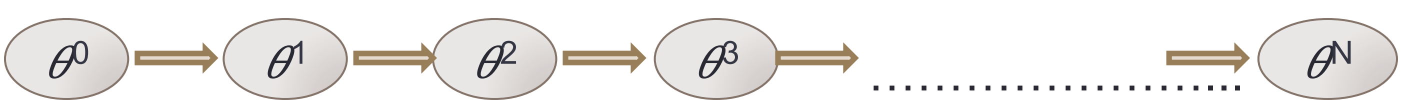
MCMC SAMPLING

Sudhir Raman
Translational Neuromodeling Unit,
UZH & ETH



Recap

Recap - Bayesian modelling

⇨ Formulation of a generative **model**

Model ⇒ Likelihood function $p(y/\theta)$

Model ⇒ prior distribution $p(\theta)$

⇨ Observation of **data**

Data ⇒ y

⇨ **Model Inversion - Update** of beliefs based upon observations, given a prior state of knowledge

$$P(\theta \mid y) = \frac{p(y \mid \theta)\ p(\theta)}{p(y)}$$

Recap - Bayesian modelling

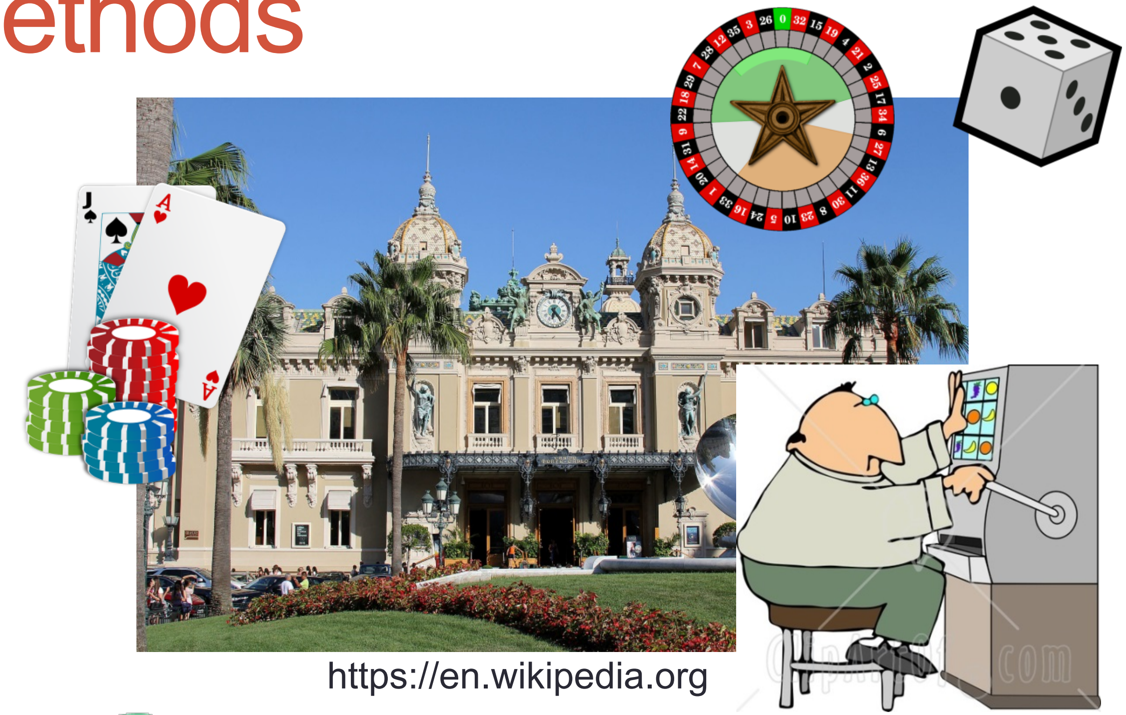
$$p(\theta \mid y) \propto p(y \mid \theta) p(\theta)$$

- Maximum a posteriori (MAP)
- Posterior Moments
- Model Evidence

Deterministic approximation	Stochastic approximation
• Variational Bayes	• Sampling Methods

