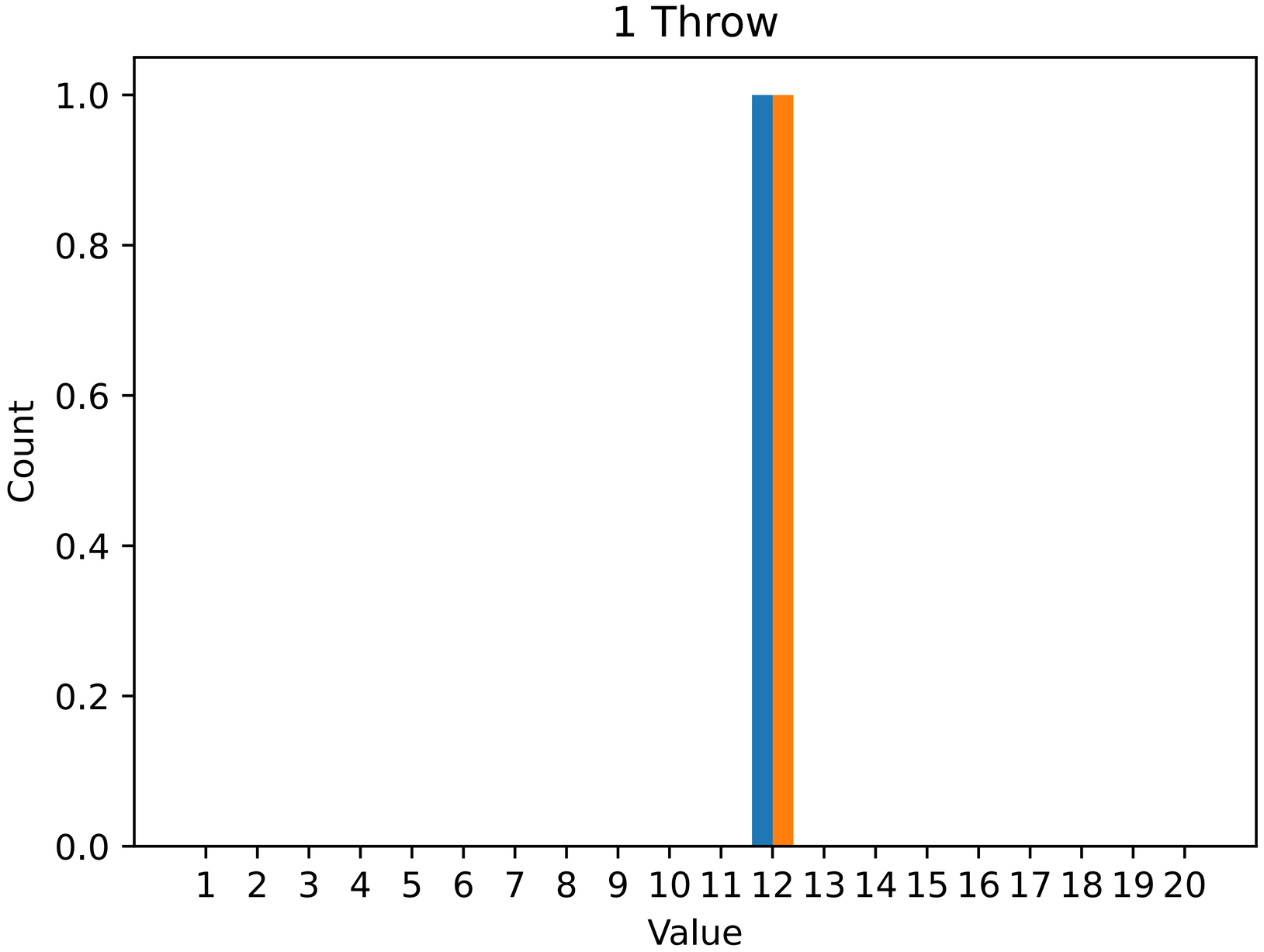
- **Draw a die**
- **Repeat**
  - → **Roll it and record result**
  - → **Roll again; if roll=1, return die to bag and draw again**

