Part 3

Markov Chain Modeling

Markov Chain Model

- Stochastic model
- Amounts to sequence of random variables

$$X_1, X_2, ..., X_t$$

- Transitions between states
- State space

$$S = \{s_1, s_2, ..., s_m\}$$

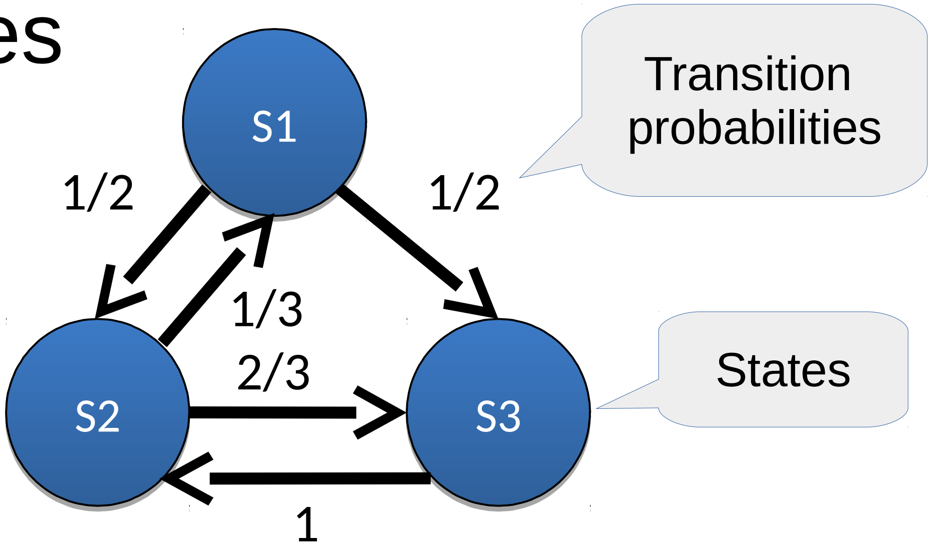
Markov Chain Model

- Stochastic model
- Amounts to sequence of random variables

$$X_1, X_2, ..., X_t$$

- Transitions between states
- State space

$$S = \{s_1, s_2, ..., s_m\}$$



Markovian property

- Next state in a sequence only depends on the current one
- Does not depend on a sequence of preceding ones

$$P(X_{t+1} = s_j | X_1 = s_{i_1}, ..., X_{t-1} = s_{i_{t-1}}, X_t = s_{i_t}) = P(X_{t+1} = s_j | X_t = s_{i_t}) = p_{i,j}$$

Transition matrix

$$\begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,j} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,j} \\ \vdots & \vdots & \ddots & \vdots \\ p_{i,1} & p_{i,2} & \cdots & p_{i,j} \end{bmatrix}$$

Transition matrix P

$$\sum_j p_{ij} = 1$$

Rows sum to 1

$$p_{i,j} = p(s_j | s_i)$$

Single transition probability

Likelihood

- Transition probabilities are parameters

$$D = x_1, x_2, x_3, ..., x_n$$

$$P(D|\theta) = p(x_n|x_{n-1})p(x_{n-1}|x_{n-2})...p(x_2|x_1)p(x_1)$$

$$= p(x_1)\prod_i \prod_j p_{i,j}^{n_{i,j}}$$

Sequence data

MC parameters

Transition count

Transition probability

Maximum Likelihood Estimation (MLE)

- Given some sequence data, how can we determine parameters?
- MLE estimation: count and normalize transitions

$$\mathcal{L}(\mathcal{P}(\mathcal{D}|\theta)) = log\left(p(x_1)\prod_i\prod_j p_{i,j}^{n_{i,j}}\right)$$
$$= logp(x_1) + \sum_i\sum_j n_{i,j}log(p_{i,j})$$