Monte Carlo Methods

Monte Carlo methods



https://en.wikipedia.org

Random Numbers

⇨ Expectation

⇨ Density

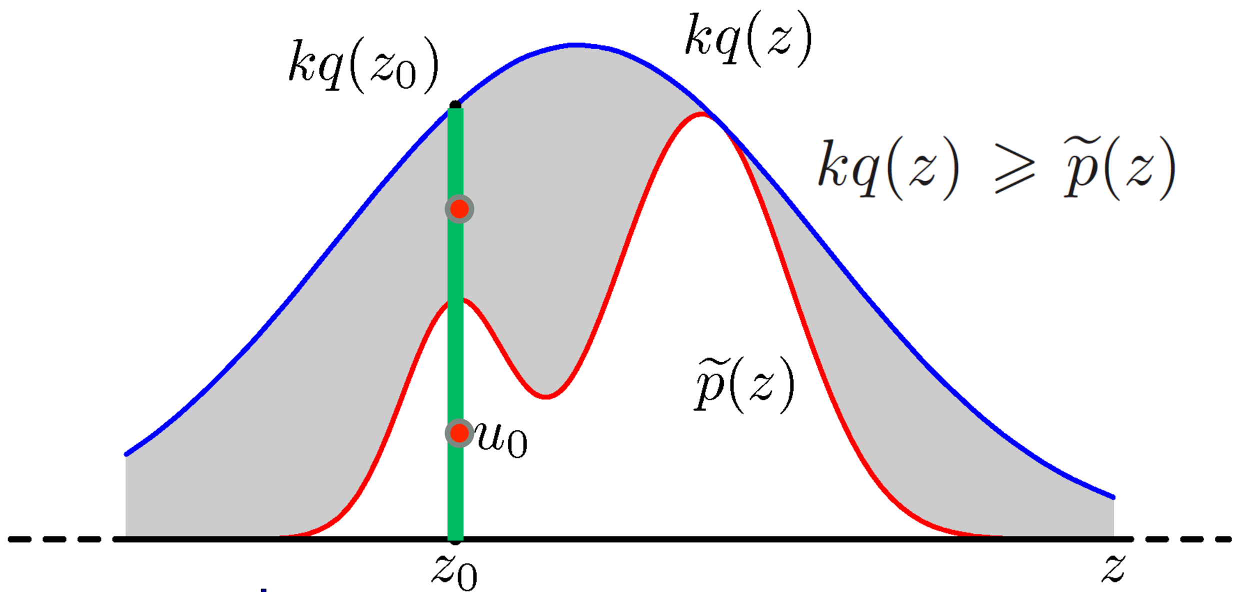
$$\int f(\theta)p(\theta)d\theta$$

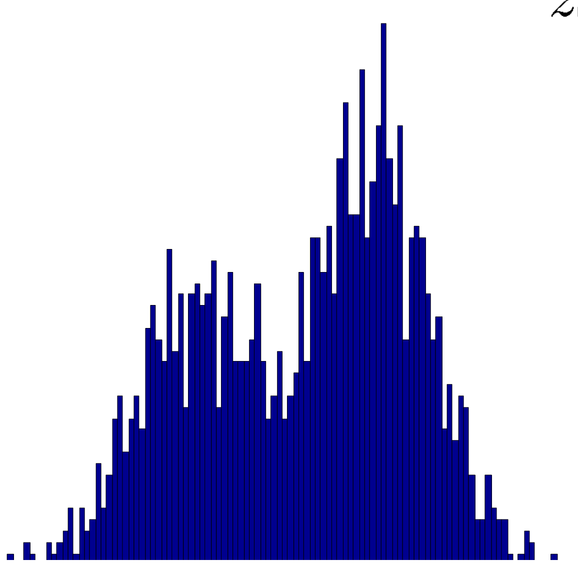
Simple Monte Carlo

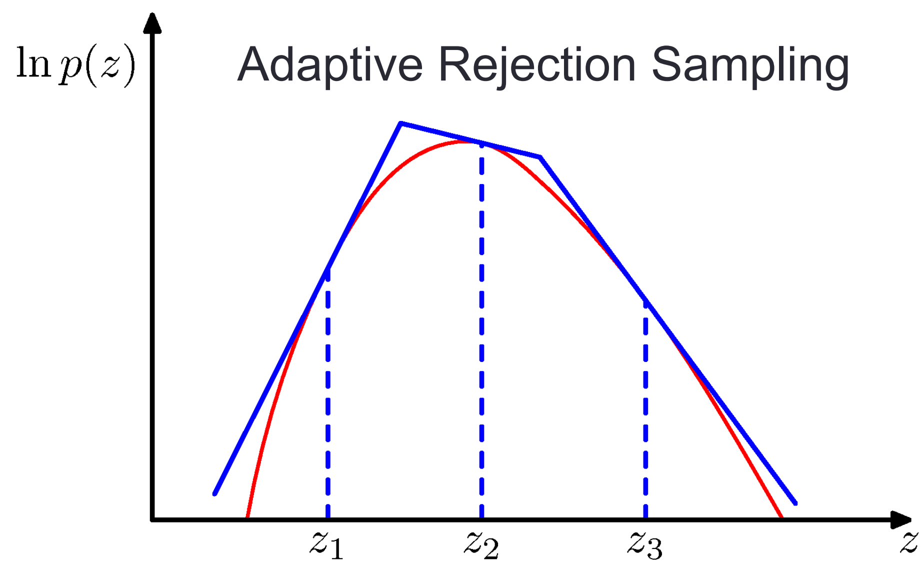
- What is the average height of people in this room ?