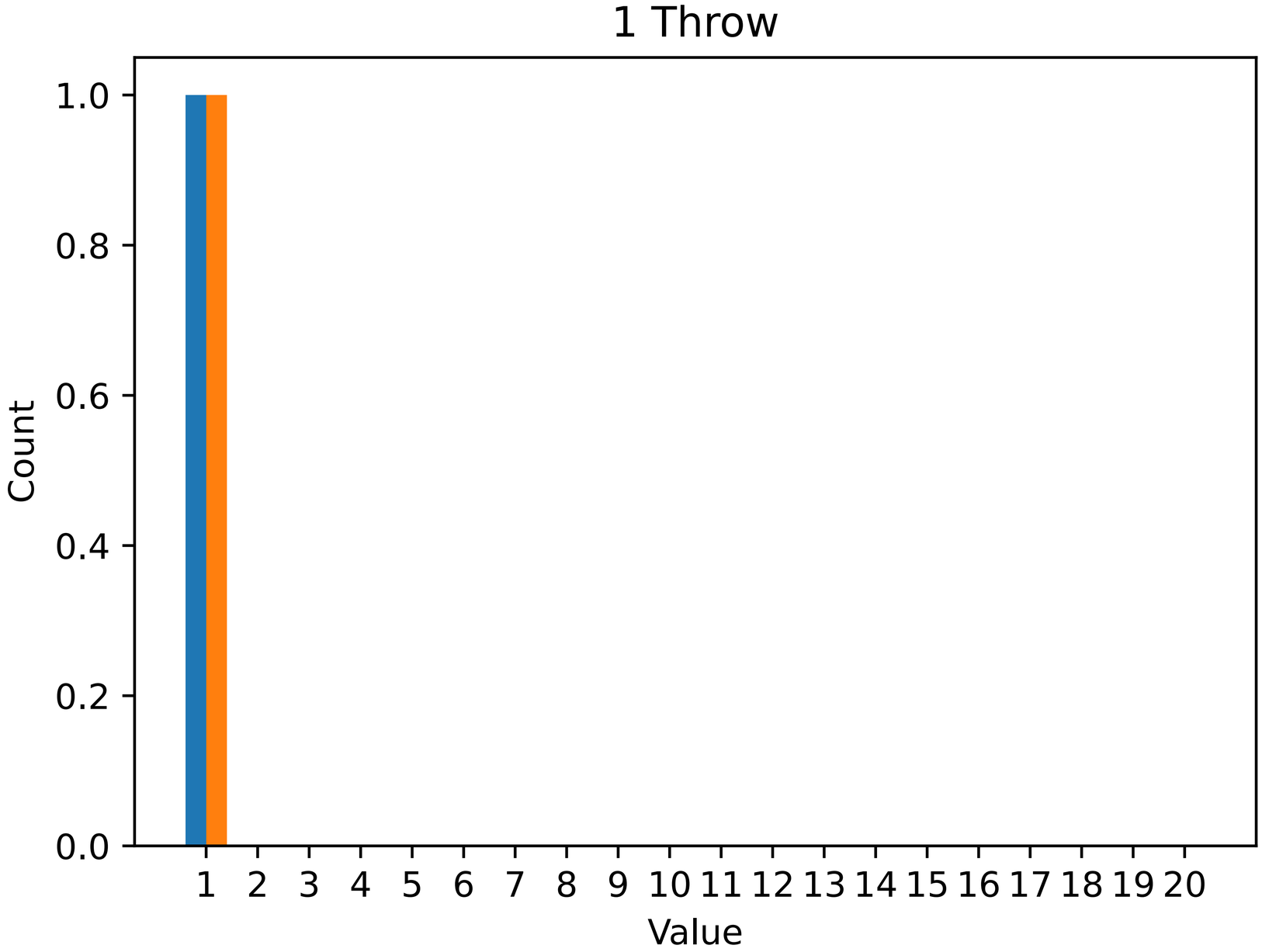
Model 1

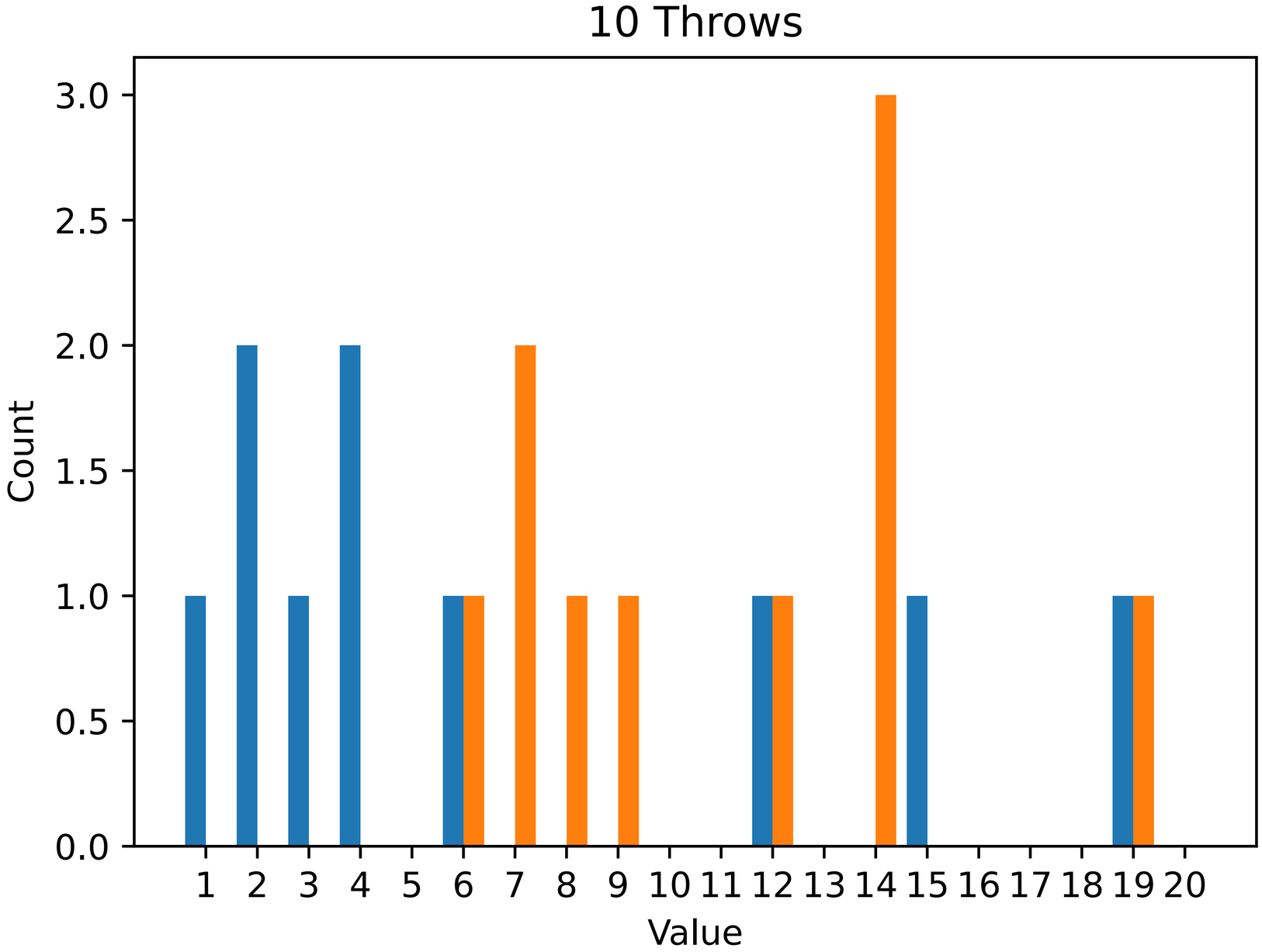
Independent Mixture

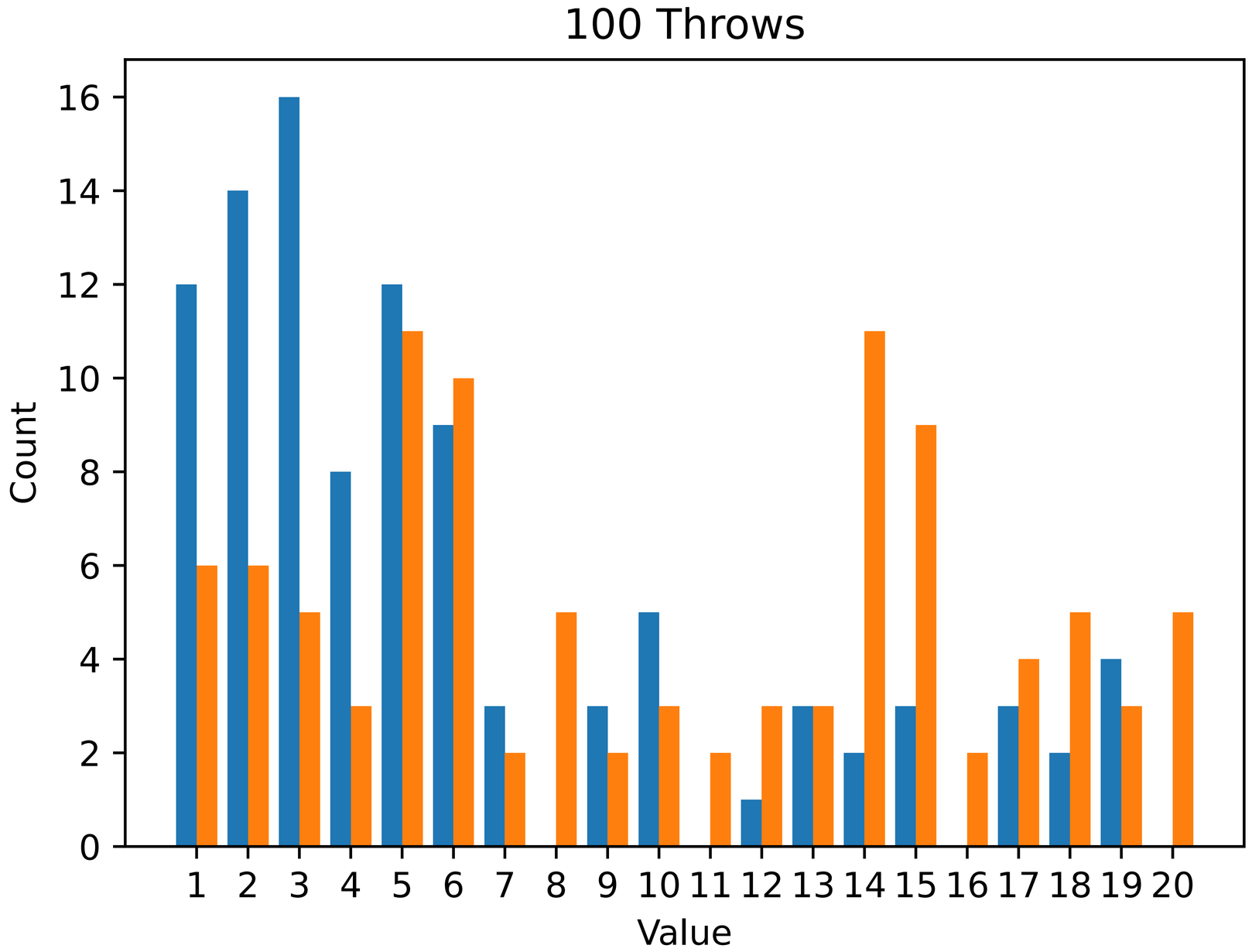
Model 2

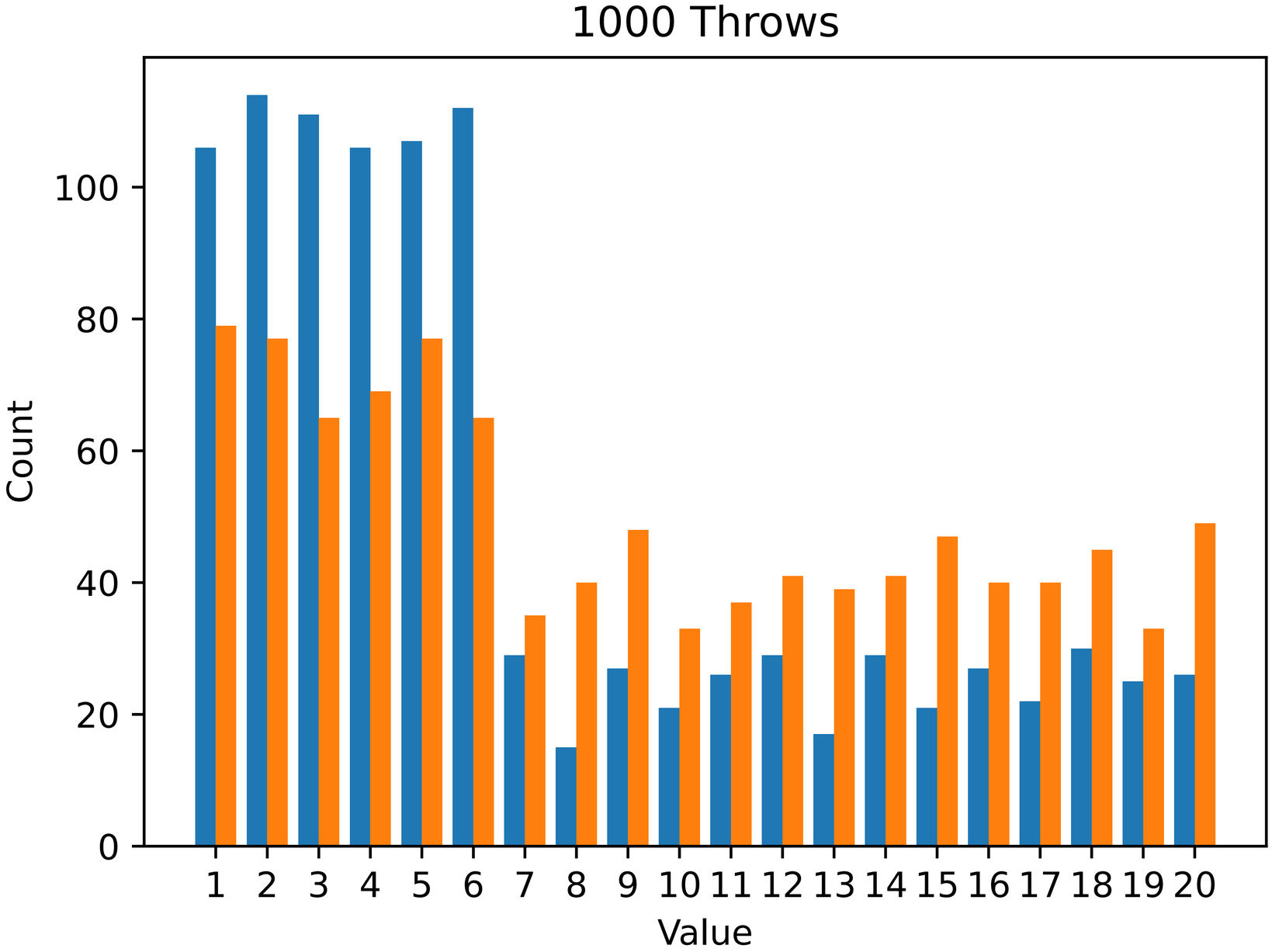
Hidden Markov

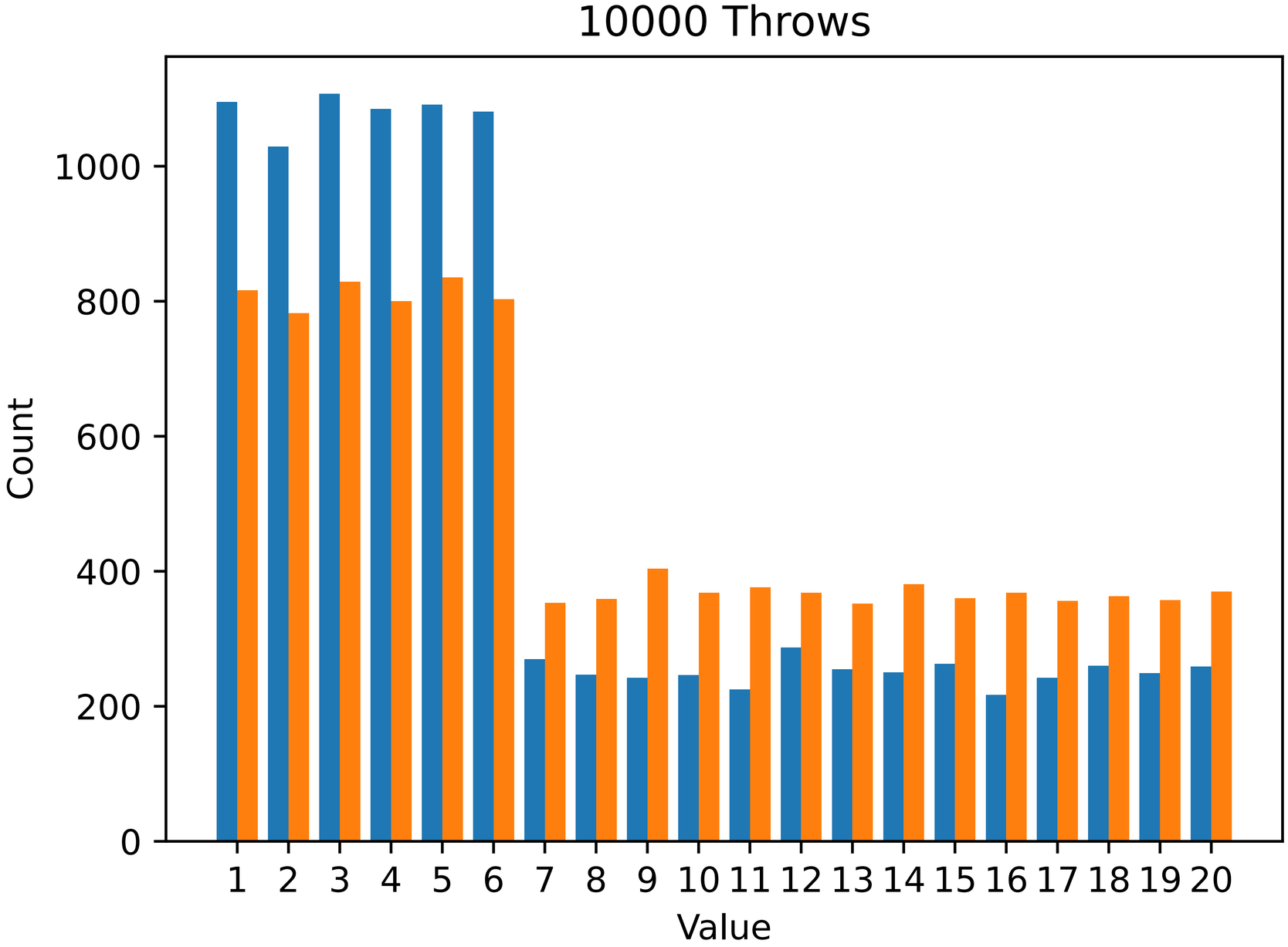

1 Throw
Count
1.0
0.8
0.6
0.4
0.2
0.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Value

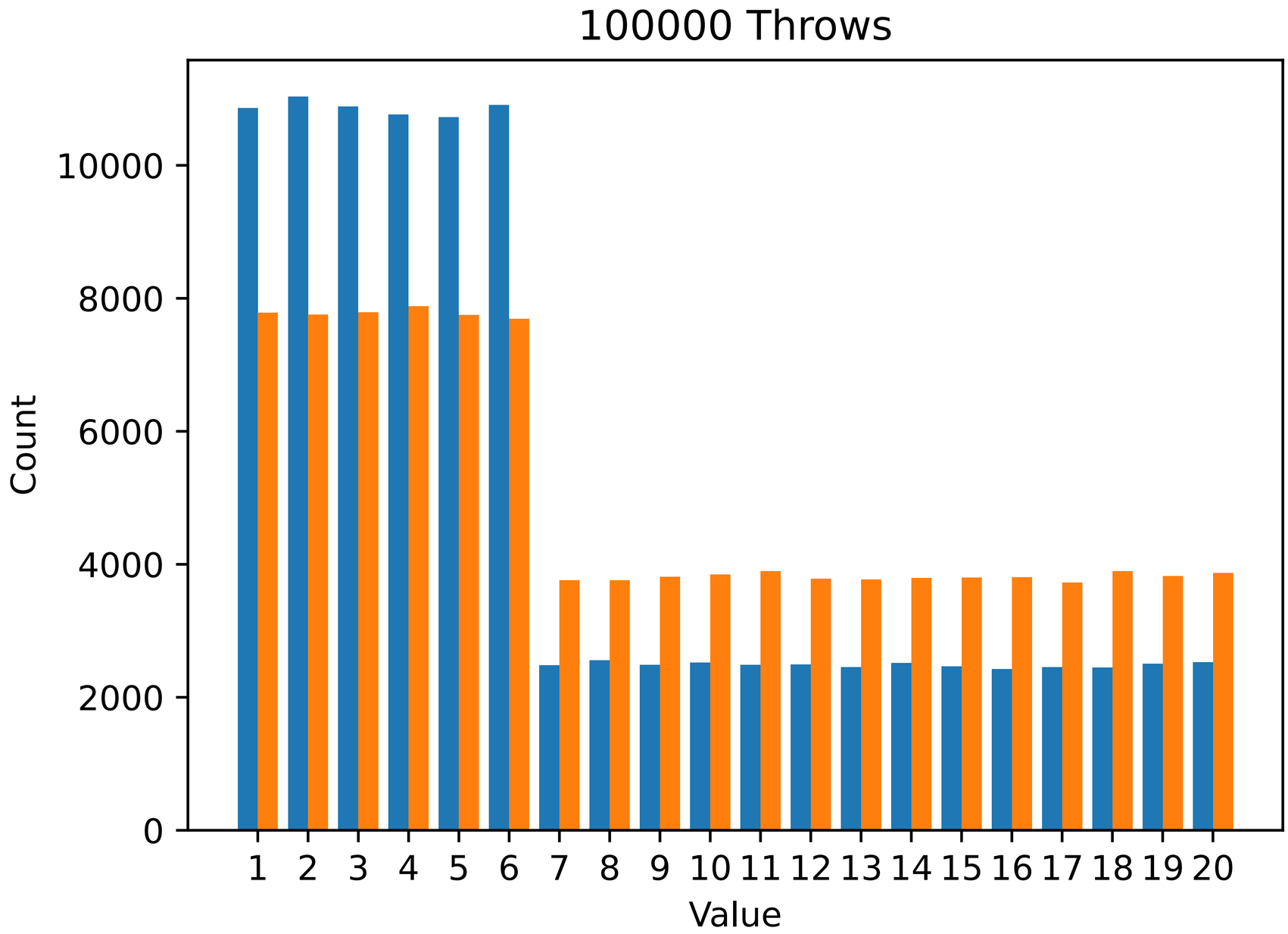

1 Throw
Count
1.0
0.8
0.6
0.4
0.2
0.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Value