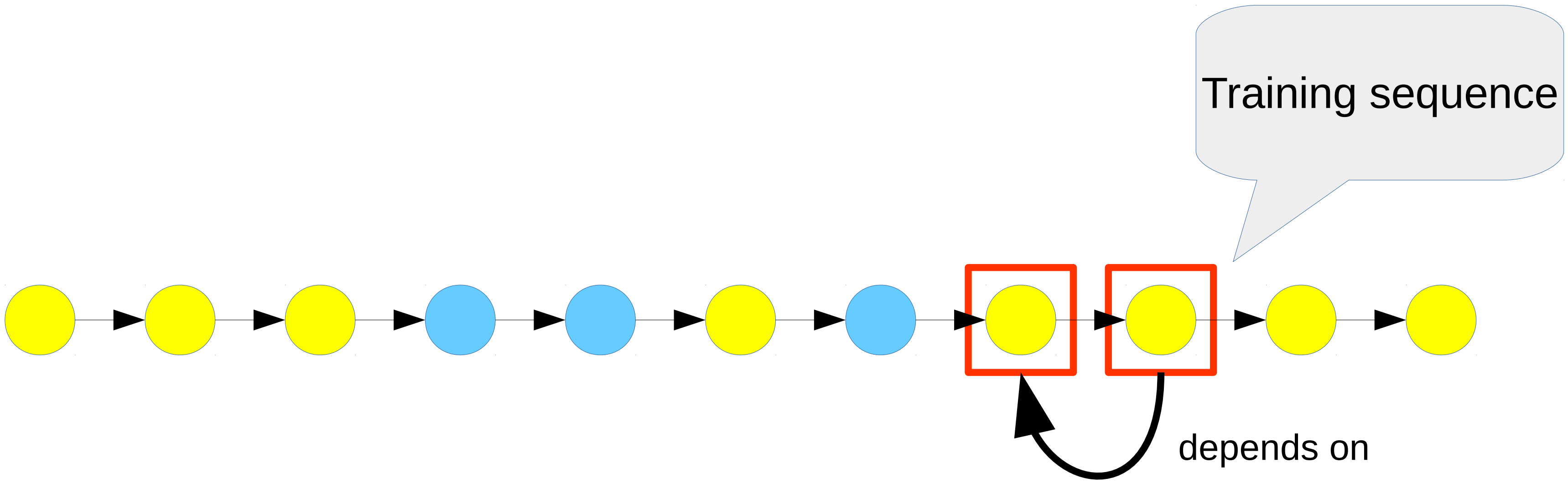
Maximize!

$$p_{i,j} = \frac{n_{i,j}}{\sum_j n_{i,j}}$$

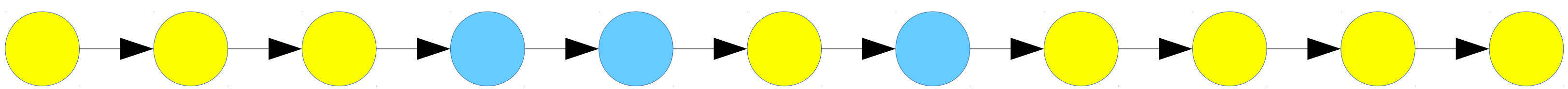
See ref [1]

[Singer et al. 2014]

Example



Example



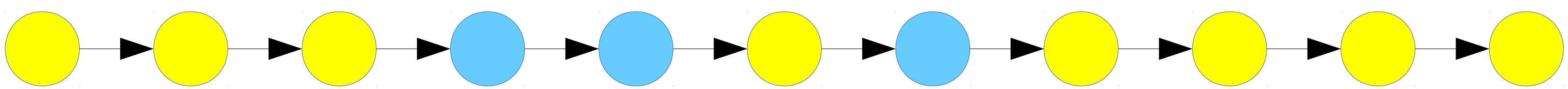
Transition counts

	5	2
	2	1

Transition matrix (MLE)

	5/7	2/7
	2/3	1/3

Example



Transition matrix (MLE)