$$\frac{1}{N}\sum_{i=0}^{N} f(z_i)$$

Rejection Sampling



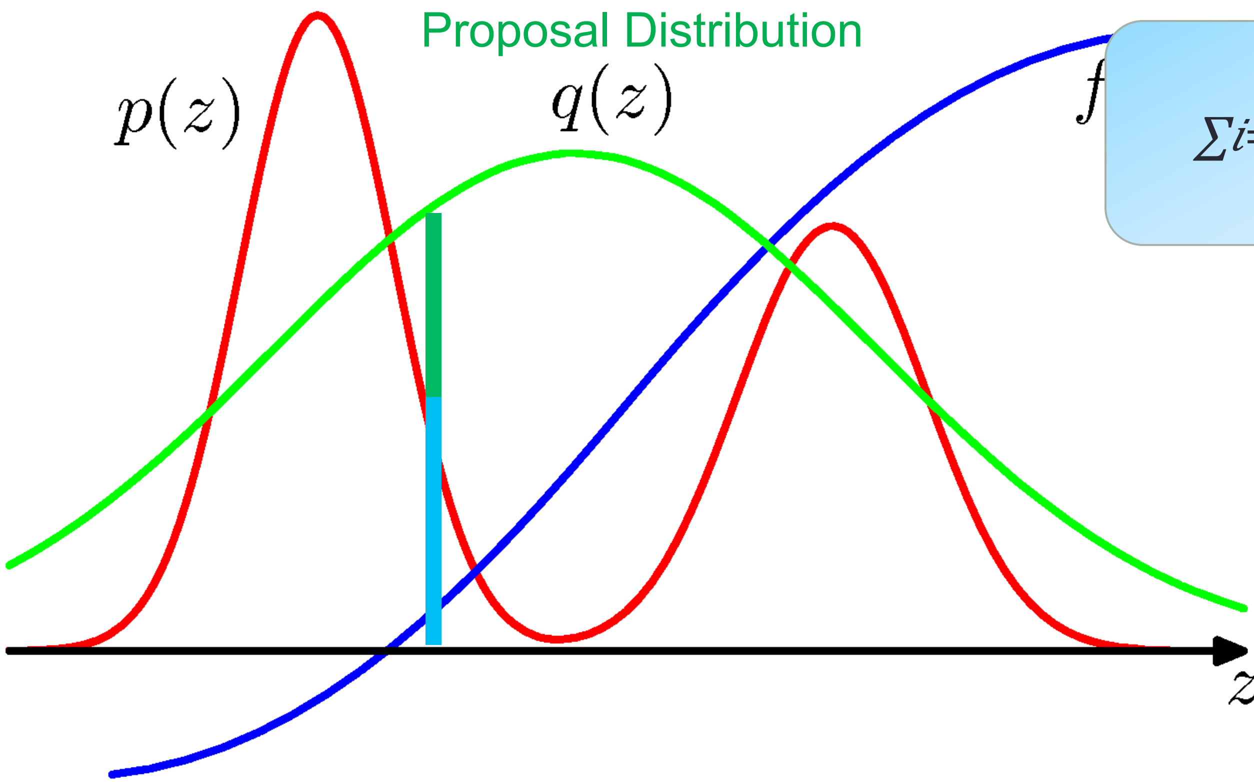




Importance Sampling

$$\int f(z)p(z)dz$$

$$1/N \sum_{i=0}^{N} f(z_i)$$

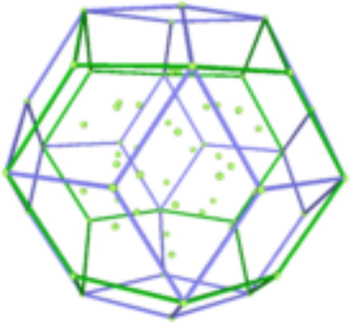


$$\mathbb{E}[f] \simeq \sum_{l=1}^{L} w_l f(\mathbf{z}^{(l)})$$

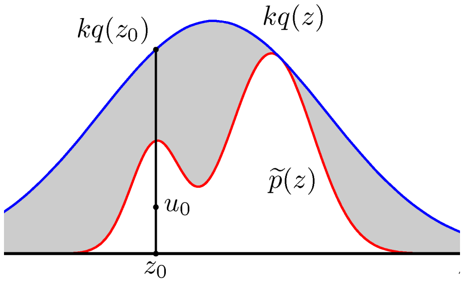
$$w_l = \frac{\widetilde{r}_l}{\sum_m \widetilde{r}_m} = \frac{\widetilde{p}(\mathbf{z}^{(l)})/q(\mathbf{z}^{(l)})}{\sum_m \widetilde{p}(\mathbf{z}^{(m)})/q(\mathbf{z}^{(m)})}.$$

Issues

Scaling to high dimensions



Choice of proposal distribution



Markov Chain Monte Carlo

Markov Chain Monte Carlo(MCMC) sampling

Andrei Markov
(1856 – 1922)
Russian mathematician

Proposal distribution

$q(\theta)$♪

$\theta^0 \Rightarrow \theta^1 \Rightarrow \theta^2 \Rightarrow \theta^3 \Rightarrow \cdots\cdots \Rightarrow \theta^N$

Transition

Posterior distribution

$p\theta y$

- A general framework for sampling from a large class of distributions
- Scales well with dimensionality of sample space
- Asymptotically convergent

Markov chain properties

- Transition probabilities – homogeneous

$$p(\theta^{t+1} \mid \theta^1,\dots,\theta^t) = p(\theta^{t+1} \mid \theta^t) = T_t(\theta^t, \theta^{t+1})$$

- Invariance

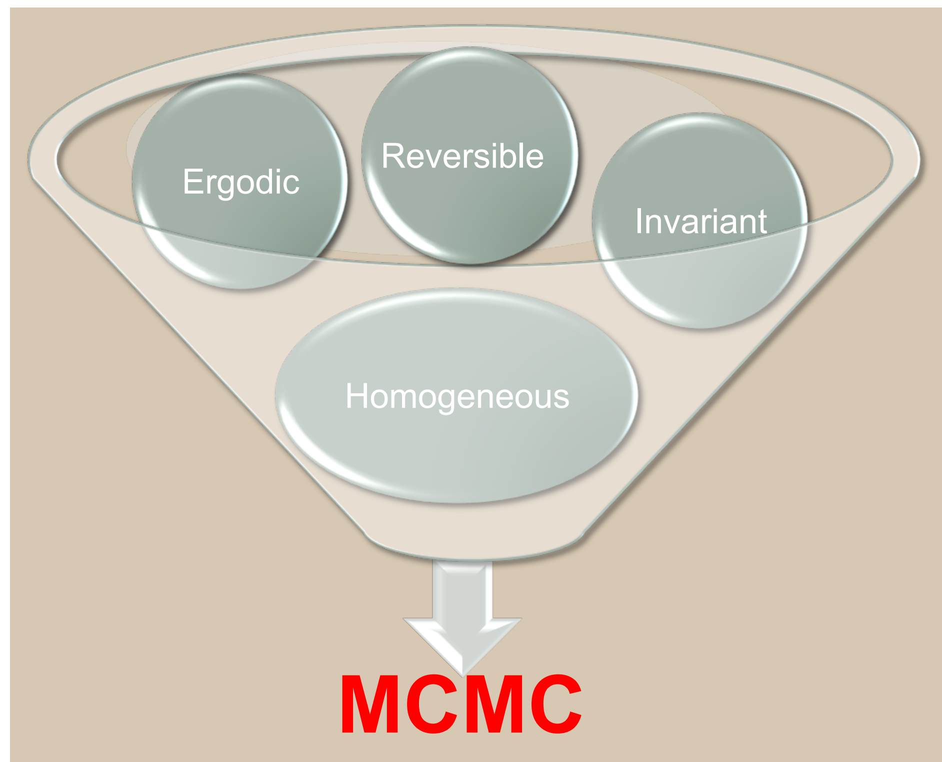
$$p^*(\theta) = \sum_{\theta'} T(\theta', \theta)\, p^*(\theta')$$

- Detailed Balance

$$T(\theta, \theta')\, p^*(\theta) = T(\theta', \theta)\, p^*(\theta')$$

- Ergodicity

$$p^*(\theta) = \lim_{n \to \infty} \left(p(\theta^n)\right) \quad \forall\, p(\theta^0)$$



Neal (1993b), Gilks et al. (1996), and Tanner (1996).