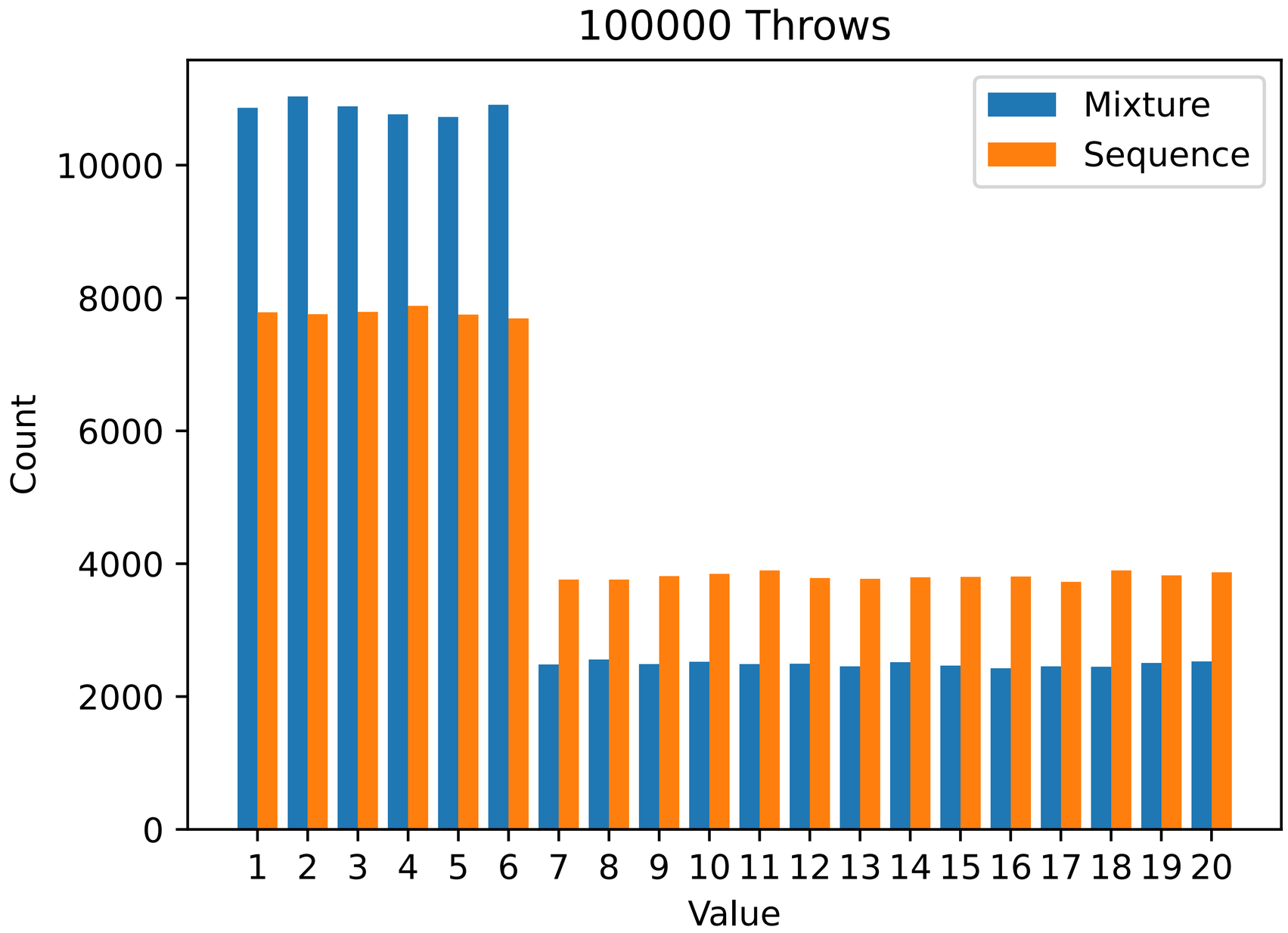

10 Throws
Count
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Value

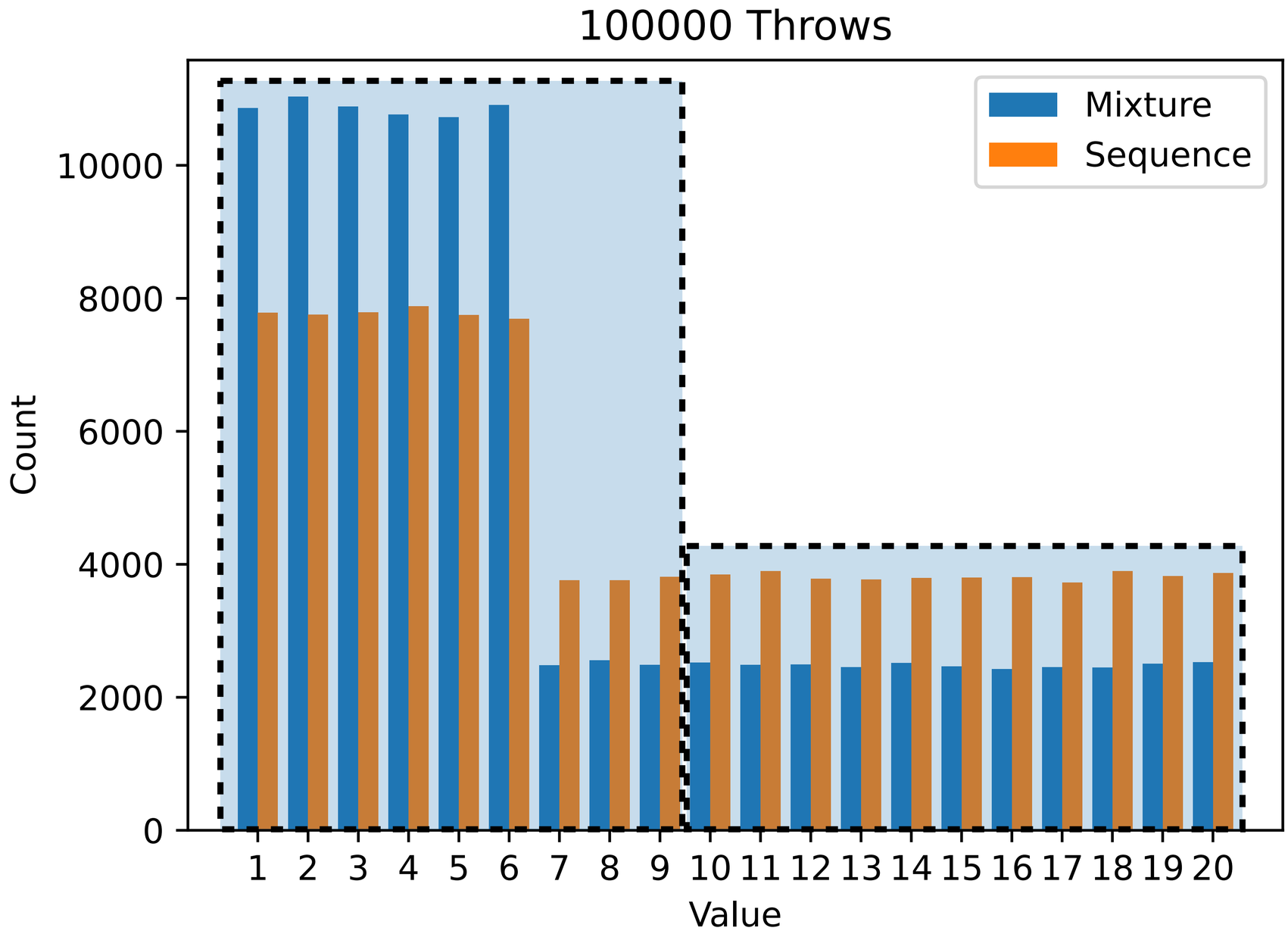

100 Throws
Count
Value
0
2
4
6
8
10
12
14
16
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

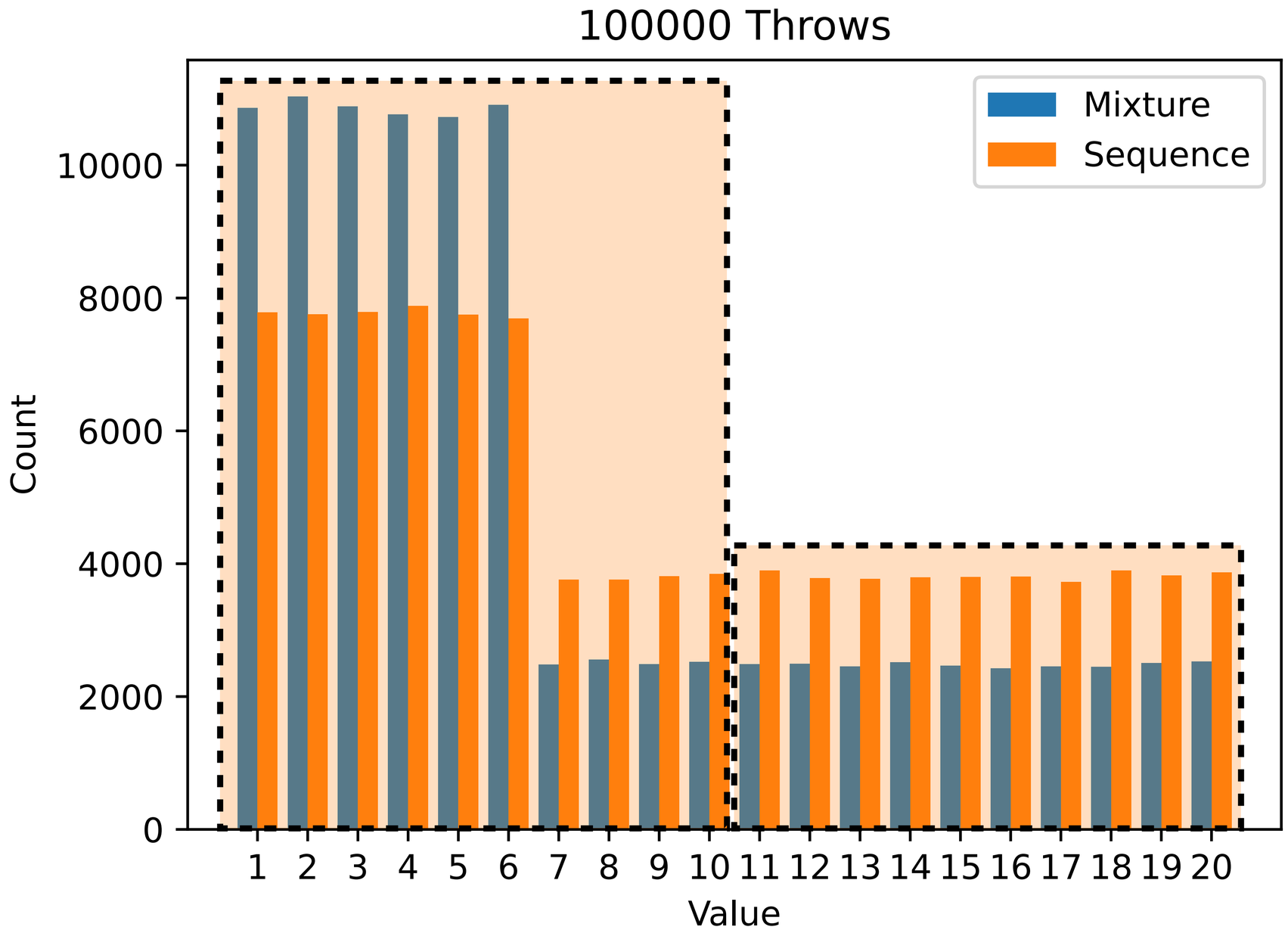

1000 Throws
Count
0
20
40
60
80
100
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Value

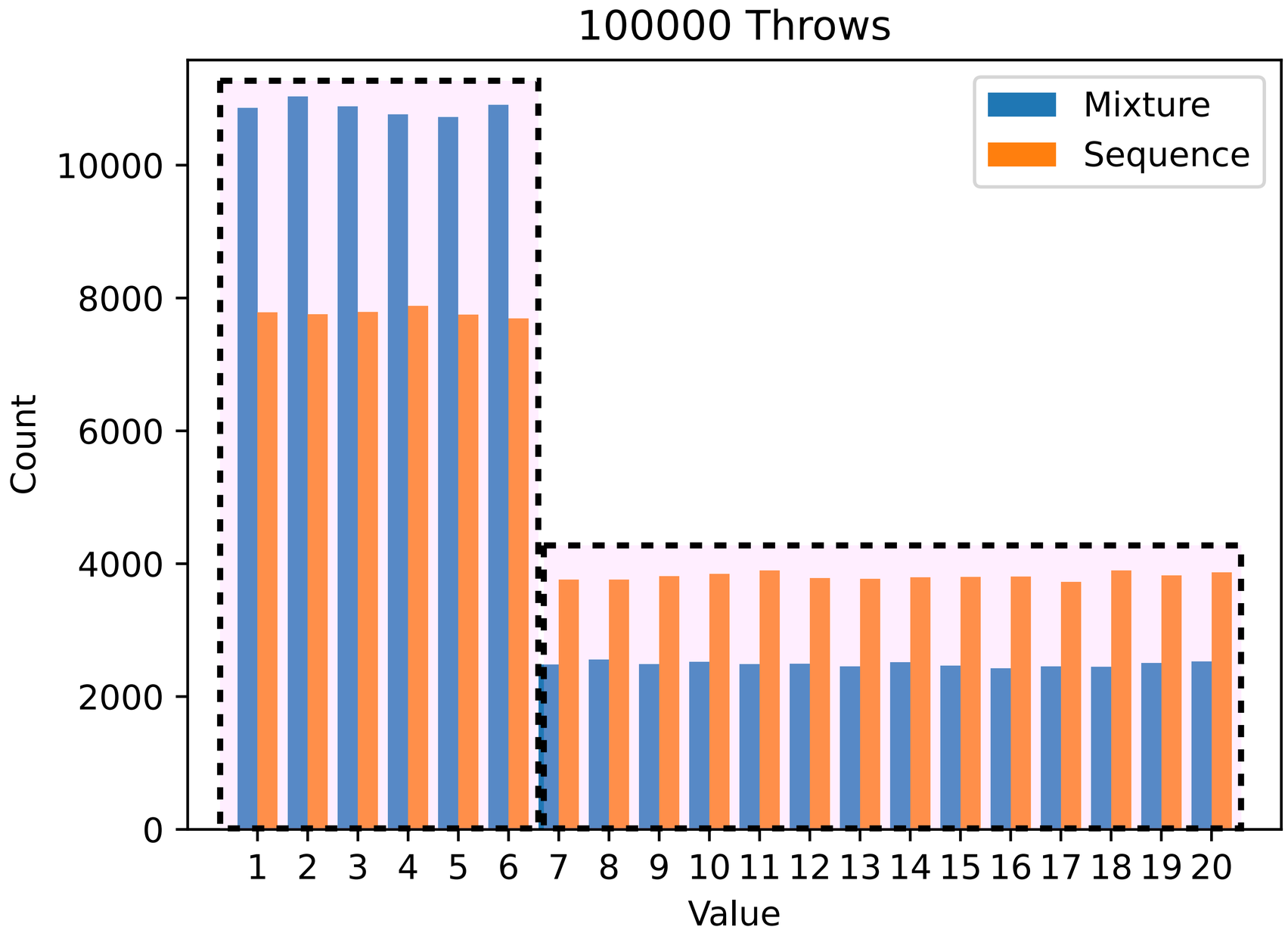

10000 Throws
Count
0
200
400
600
800