	5/7	**2/7**
	2/3	**1/3**

Likelihood of given sequence

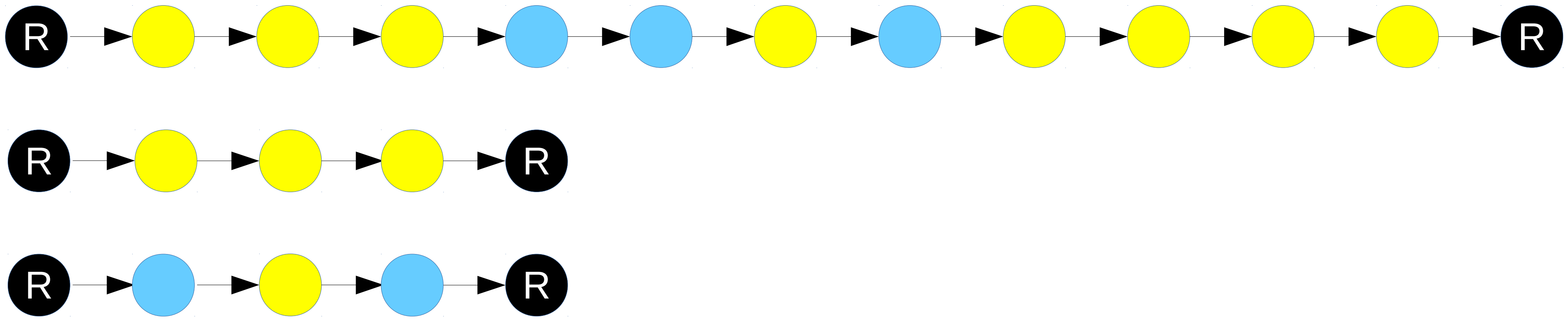
$$(5/7)^5 * (2/7)^2 * (2/3)^2 * (1/3)^1 = 0.002248$$

$$5 * ln(5/7) + 2 * ln(2/7) + 2 * ln(2/3) + 1 * ln(1/3) = -6.0974$$

We calculate the probability of the sequence with the assumption that we start with the yellow state.

Reset state

- Modeling start and end of sequences
- Specifically useful if many individual sequences



[Chierichetti et al. WWW 2012]

Properties

- Reducibility
  - State j is accessible from state i if it can be reached with non-zero probability
  - Irreducible: All states can be reached from any state (possibly multiple steps)
- Periodicity
  - State i has period k if any return to the state is in multiples of k
  - If k=1 then it is said to be aperiodic
- Transcience
  - State i is transient if there is non-zero probability that we will never return to the state
  - State is recurrent if it is not transient
- Ergodicity
  - State i is ergodic if it is aperiodic and positive recurrent
- Steady state
  - Stationary distribution over states
  - Irreducible and all states positive recurrent → one solution
  - Reverting a steady-state [Kumar et al. 2015]

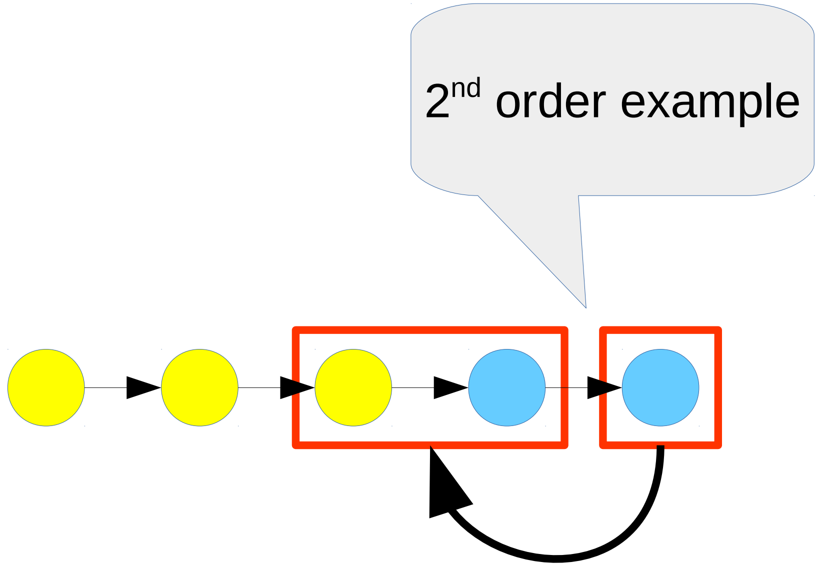
Higher Order Markov Chain Models

- Drop the memoryless assumption?
- Models of increasing order
  - 2nd order MC model
  - 3rd order MC model
  - ...

$$P(X_{t+1} = s_j | X_1 = s_{i_1}, ..., X_{t-1} = s_{i_{t-1}}, X_t = s_{i_t}) =$$
$$P(X_{t+1} = s_j | X_t = s_{i_t}, X_{t-1} = s_{i_{t-1}}, ..., X_{t-k+1} = s_{i_{t-k+1}})$$

Higher Order Markov Chain Models

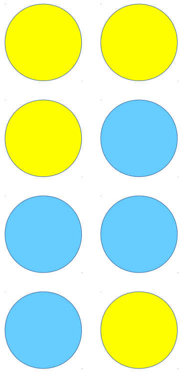
- Drop the memoryless assumption?
- Models of increasing order
  - 2nd order MC model
  - 3rd order MC model
  - ...