Metropolis-Hastings Algorithm

$$\boldsymbol{p}(\boldsymbol{\theta}|\boldsymbol{y})$$

- Initialize θ at step 1 - for example, sample from prior
- At step t, sample from the proposal distribution:

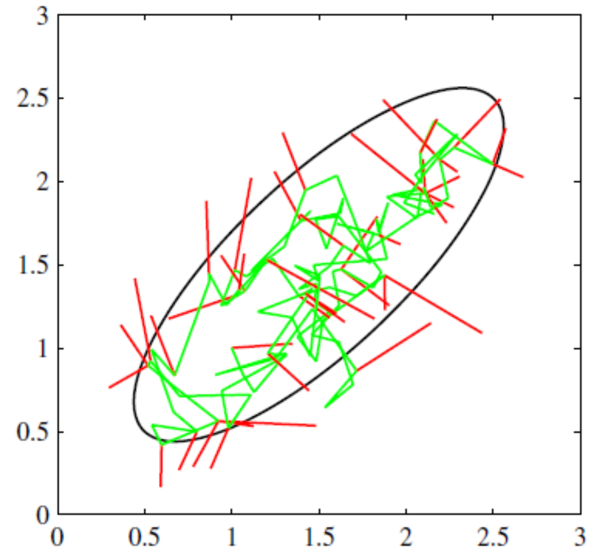
$$\theta^* \sim q(\theta^* | \theta^t)$$

- Accept with probability:

$$A(\theta^*, \theta^t) \sim min(1, p(\theta^* | y)\, q(\theta^t | \theta^*) / p(\theta^t | y)\, q(\theta^* | \theta^t))$$

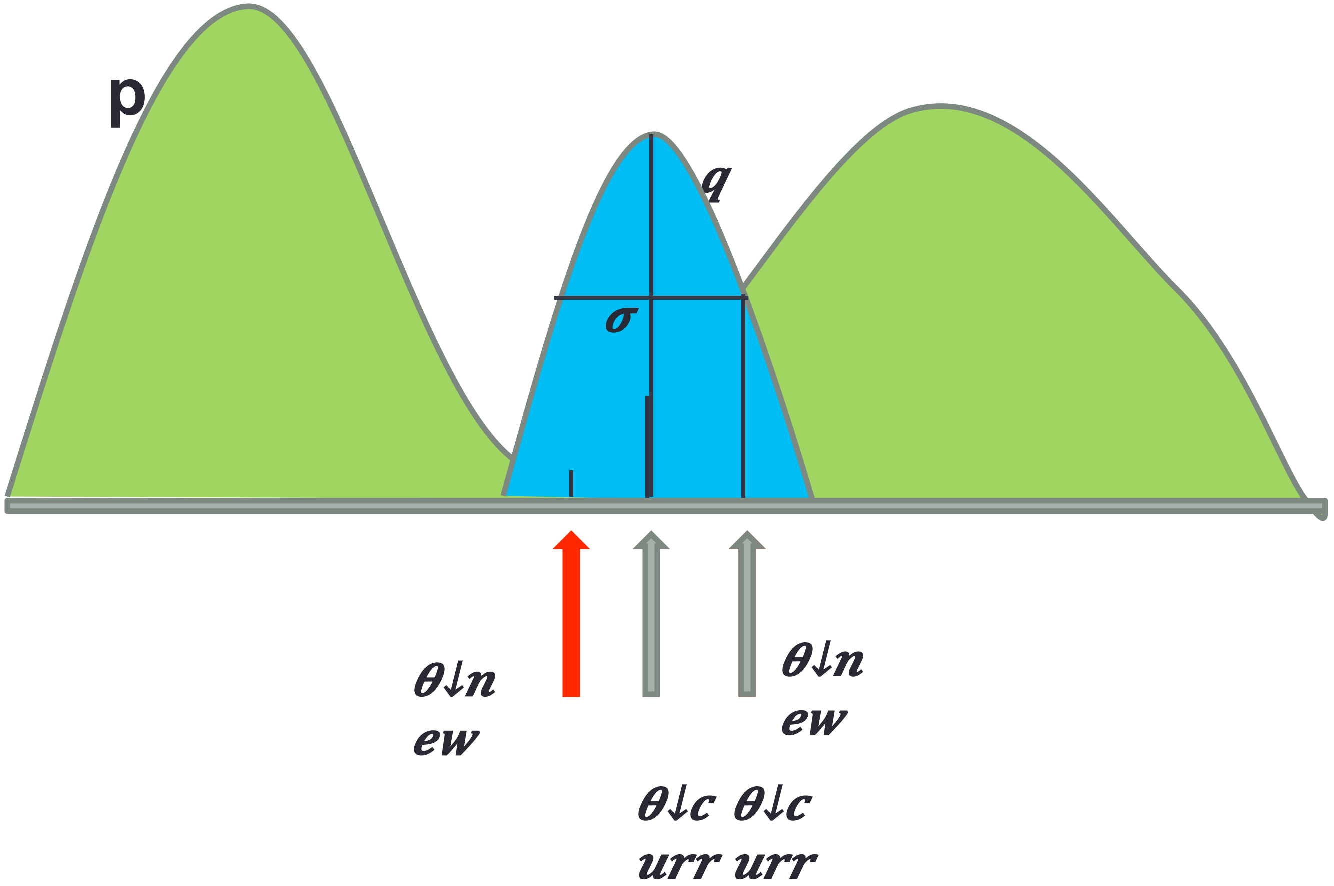
- Metropolis – Symmetric proposal distribution

