Multimodal MCMC with Replica Exchange and TensorFlow Probability

Simeon Carstens

www.tweag.io

Tweag I/O

Software engineering lab based in Paris with employees all around the world.

We specialize in

- software engineering, with a focus on functional programming
- DevOps, with a focus on reproducible software systems and builds
- data science – from the first model to production deployment

Industries: among others finance, biotech, automotive

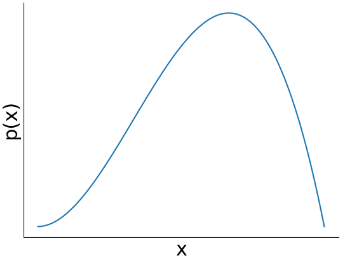
Need help with your project? Want to work with us?

www.tweag.io

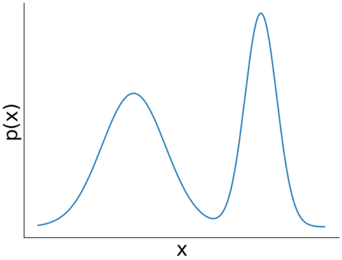
hello@tweag.io

Multimodality

Unimodal distribution:



Bimodal distribution:



Multimodality: common occurences

mixture models

For GMM parameterized by mixture weight θ and μ_1, μ_2, σ:

$$p(\theta, \mu_1, \mu_2, \sigma | y) = p(1 - \theta, \mu_2, \mu_1, \sigma | y)$$

models invariant under reflections

e.g., biomolecular structure determination:

$$L(\mathbf{x}_1, \dots \mathbf{x}_N) = p(D | \mathbf{x}_1, \dots \mathbf{x}_N) = L(\dots, |\mathbf{x}_i - \mathbf{x}_j| \dots)$$

latent variable models

e.g., probabilistic PCA: symmetry w.r.t. rotation of latent variable coordinates

Refresher: Metropolis-Hastings

Markov chain

Random process with

$$p(x_{i+1}|x_i, x_{i-1}, \ldots, x_1) = p(x_{i+1}|x_i)$$

$\rightarrow$ a Markov chain has no "memory"

In some conditions: converges to a unique invariant distribution $\pi(x)$

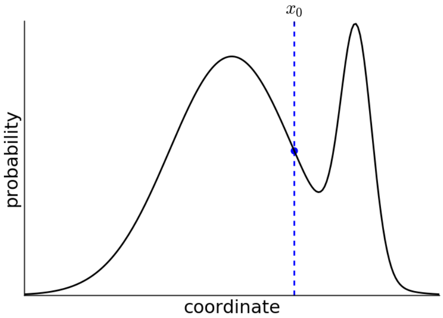
Metropolis-Hastings algorithm

Construct Markov chain with invariant distribution $\pi(x) = p(x)$:

1. starting at state, x_i, propose a new state x^*_{i+1} from $q(x^*_{i+1}|x_i)$
2. calculate acceptance probability p_{acc}
3. draw $u \sim \mathcal{U}(0, 1)$
4. if $u < p_{\text{acc}}$: $x_{i+1} = x^*_{i+1}$, else $x_{i+1} = x_i$

Metropolis (-Hastings)

Initialize with any state x_0



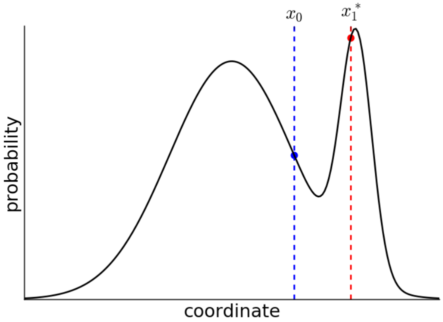
Sequence of states:
(x_0)

Metropolis et al., J. Chem. Phys (1953); Hastings, Biometrika (1970)

Metropolis (-Hastings)

Initial state: x_0

1. calculate a proposal state x_1^* by randomly perturbing x_0



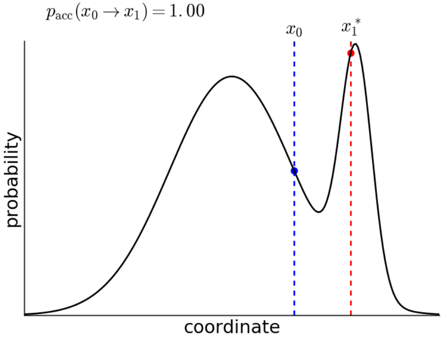
Metropolis et al., J. Chem. Phys (1953); Hastings, Biometrika (1970)