$$P(X_{t+1} = s_j | X_1 = s_{i_1}, ..., X_{t-1} = s_{i_{t-1}}, X_t = s_{i_t}) =$$
$$P(X_{t+1} = s_j | X_t = s_{i_t}, X_{t-1} = s_{i_{t-1}}, ..., X_{t-k+1} = s_{i_{t-k+1}})$$

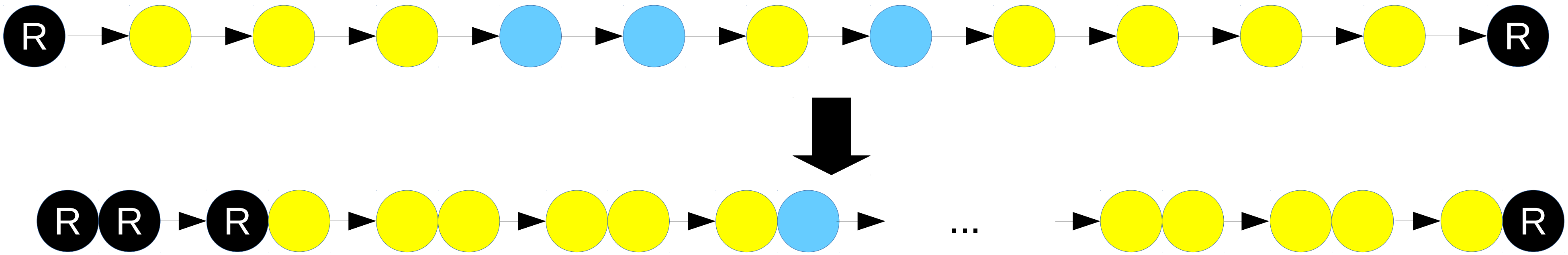
Higher order to first order transformation

- Transform state space
- 2nd order example – new compound states

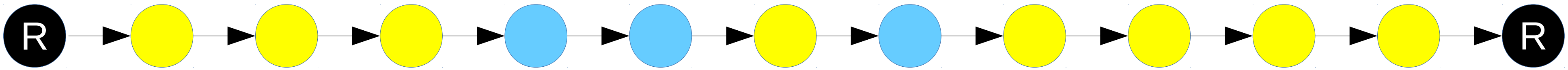


Higher order to first order transformation

- Transform state space
- 2nd order example – new compound states
- Prepend (nr. of order) and append (one) reset states