$$A(\theta^*, \theta^t) \sim min(1, p(\theta^* | y) / p(\theta^t | y))$$



Bishop (2006) PRML, p. 539

Visual Example



Gibbs Sampling Algorithm

$$p(\boldsymbol{\theta}) = p(\theta_1, \theta_2, \ldots, \theta_n)$$

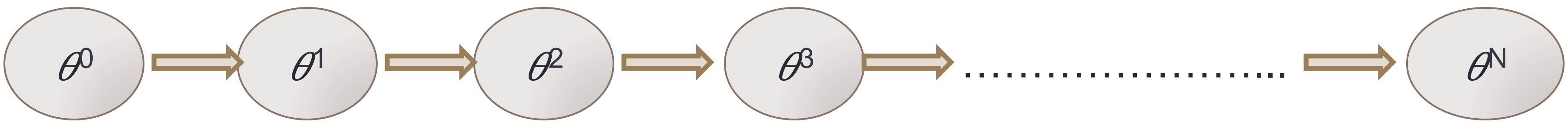
- Special case of Metropolis Hastings
- At step t, sample from the conditional distribution:

$$
\begin{aligned}
\theta_1^{t+1} &\sim p(\theta_1 | \theta_2^t, \ldots\ldots, \theta_n^t) \\
\theta_2^{t+1} &\sim p(\theta_2 | \theta_1^{t+1}, \ldots\ldots, \theta_n^t) \\
&\vdots \\
\theta_n^{t+1} &\sim p(\theta_3 | \theta_1^{t+1}, \ldots\ldots, \theta_{n-1}^{t+1})
\end{aligned}
$$

- Acceptance probability = 1
- Blocked Sampling

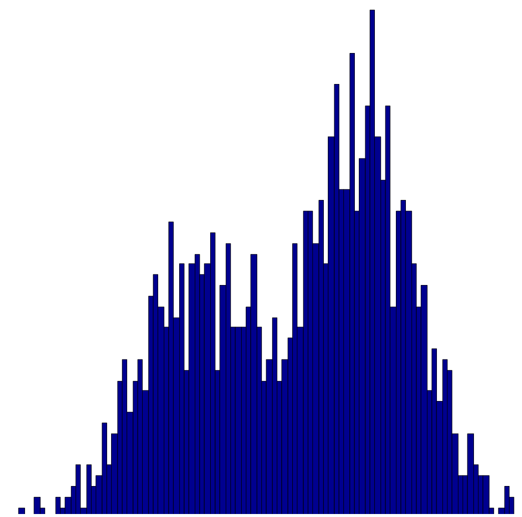
$$\theta_1^{t+1}, \theta_2^{t+1} \sim p(\theta_1, \theta_2 | \theta_3^t, \ldots\ldots, \theta_n^t)$$

Posterior analysis from MCMC



Obtain independent samples:

- Generate samples based on MCMC sampling.
- Discard initial “burn-in” period samples to remove dependence on initialization.
- Thinning- select every mth sample to reduce correlation .
- Inspect sample statistics (e.g., histogram, sample quantiles, …)

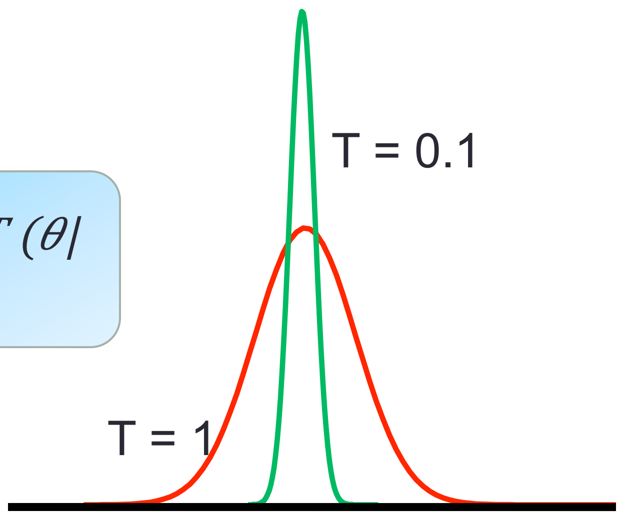


MAP estimate via Simulated Annealing

- Add a temperature parameter and schedule to update it

$$p^{1/T}(\theta|y)$$

- Algorithm
  - Set T = 1
  - Until convergence
    - For every K iterations sample from:
    - Reduce T



Convergence Analysis

- Single chain methods
  - Geweke (1992)
  - Raftery-Lewis (1992)

$$\theta^0 \Rightarrow \theta^1 \Rightarrow \theta^2 \Rightarrow \theta^3 \Rightarrow \cdots\cdots \Rightarrow \theta^N$$

- Multi-chain methods
  - Gelman-Rubin – (1992)
    - Potential Scale Reduction factor

$$\theta^0 \Rightarrow \theta^1 \Rightarrow \theta^2 \Rightarrow \theta^3 \Rightarrow \cdots\cdots \Rightarrow \theta^N$$

$$\theta^0 \Rightarrow \theta^1 \Rightarrow \theta^2 \Rightarrow \theta^3 \Rightarrow \cdots\cdots \Rightarrow \theta^N$$

$$\theta^0 \Rightarrow \theta^1 \Rightarrow \theta^2 \Rightarrow \theta^3 \Rightarrow \cdots\cdots \Rightarrow \theta^N$$

Model Evidence

Model evidence using MCMC

- Importance Sampling

$$p(D \mid M) = \frac{E_g\left[\frac{p(D|\theta,M)p(\theta|M)}{g(\theta)}\right]}{E_g\left[\frac{p(\theta|M)}{g(\theta)}\right]},$$

- Prior arithmetic mean

$$\widehat{f(Y)} = \frac{1}{M}\sum_{m=1}^{M} p(Y|\theta_m)$$

- Posterior harmonic mean

$$\widehat{f(Y)} = \frac{1}{\frac{1}{M}\sum_{m=1}^{M}\frac{1}{p(\mathbf{Y}|\bullet)}},$$