1000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Value

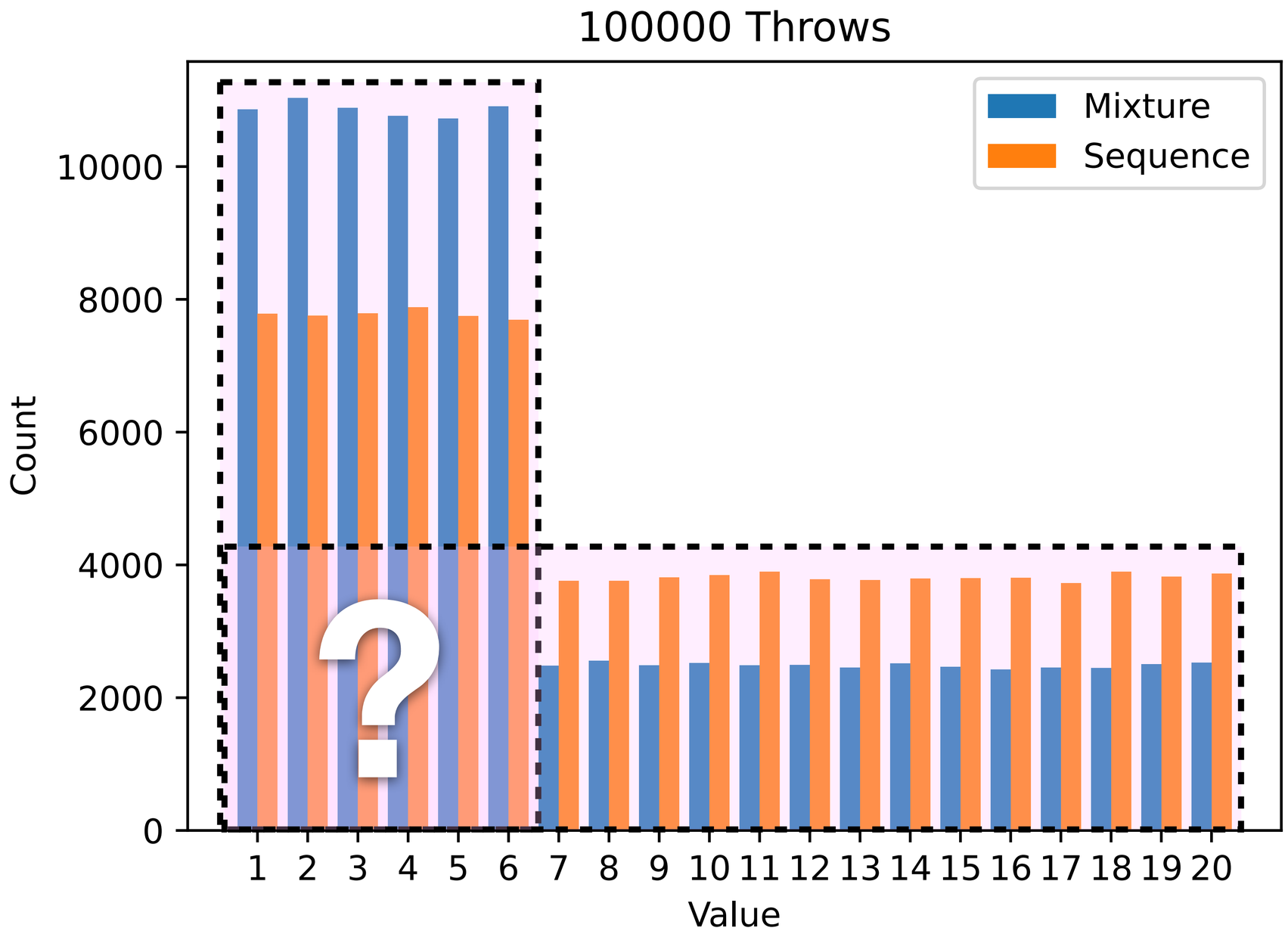

100000 Throws
Count
0
2000
4000
6000
8000
10000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Value


100000 Throws
Mixture
Sequence
Count
0
2000
4000
6000
8000
10000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Value


100000 Throws
Mixture
Sequence
Count
0
2000
4000
6000
8000
10000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Value

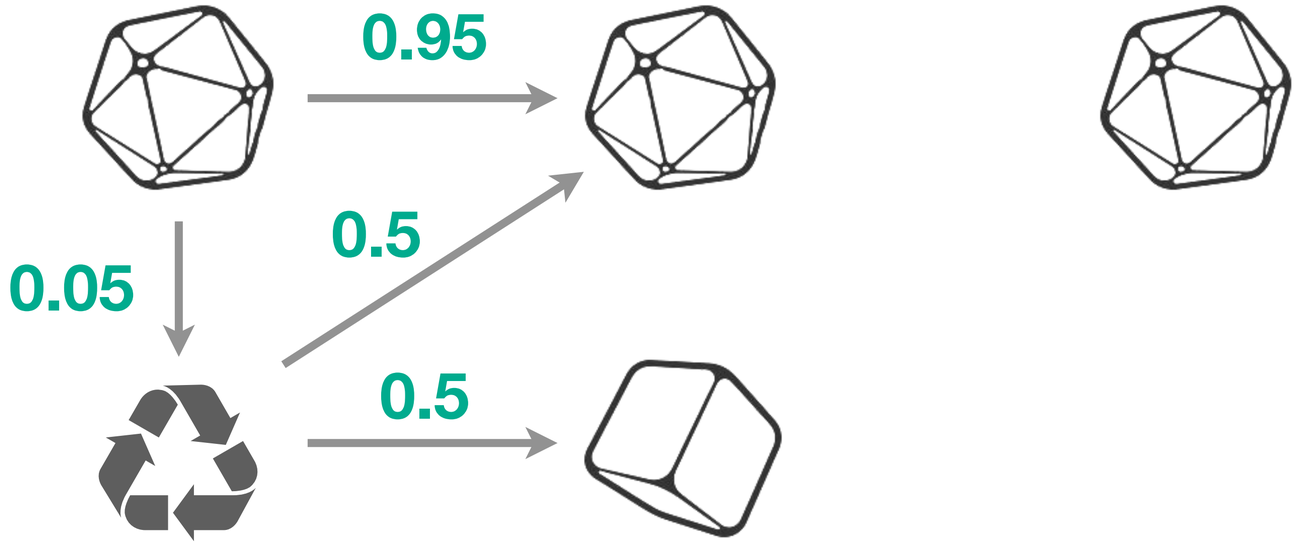