Metropolis (-Hastings)

Initial state: x_0

1. calculate a proposal state x_1^* by randomly perturbing x_0
2. calculate acceptance probability

$$p_{\text{acc}} = \min\left(1, \frac{p(x_1^*)}{p(x_0)}\right)$$



Metropolis et al., J. Chem. Phys (1953); Hastings, Biometrika (1970)

Metropolis (-Hastings)

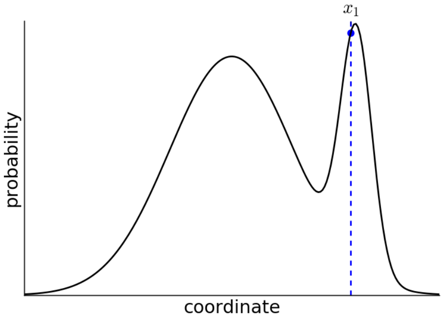
Initial state: x_0

1. calculate a proposal state x_1^* by randomly perturbing x_0
2. calculate acceptance probability

$$p_{\text{acc}} = \min\left(1, \frac{p(x_1^*)}{p(x_0)}\right)$$

3. with probability p_{acc}, accept proposal state x_1^* as the next state x_1, else copy x_0

Sequence of states:
(x_0, x_1)

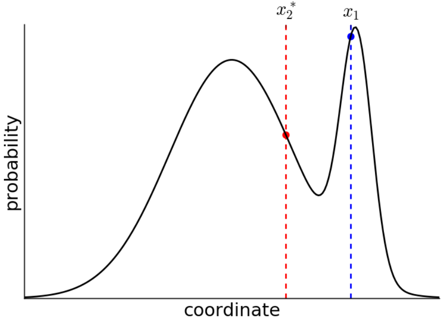


Metropolis et al., J. Chem. Phys (1953); Hastings, Biometrika (1970)

Metropolis (-Hastings)

Current state: x_1

1. calculate a proposal state x_2^* by randomly perturbing x_1



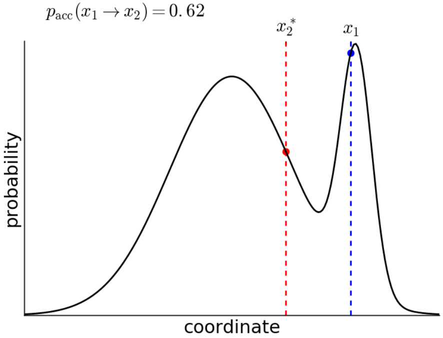
Metropolis et al., J. Chem. Phys (1953); Hastings, Biometrika (1970)

Metropolis (-Hastings)

Current state: x_1

1. calculate a proposal state x_2^* by randomly perturbing x_1
2. calculate acceptance probability

$$p_{\text{acc}} = \min\left(1, \frac{p(x_2^*)}{p(x_1)}\right)$$



Metropolis et al., J. Chem. Phys (1953); Hastings, Biometrika (1970)

Metropolis (-Hastings)

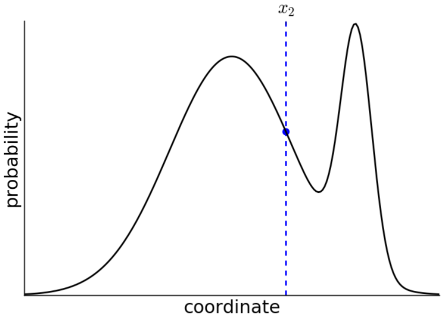
Current state: x_1

1. calculate a proposal state x_2^* by randomly perturbing x_1
2. calculate acceptance probability

$$p_{\text{acc}} = \min\left(1, \frac{p(x_2^*)}{p(x_1)}\right)$$

3. with probability p_{acc}, accept proposal state x_2^* as the next state x_2, else copy x_1

Sequence of states:
(x_0, x_1, x_2)



Metropolis et al., J. Chem. Phys (1953); Hastings, Biometrika (1970)

Metropolis (-Hastings)