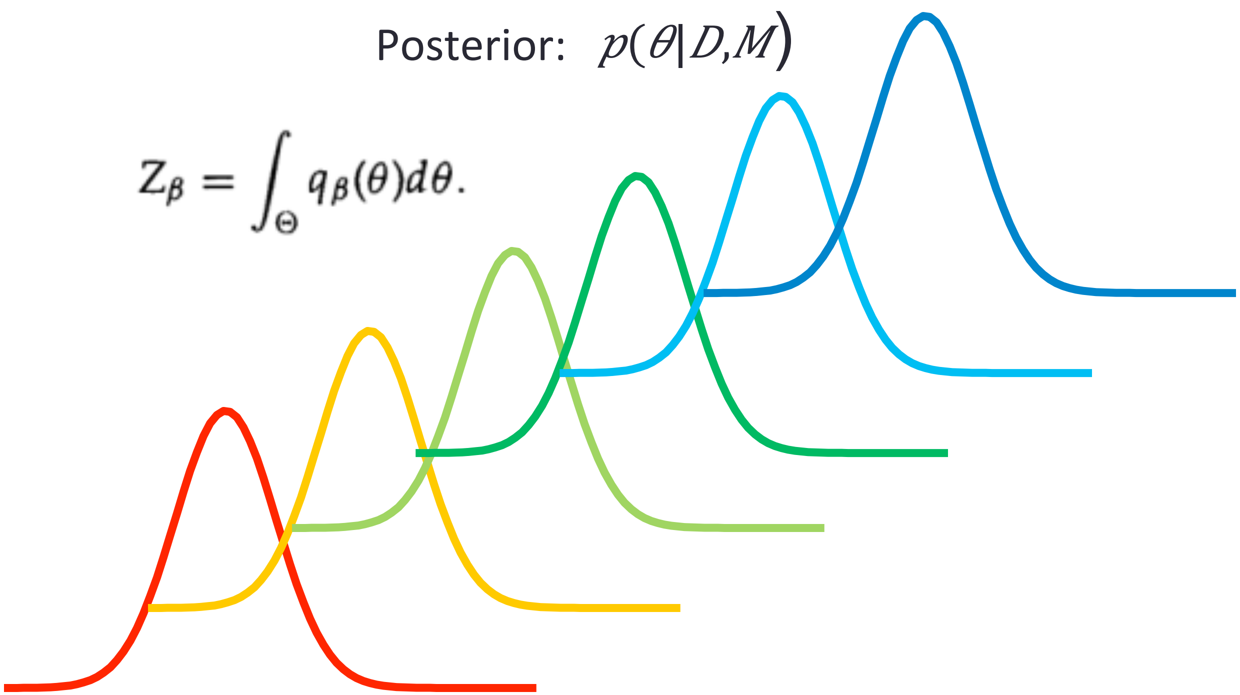
Thermodynamic Integration

- Use multiple distributions

$$q_\beta(\theta) = p(D \mid \theta, M)^\beta \, p(\theta \mid M).$$

Posterior: $p(\theta|D,M)$

$$Z_\beta = \int_\Theta q_\beta(\theta) d\theta.$$



Prior: $p(\theta|M)$

$$p_\beta(\theta) = \frac{1}{Z_\beta} q_\beta(\theta), \tag{15}$$

$$Z_\beta = \int_\Theta q_\beta(\theta) d\theta. \tag{16}$$

When β tends to 0 (resp. 1), p_β converges pointwise to p_0 (resp. p_1), and Z_β to Z_0 (resp. Z_1).
Taking the derivative of $\ln Z_\beta$ with respect to β:

$$\frac{\partial \ln Z_\beta}{\partial \beta} = \frac{1}{Z_\beta} \frac{\partial Z_\beta}{\partial \beta} \tag{17}$$

$$= \frac{1}{Z_\beta} \frac{\partial}{\partial \beta} \int_\Theta q_\beta(\theta) d\theta \tag{18}$$

$$= \frac{1}{Z_\beta} \int_\Theta \frac{\partial q_\beta(\theta)}{\partial \beta} d\theta \tag{19}$$

$$= \int_\Theta \frac{1}{q_\beta(\theta)} \frac{\partial q_\beta(\theta)}{\partial \beta} \frac{q_\beta(\theta)}{Z_\beta} d\theta \tag{20}$$

$$= \int_\Theta \frac{\partial \ln q_\beta(\theta)}{\partial \beta} p_\beta(\theta) d\theta \tag{21}$$

$$= E_\beta \left[\frac{\partial \ln q_\beta(\theta)}{\partial \beta} \right], \tag{22}$$

p_β. Defining the *potential*

$$U(\theta) = \frac{\partial \ln q_\beta(\theta)}{\partial \beta}, \tag{23}$$

one has thus the first moment identity:

$$\frac{\partial \ln Z_\beta}{\partial \beta} = E_\beta[U]. \tag{24}$$

Integrating over [0, 1] yields the log-ratio one is looking for:

$$\mu = \ln Z_1 - \ln Z_0 \tag{25}$$

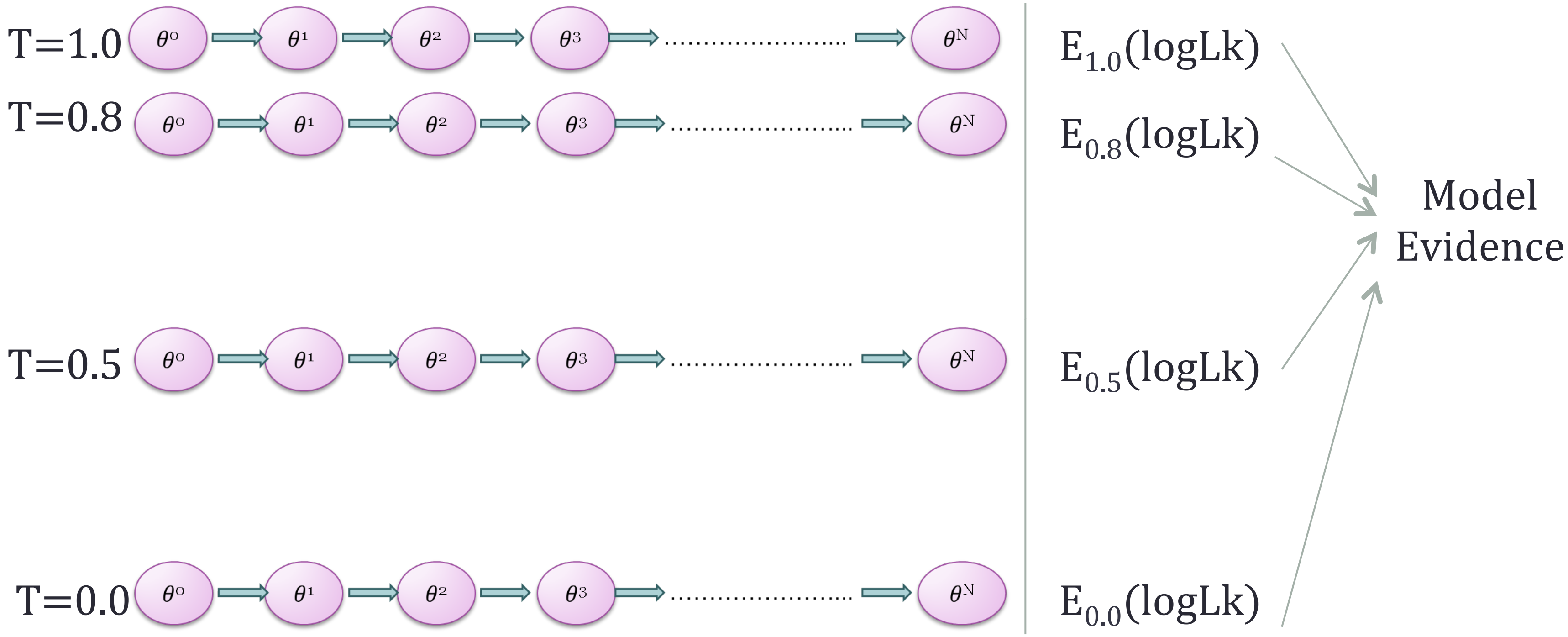
$$= \int_0^1 \frac{\partial \ln Z_\beta}{\partial \beta} d\beta \tag{26}$$

$$= \int_0^1 E_\beta[U] d\beta. \tag{27}$$

Thermodynamic Integration

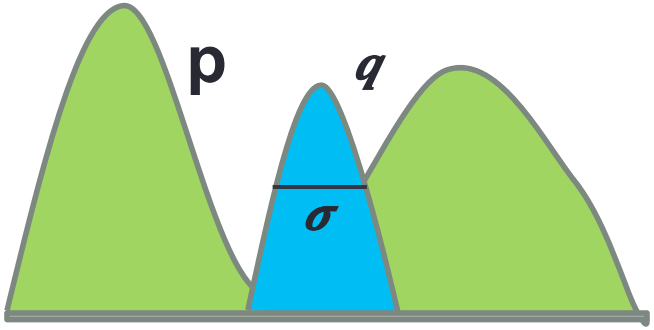
- Path Sampling (Thermodynamic Integration)