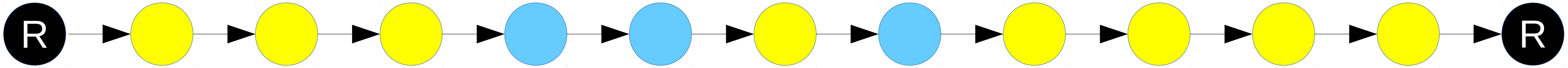


Example



gesis

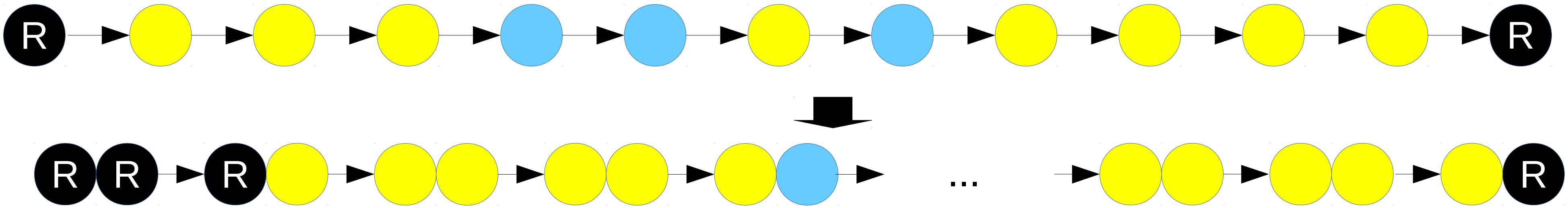
Example



			R
	5/8	2/8	1/8
	2/3	1/3	0/3
R	1/1	0/1	0/1

1st order parameters

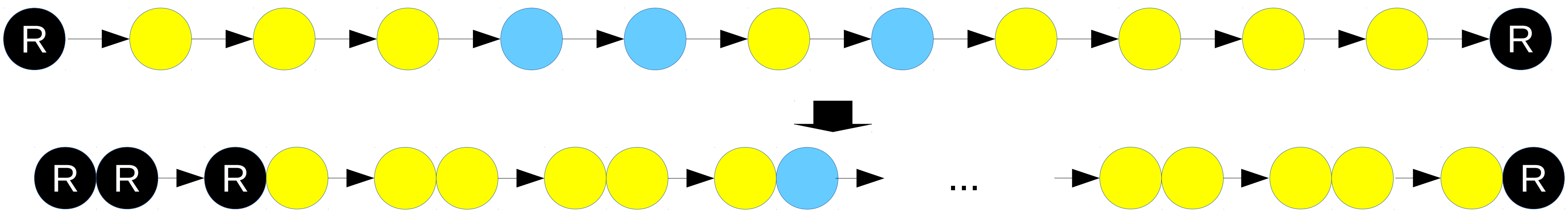
Example



			R
	5/8	2/8	1/8
	2/3	1/3	0/3
R	1/1	0/1	0/1

1st order parameters

Example



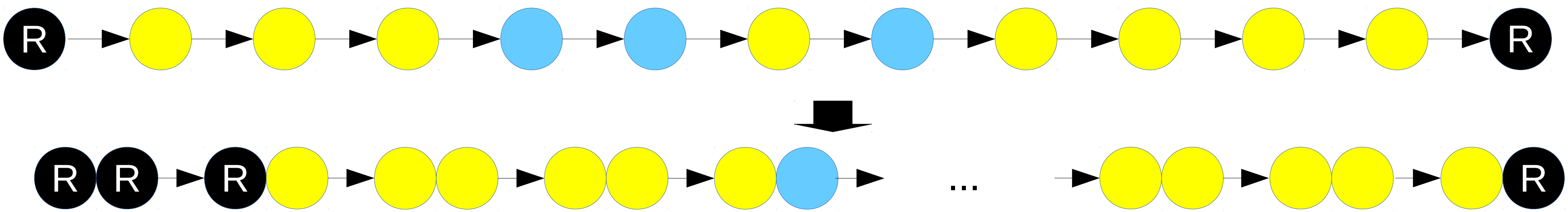
			R
	5/8	2/8	1/8
	2/3	1/3	0/3
R	1/1	0/1	0/1

1st order parameters

			R
	3/5	1/5	1/5
	1/2	1/2	0
	0	1/1	0
	1/2	1/2	0
R R	1/1	0	0
R	1/1	0	0
R	0	0	0
R	0	0	0
R	0	0	0

2nd order parameters

Example



$$ln(P(D|\theta_1) = -9.11$$

			R
	5/8	2/8	1/8
	2/3	1/3	0/3
R	1/1	0/1	0/1

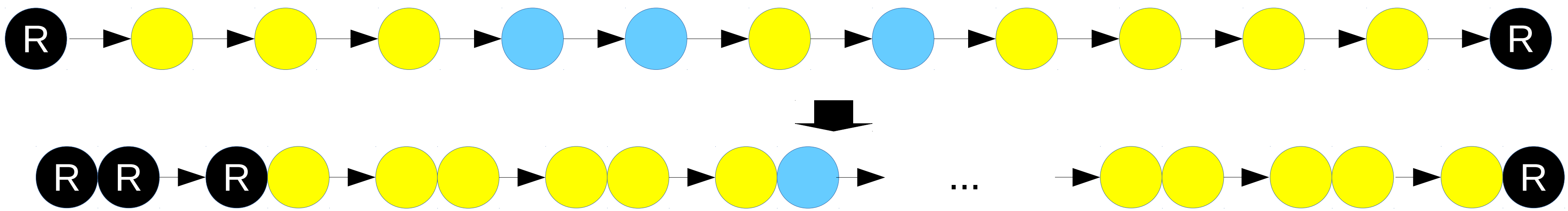
1st order parameters

			R
	3/5	1/5	1/5
	1/2	1/2	0
	0	1/1	0
	1/2	1/2	0
R R	1/1	0	0
R	1/1	0	0
R	0	0	0
R	0	0	0
R	0	0	0

2nd order parameters

$$ln(P(D|\theta_2) = -7.52$$

Example



$$ln(P(D|\theta_1) = -9.11$$

			R
	5/8	2/8	1/8
	2/3	1/3	0/3
R	1/1	0/1	0/1

6 free parameters

			R
	3/5	1/5	1/5
	1/2	1/2	0
	0	1/1	0
	1/2	1/2	0
R R	1/1	0	0
R	1/1	0	0
R	0	0	0
R	0	0	0
R	0	0	0

18 free parameters

$$ln(P(D|\theta_2) = -7.52$$

Model Selection

- Which is the “best” model?
- 1st vs. 2nd order model
- Nested models → higher order always fits better
- Statistical model comparison
- Balance goodness of fit with complexity