100000 Throws
Mixture
Sequence
Count
0
2000
4000
6000
8000
10000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Value


100000 Throws
Mixture
Sequence
Count
0
2000
4000
6000
8000
10000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Value


100000 Throws
Mixture
Sequence
Count
10000
8000
6000
4000
2000
0
?
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Value

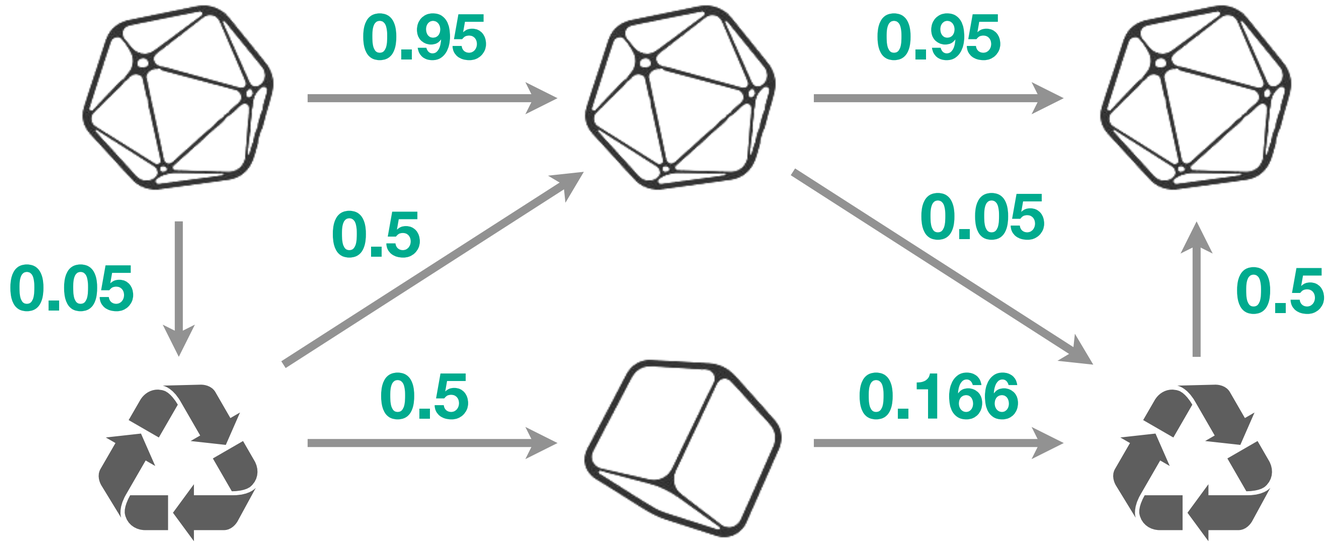
... 13 2 17 ...

... 13 2 17 ...



“Filtering”

... 13 2 17 ...