$$q_\beta(\theta) = p(D \mid \theta, M)^\beta p(\theta \mid M).$$

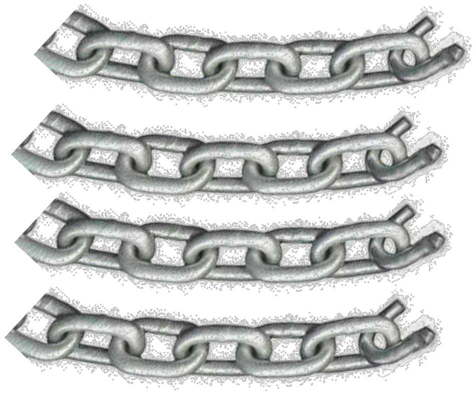


Ogata ,Y. 1989. Num.Math.55:137-157,
Gelman,A.1998. . Stat . Sci . 13:163 — 185

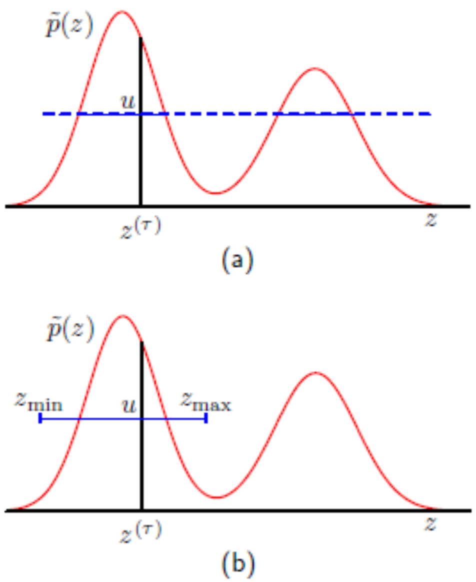
Other MCMC variants



Adaptive MCMC

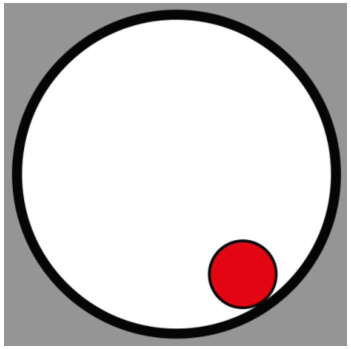


Population MCMC

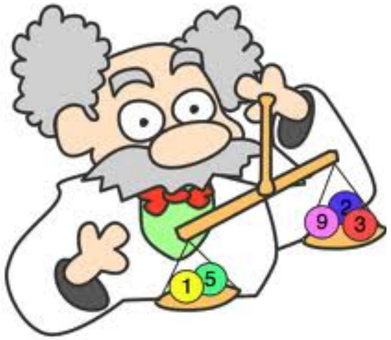


Slice Sampling

Hybrid Monte Carlo

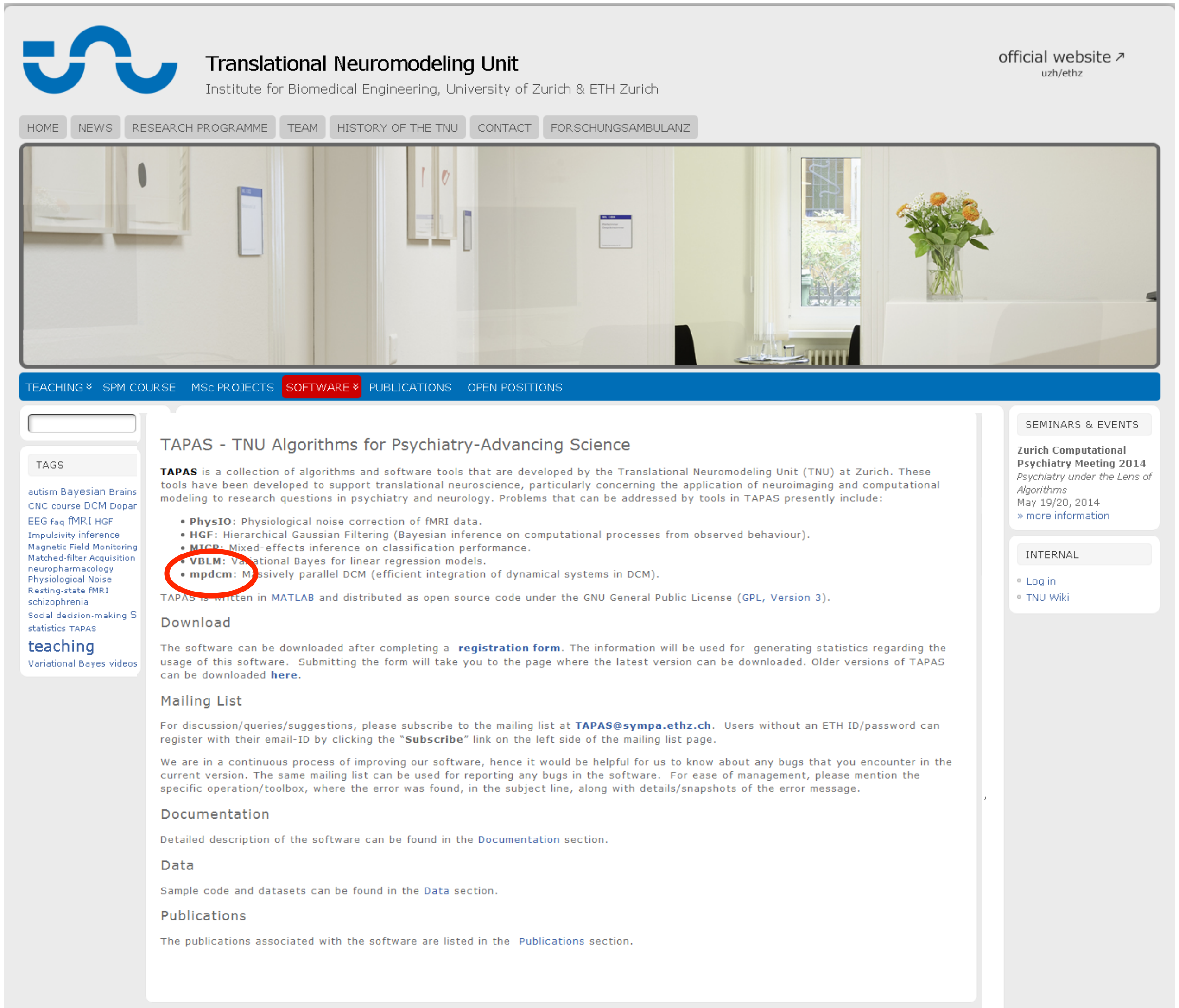


Comparison



Criteria	Variational Bayes	MCMC Sampling
Method	Find a proxy posterior using a hypothesis class	Collect samples
Speed		
Model Evidence	Free	Computationally Expensive
Convergence	Local minima	Convergence diagnostics required
Update Equations		

TAPAS



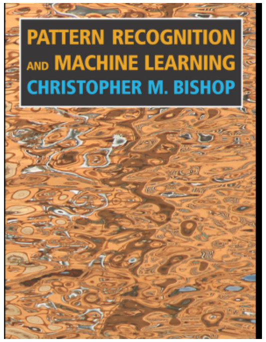
- mpdcm – GPU based Population MCMC for DCM
  - http://www.translationalneuromodeling.org/tapas/

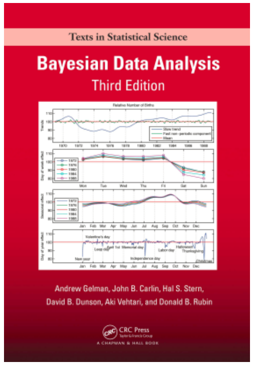
Conclusion

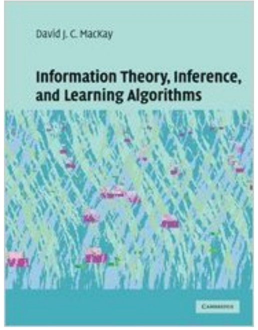
- Simple, Asymptotically Exact
- Computationally Expensive
- Profiling, Parallelizing
- Testing, Testing, Testing, Testing

References

- Pattern Recognition and Machine Learning (2006). Christopher M. Bishop
- Information theory, pattern recognition and learning algorithms. David McKay
- Videolectures.net – Bayesian inference, MCMC , Variational Bayes
- Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Second Edition (Chapman & Hall/CRC Texts in Statistical Science) Dani Gamerman, Hedibert F. Lopes.
- N. Lartillot and H. Philippe, "Computing bayes factors using thermodynamic integration," Systematic Biology, vol. 55, no. 2, pp. 195–207, 2006.
- Andrew Gelman, Donald B. Rubin. Inference from iterative simulation using multiple sequences Stat. Sci., 7 (4) (1992), pp. 457–472
- J. Geweke. Evaluating the accuracy of sampling-based approaches to calculating posterior moments Bayesian Statistics 4, (1992), pp. 169–193
- Adrian E. Raftery, Steven Lewis. How many iterations in the Gibbs sampler? Bayesian Statistics 4, Oxford University Press (1992), pp. 763–773







Thank You

http://www.translationalneuromodeling.org/tapas/