Model Selection Criteria

- Likelihood ratio test
  - Ratio between likelihoods for order m and k $_k\eta_m = -2(\mathcal{L}(\mathcal{P}(\mathcal{D}|\theta_k)) - \mathcal{L}(\mathcal{P}(\mathcal{D}|\theta_m)))$
  - Follows chi2 distribution with dof $(|S|^m - |S|^k)(|S| - 1)$
  - Nested models only
- Akaike Information Criterion (AIC)

$$AIC(k) = 2 * (|S|^k)(|S| - 1) - 2(\mathcal{L}(\mathcal{P}(\mathcal{D}|\theta_k))$$

- Bayesian Information Criterion (BIC)

$$BIC(k) = (|S|^k)(|S| - 1) * ln(n) - 2(\mathcal{L}(\mathcal{P}(\mathcal{D}|\theta_k))$$

- **Bayes factors**
- Cross Validation

[Singer et al. 2014], [Strelioff et al. 2007], [Anderson & Goodman 1957]

Bayesian Inference

- Probabilistic statements of parameters
- Prior belief updated with observed data

$$\overbrace{P(\theta|D,M)}^{\text{posterior}} = \frac{\overbrace{P(D|\theta,M)}^{\text{likelihood}}\,\overbrace{P(\theta|M)}^{\text{prior}}}{\underbrace{P(D|M)}_{\text{marginal likelihood}}}$$

Bayesian Model Selection

- Probability theory for choosing between models
- Posterior probability of model M given data D

Evidence

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)}$$

$$\overbrace{P(\theta|D,M)}^{\text{posterior}} = \frac{\overbrace{P(D|\theta,M)}^{\text{likelihood}}\,\overbrace{P(\theta|M)}^{\text{prior}}}{\underbrace{P(D|M)}_{\text{marginal likelihood}}}$$

Evidence

Bayes Factor

- Comparing two models

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)}$$

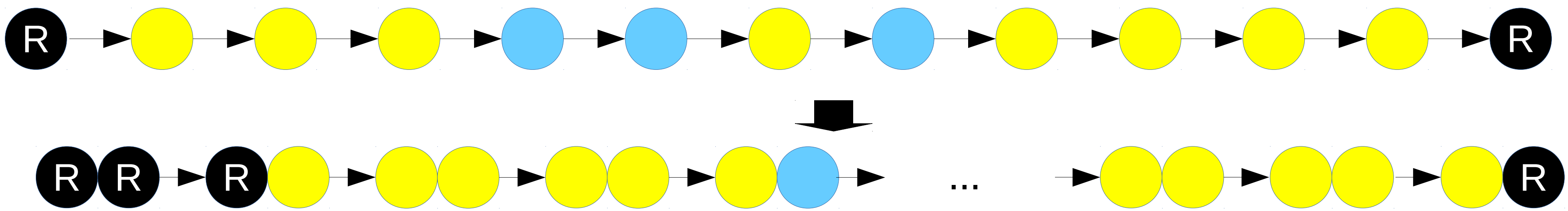
$$\frac{P(D|M_1)}{P(D|M_2)} = \frac{\int P(\theta_1|M_1)P(D|\theta_1, M_1)d\theta}{\int P(\theta_2|M_2)P(D|\theta_2, M_2)d\theta}$$

$$P(D|M) = \prod_i \frac{\Gamma(\sum_j \alpha_{i,j})}{\prod_j \Gamma(\alpha_{i,j})} \frac{\prod_j \Gamma(n_{i,j} + \alpha_{i,j})}{\Gamma(\sum_j (n_{i,j} + \alpha_{i,j}))}$$

- Evidence: Parameters marginalized out
- Automatic penalty for model complexity
- Occam's razor
- Strength of Bayes factor: Interpretation table