“Smoothing”

Gaussian Mixture Models

$$\tilde{\mathbf{x}}_i = \{ \mathbf{x}_i , z_i \}$$

$$\mathbf{x}_i \in \mathbb{R}^d \longrightarrow \text{ observed}$$

$$z_i \in \{1, 2, ..., k\} \longrightarrow \text{ invisible}$$

$$z_i \sim \text{Cat}(k, \boldsymbol{\alpha})$$

$$\mathbf{x}_i | (z_i = j) \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

$$P(\mathbf{x}) = \sum_{i}^{k} \alpha_i \phi(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

$$P(\mathbf{x}) = \sum_{i}^{k} \alpha_i \phi(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

$$P(\mathbf{x}) = \sum_{i}^{k} \alpha_i \phi(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

$$\phi(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^{d}|\boldsymbol{\Sigma}|}}$$

$$P(\mathbf{x}) = \sum_{i}^{k} \alpha_i \phi(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

$$L(\boldsymbol{\theta}) = \sum_{i}^{k} \alpha_i \phi(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

$$\boldsymbol{\theta} = \{\boldsymbol{\alpha}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_k\}$$

$$l(\boldsymbol{\theta}) = \log\left(\sum_{i}^{k} \alpha_i \phi(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\right)$$

Maximum Likelihood Estimators

Gaussian

μ = sample mean

$\sum$ = sample covariance

Categorical

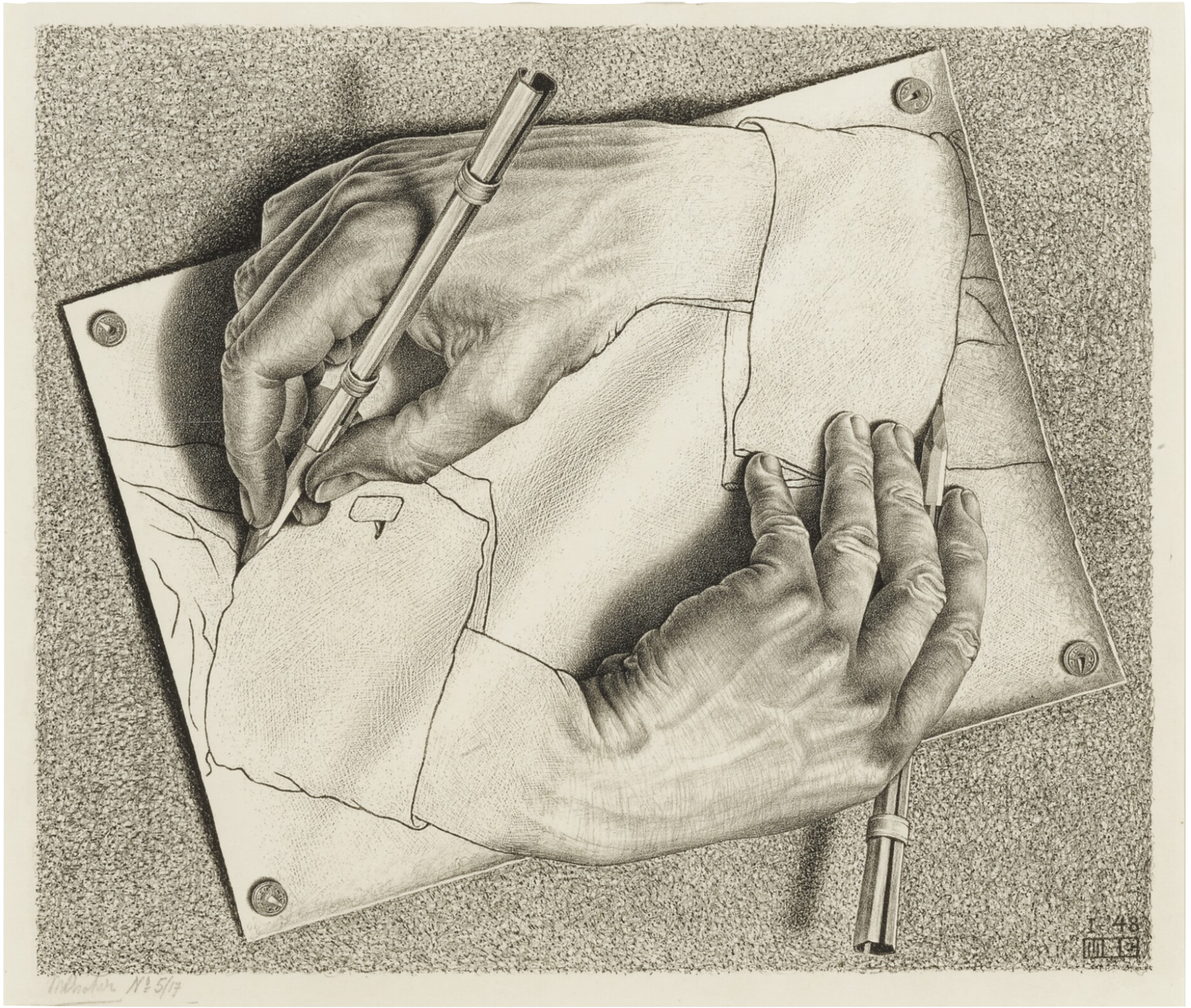
α = sample proportions

$$\alpha_j = \frac{\sum_i \mathbf{1}(z_i = j)}{n}$$

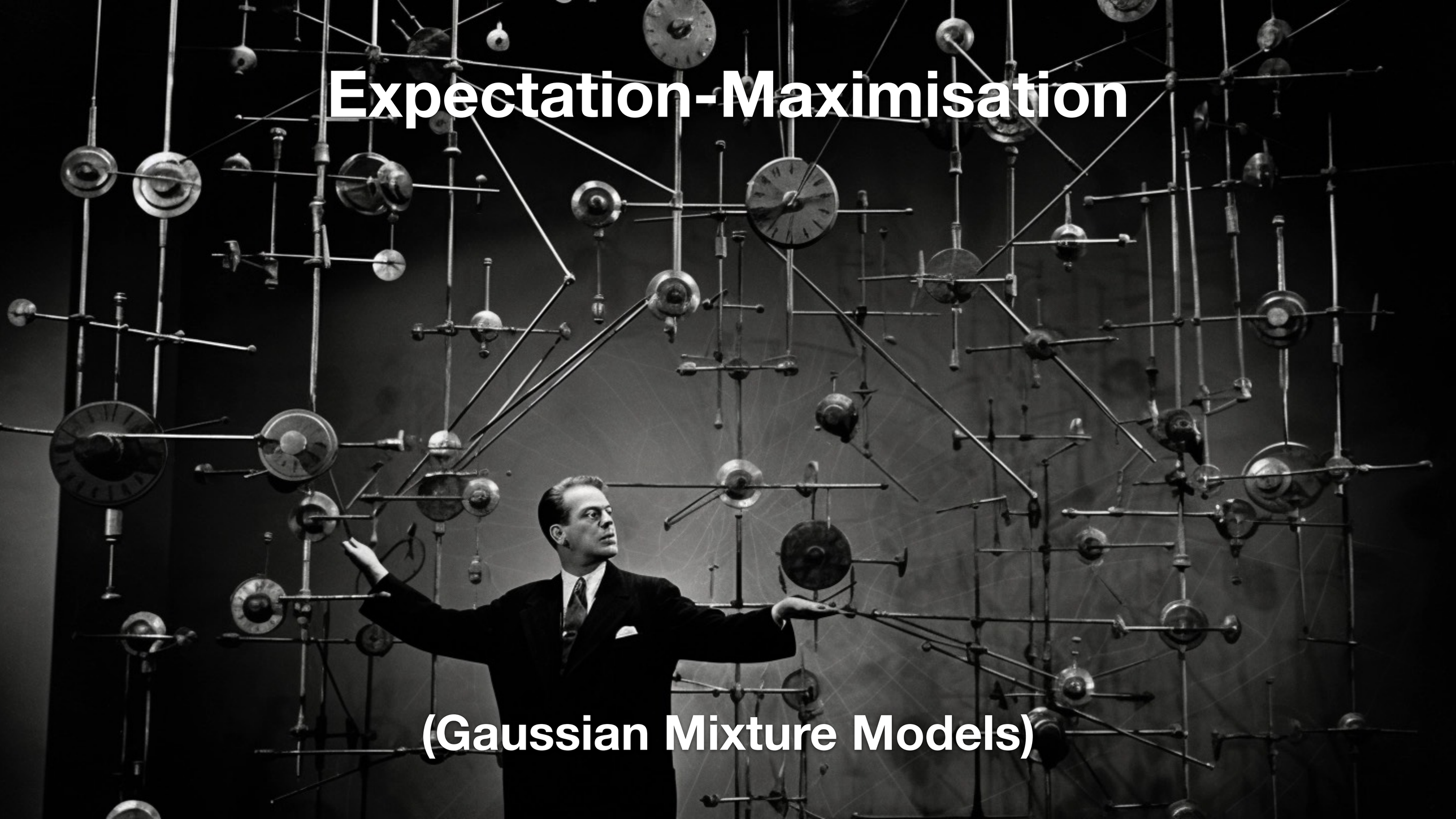
$$\boldsymbol{\mu}_j = \frac{\sum_i \mathbf{1}(z_i = j)\mathbf{x}_i}{\sum_i \mathbf{1}(z_i = j)}$$

$$\boldsymbol{\Sigma}_j = \frac{\sum_i \mathbf{1}(z_i = j)(\mathbf{x}_i - \boldsymbol{\mu}_j)(\mathbf{x}_i - \boldsymbol{\mu}_j)^\mathsf{T}}{\sum_i \mathbf{1}(z_i = j)}$$

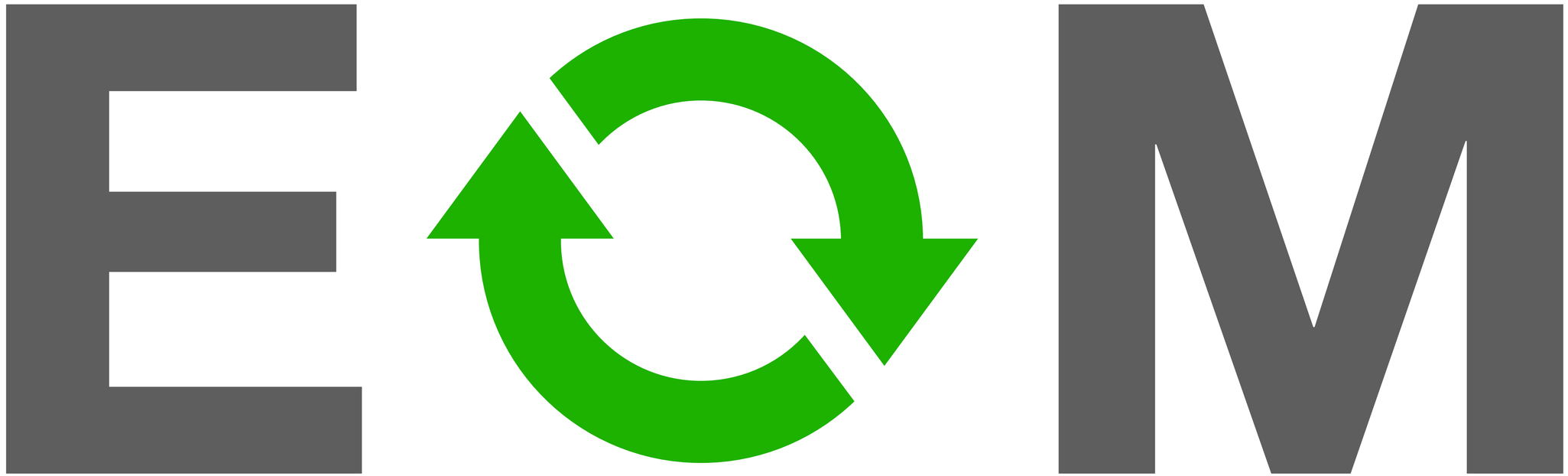
$$z_i = \underset{z}{\operatorname{argmax}}\, P(\mathbf{x}_i | z)$$



Expectation-Maximisation



(Gaussian Mixture Models)


EOM

E-step

Estimate distribution of Z

$$\gamma_{i,j} = \frac{P(\mathbf{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\alpha_j}{\sum_l P(\mathbf{x}_i|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)\alpha_l}$$

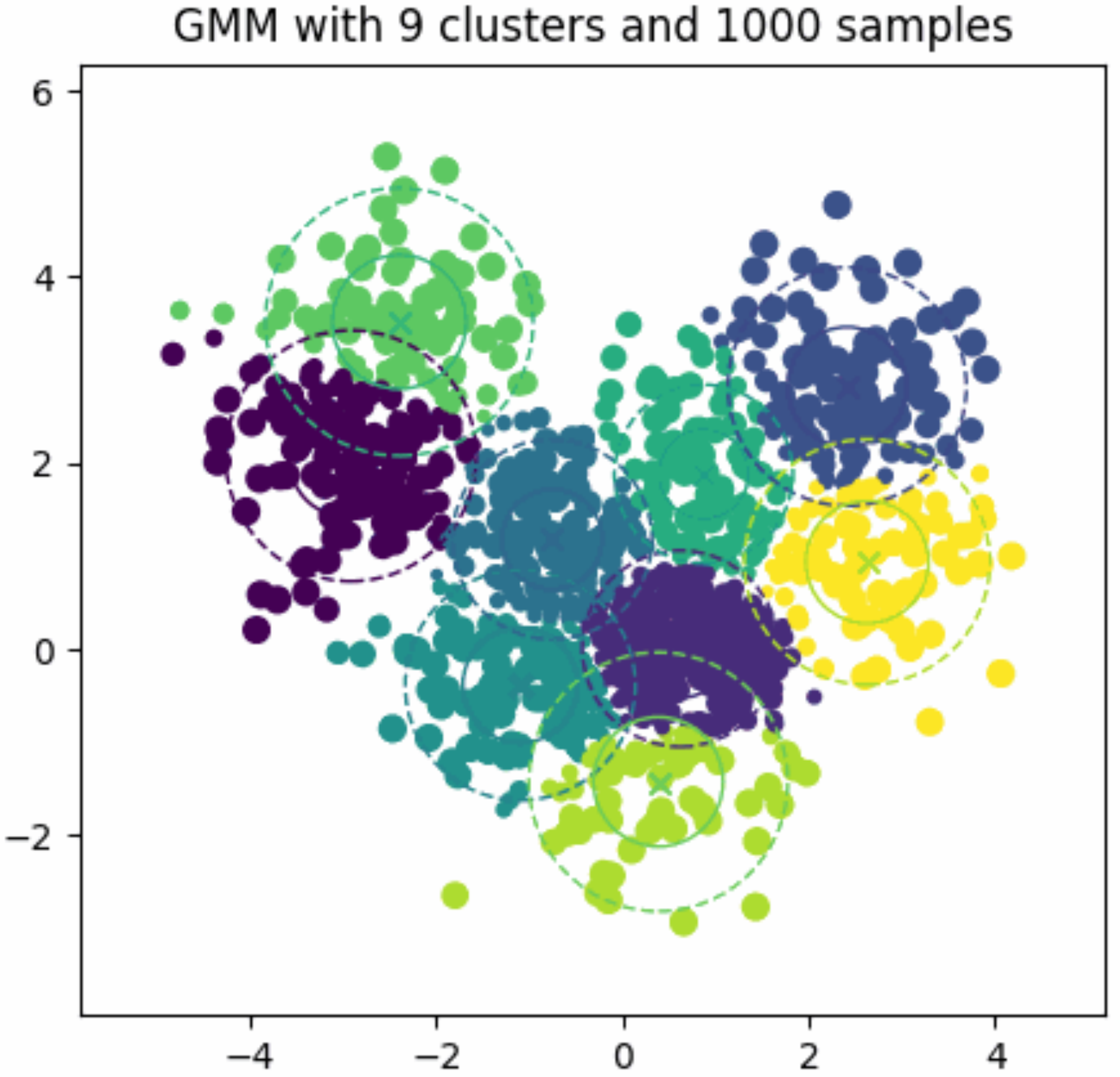
M-step

Maximise likelihood of θ

$$\alpha_j = \frac{\sum_i \gamma_{i,j}}{n}$$

$$\boldsymbol{\mu}_j = \frac{\sum_i \gamma_{i,j} \mathbf{x}_i}{\sum_i \gamma_{i,j}}$$

$$\boldsymbol{\Sigma}_j = \frac{\sum_i \gamma_{i,j} (\mathbf{x}_i - \boldsymbol{\mu}_j)(\mathbf{x}_i - \boldsymbol{\mu}_j)^\mathsf{T}}{\sum_i \gamma_{i,j}}$$


GMM with 9 clusters and 1000 samples
6
4
2
0
−2
−4
−2
0
2
4

Expectation-Maximisation

(Latent Variable Models)

$$\tilde{\mathbf{x}}_i = \left\{ \mathbf{x}_i \, , \, \mathbf{z}_i \right\}$$

$$\mathbf{x}_i \in \mathbb{R}^d \longrightarrow \text{ observed}$$

$$\mathbf{z}_i \in \mathcal{Z} \longrightarrow \text{ invisible}$$

$$\tilde{\mathbf{x}} = \{\mathbf{X}, \mathbf{Z}\}$$

$$P(\tilde{\mathbf{X}};\boldsymbol{\theta}) = P(\mathbf{X},\mathbf{Z};\boldsymbol{\theta})$$
$$= P(\mathbf{X}|\mathbf{Z};\boldsymbol{\theta})P(\mathbf{Z};\boldsymbol{\theta})$$

$$L(\boldsymbol{\theta}) = P(\mathbf{X}|\mathbf{Z};\boldsymbol{\theta})P(\mathbf{Z};\boldsymbol{\theta})$$

$$P(\mathbf{Z}|\mathbf{X};\boldsymbol{\theta}) = \frac{P(\mathbf{X}|\mathbf{Z};\boldsymbol{\theta})P(\mathbf{Z};\boldsymbol{\theta})}{P(\mathbf{X};\boldsymbol{\theta})}$$

E-step

Estimate distribution of Z

$$Q^{(t)}(\mathbf{Z}) = P(\mathbf{Z}|\mathbf{X}; \boldsymbol{\theta}^{(t-1)})$$

M-step

Maximise likelihood of θ

$$l^{(t)}(\boldsymbol{\theta}) = \sum_{\mathbf{Z} \in \mathcal{Z}} Q^{(t)}(\mathbf{Z}) \log P(\mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}^{(t-1)})$$

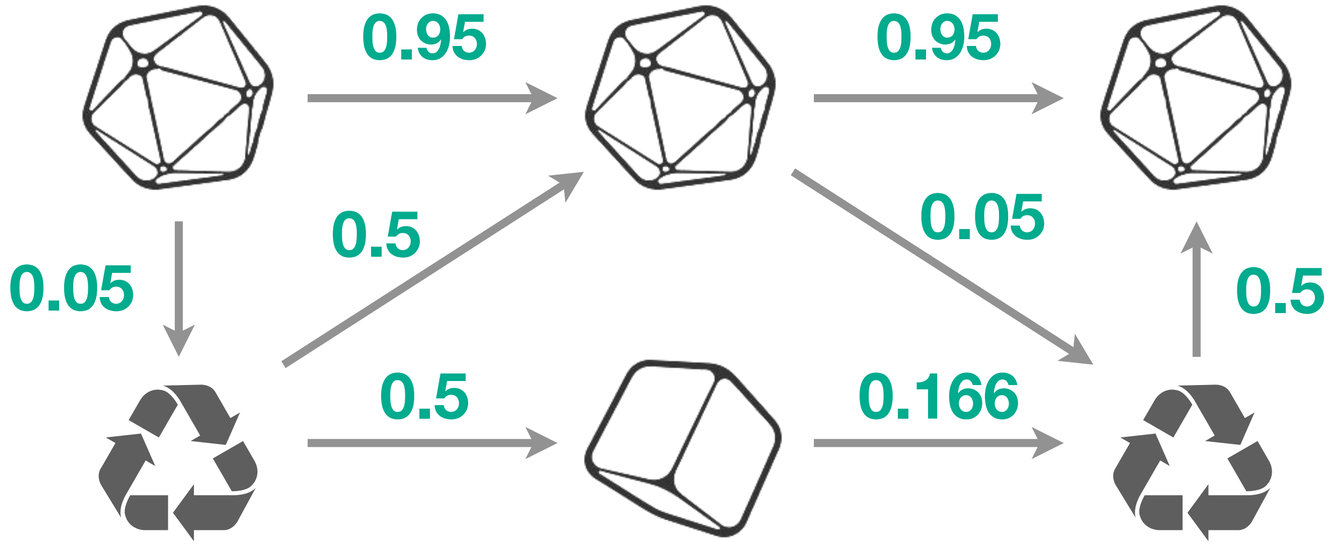
$$l^{(t)}(\boldsymbol{\theta}) = \sum_{\mathbf{Z} \in \mathcal{Z}} Q^{(t)}(\mathbf{Z}) \log P(\mathbf{X}, \mathbf{Z}; \boldsymbol{\theta}^{(t-1)})$$