[Kass & Raftery 1995]

Example



			R
	5/8	2/8	1/8
	2/3	1/3	0/3
R	1/1	0/1	0/1

$$P(D|M_1) = -13.43$$

			R
	3/5	1/5	1/5
	1/2	1/2	0
	0	1/1	0
	1/2	1/2	0
R R	1/1	0	0
R	1/1	0	0
R	0	0	0
R	0	0	0
R	0	0	0

$$P(D|M_2) = -14.31$$

Hands-on jupyter notebook

Methodological extensions/adaptions

- Variable-order Markov chain models
  - Example: AAABCAAABC
  - Order dependent on context/realization
  - Often huge reduction of parameter space
  - [Rissanen 1983, Bühlmann & Wyner 1999, Chierichetti et al. WWW 2012]
- Hidden Markov Model [Rabiner1989, Blunsom 2004]
- Markov Random Field [Li 2009]
- MCMC [Gilks 2005]

Some applications

- Sequence of letters [Markov 1912, Hayes 2013]
- Weather data [Gabriel & Neumann 1962]
- Computer performance evaluation [Scherr 1967]
- Speech recognition [Rabiner 1989]
- Gene, DNA sequences [Salzberg et al. 1998]
- Web navigation, PageRank [Page et al. 1999]

What have we learned?

- Markov chain models
- Higher-order Markov chain models
- Model selection techniques: Bayes factors

Questions?

References 1/2

[Singer et al. 2014] Singer, P., Helic, D., Taraghi, B., & Strohmaier, M. (2014). Detecting memory and structure in human navigation patterns using markov chain models of varying order. PloS one, 9(7), e102070.

[Chierichetti et al. WWW 2012] Chierichetti, F., Kumar, R., Raghavan, P., & Sarlos, T. (2012, April). Are web users really markovian?. In Proceedings of the 21st international conference on World Wide Web (pp. 609-618). ACM.

[Strelioff et al. 2007] Strelioff, C. C., Crutchfield, J. P., & Hübler, A. W. (2007). Inferring markov chains: Bayesian estimation, model comparison, entropy rate, and out-of-class modeling. Physical Review E, 76(1), 011106.

[Andersoon & Goodman 1957] Anderson, T. W., & Goodman, L. A. (1957). Statistical inference about Markov chains. The Annals of Mathematical Statistics, 89-110.

[Kass & Raftery 1995] Kass, R. E., & Raftery, A. E. (1995). Bayes factors. Journal of the american statistical association, 90(430), 773-795.

[Rissanen 1983] Rissanen, J. (1983). A universal data compression system. IEEE Transactions on information theory, 29(5), 656-664.

[Bühlmann & Wyner 1999] Bühlmann, P., & Wyner, A. J. (1999). Variable length Markov chains. The Annals of Statistics, 27(2), 480-513.

[Gabriel & Neumann 1962] Gabriel, K. R., & Neumann, J. (1962). A Markov chain model for daily rainfall occurrence at Tel Aviv. Quarterly Journal of the Royal Meteorological Society, 88(375), 90-95.

gesis

References 2/2

[Blunsom 2004] Blunsom, P. (2004). Hidden markov models. Lecture notes, August, 15, 18-19.

[Li 2009] Li, S. Z. (2009). Markov random field modeling in image analysis. Springer Science & Business Media.

[Gilks 2005] Gilks, W. R. (2005). Markov chain monte carlo. John Wiley & Sons, Ltd.

[Page et al. 1999] Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: bringing order to the web.

[Rabiner 1989] Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2), 257-286.

[Markov 1912] Markov, A. A. (1912). Wahrscheinlichkeits-rechnung. Рипол Классик.

[Salzberg et al. 1998] Salzberg, S. L., Delcher, A. L., Kasif, S., & White, O. (1998). Microbial gene identification using interpolated Markov models. Nucleic acids research, 26(2), 544-548.

[Scherr 1967] Scherr, A. L. (1967). An analysis of time-shared computer systems (Vol. 71, pp. 383-387). Cambridge (Mass.): MIT Press.

[Kumar et al. 2015] Kumar, R., Tomkins, A., Vassilvitskii, S., & Vee, E. (2015. Inverting a Steady-State. In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (pp. 359-368). ACM.

[Hayes 2013] Hayes, B. (2013). First links in the Markov chain. American Scientist, 101(2), 92-97.