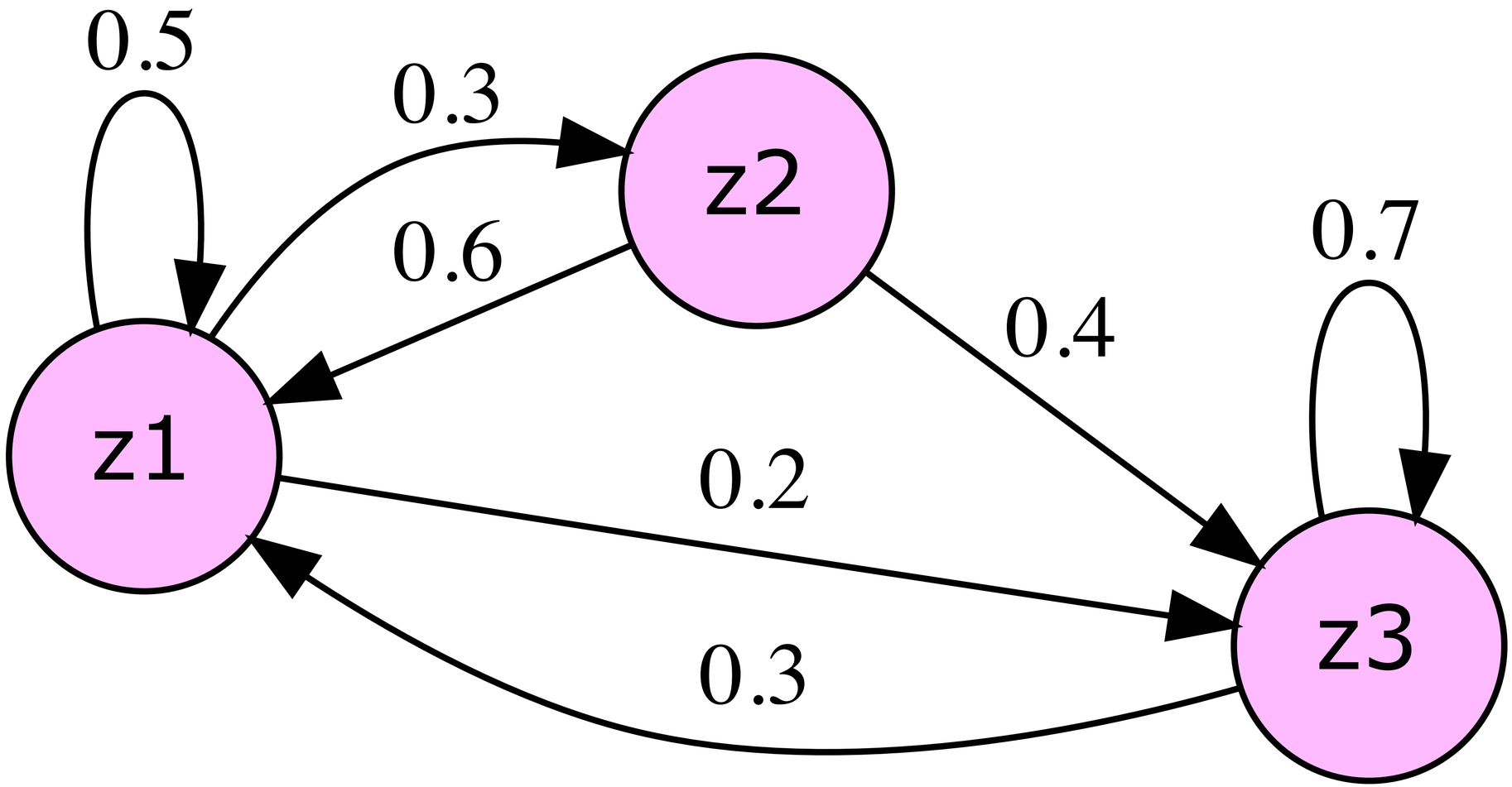
$$\boldsymbol{\theta}^{(t)} = \underset{\theta}{\operatorname{argmax}}\, l^{(t)}(\boldsymbol{\theta})$$

$$\boldsymbol{\theta}^{(0)} = ???$$

Hidden Markov Models

... 13 2 17 ...




0.5
0.3
z2
0.6
0.7
0.4
z1
0.2
z3
0.3

state at $t+1$

state at t	z1	z2	z3
z1	0.5	0.3	0.2
z2	0.6	0	0.4
z3	0.3	0	0.7

state at $t+1$

state at t	d6	d20
d6	0.917	0.083
d20	0.025	0.975

$$a_{i,j} = P(z_{t+1} = j | z_t = i)$$

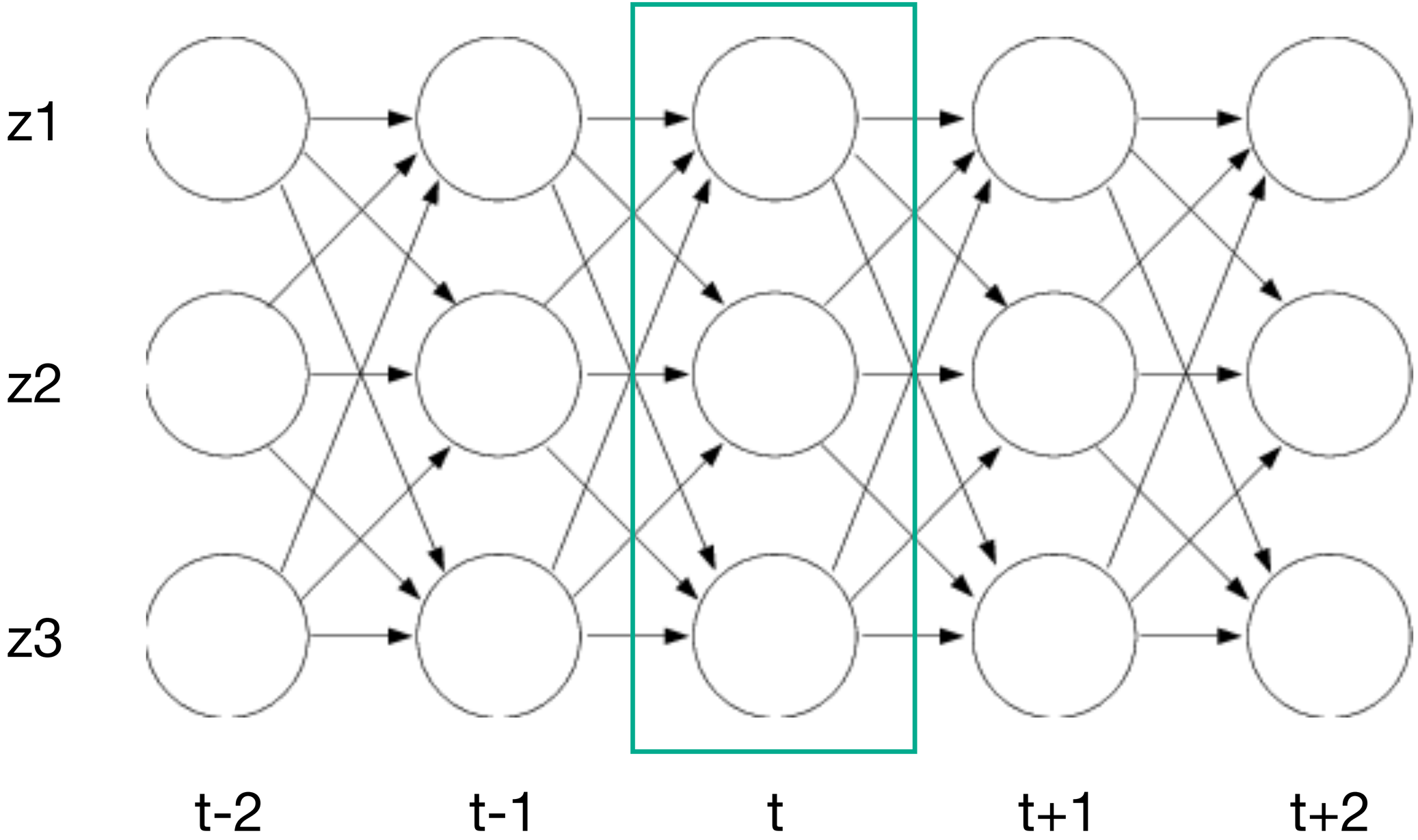
- Observations: $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T \in \mathbb{R}^d$
- Latent states: $z_1, z_2, ..., z_T \in \{1, 2, ..., k\}$
- Emission distribution: $P(\mathbf{x}|z; \boldsymbol{\theta})$
- Initial probabilities: $\boldsymbol{\pi} = [\pi_1, \pi_2, ..., \pi_k]$
- Transition matrix: $\mathbf{A} \in \mathbb{R}^{k \times k}$

- Probability of an observation or sequence
  - → given (**X**, **π**, **A**, **θ**), estimate P(**X**)
- Likelihood of parameters
  - → given (**X**, **π**, **A**, **θ**), estimate L(**π**, **A**, **θ**)
- Parameter fitting
  - → given (maybe lots of) **X**, estimate (**π**, **A**, **θ**)
- Infer hidden state sequence
  - → given (**X**, **π**, **A**, **θ**), estimate (at least some of) **Z**

$$P(\mathbf{x}_{1:t}, z_t) = P(\mathbf{x}_t|z_t) \sum_{z_{t-1}} P(z_t|z_{t-1}) P(\mathbf{x}_{1:t-1}, z_{t-1})$$

$$P(\mathbf{x}_{1:t}, z_t) = \boxed{P(\mathbf{x}_t|z_t)} \sum_{z_{t-1}} P(z_t|z_{t-1}) P(\mathbf{x}_{1:t-1}, z_{t-1})$$

$$P(\mathbf{x}_{1:t}, z_t) = P(\mathbf{x}_t|z_t) \sum_{z_{t-1}} P(z_t|z_{t-1}) P(\mathbf{x}_{1:t-1}, z_{t-1})$$


z1
z2
z3
t-2
t-1
t
t+1
t+2

Forward

$$\alpha_t(j) = P(\mathbf{x}_{1:t}, z_t = j) = \sum_i^k \alpha_{t-1}(i)\, a_{i,j}\, P(\mathbf{x}_t; \boldsymbol{\theta}_j)$$

Forward-Backward

$$\beta_t(j) = P(\mathbf{x}_{t:T}, z_t = j) = \sum_i^k \beta_{t+1}(i)\, a_{i,j}\, P(\mathbf{x}_{t+1}; \boldsymbol{\theta}_j)$$

Viterbi

$$\alpha_t(j) = \max_i \ \alpha_{t-1}(i)\, a_{i,j}\, P(\mathbf{x}_t; \boldsymbol{\theta}_j)$$

$$\zeta_t(j) = \operatorname*{argmax}_i \ \alpha_{t-1}(i)\, a_{i,j}\, P(\mathbf{x}_t; \boldsymbol{\theta}_j)$$

Baum-Welch

Questions?

Next: Deep Learning Applications