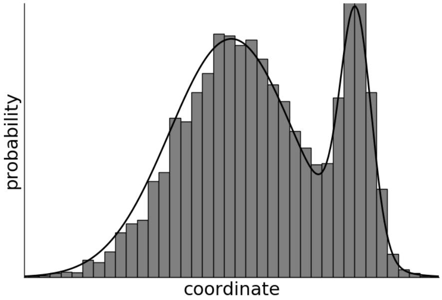
Current state: x_1

1. calculate a proposal state x_2^* by randomly perturbing x_1
2. calculate acceptance probability

$$p_{\text{acc}} = \min\left(1, \frac{p(x_2^*)}{p(x_1)}\right)$$

3. with probability p_{acc}, accept proposal state x_2^* as the next state x_2, else copy x_1

Sequence of states:
$(x_0, x_1, x_2, \ldots, x_n)$

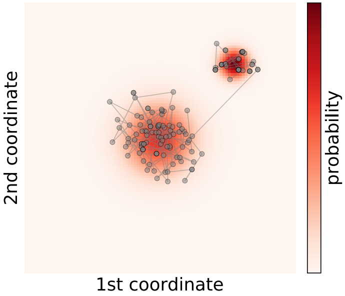


Metropolis et al., J. Chem. Phys (1953); Hastings, Biometrika (1970)

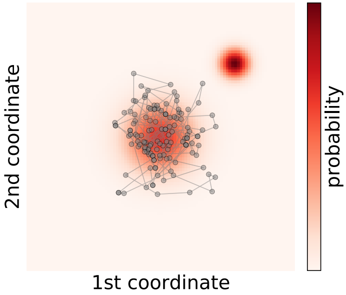
Multimodality: hard to sample

Markov chain can get stuck in modes:

Lucky us:



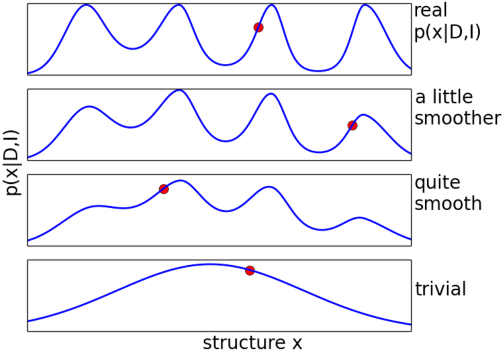
Unlucky us:



In higher dimensions: long time until all modes are discovered, if ever

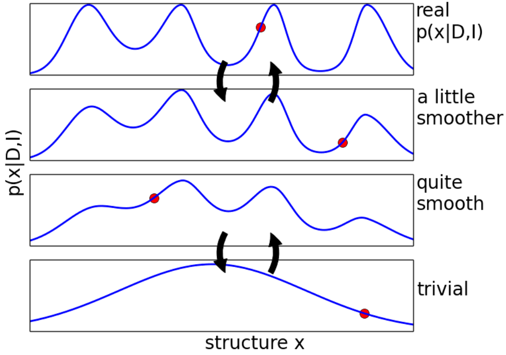
Replica Exchange

Simulate “flatter” versions of probability distribution...



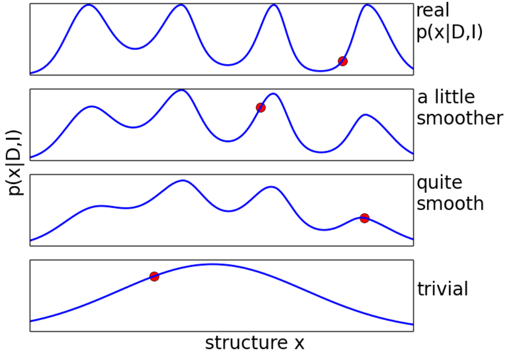
Replica Exchange

Simulate “flatter” versions of probability distribution and exchange states between Markov chains



Replica Exchange

Simulate “flatter” versions of probability distribution and exchange states between Markov chains



Replica Exchange: acceptance criterion (informal motivation)

Plain exchanges disturb equilibrium distributions
$\rightarrow$ correct with acceptance criterion

$$
\begin{tikzcd}
p_i(x): & x^i_{k-1} \arrow[r] & x^i_k \arrow[dr, "p_{i\to j}"] & x^i_{k+1} \arrow[r] & x^i_{k+2} \\
p_j(x): & x^j_{k-1} \arrow[r] & x^j_k & x^j_{k+1} \arrow[r] & x^j_{k+2}
\end{tikzcd}
$$

Probability of accepting x^i_k as $(k+1)$st
state of $p_j(x)$ chain:

$$p_{i\to j} = \frac{p_j(x^i_k)}{p_j(x^j_k)}$$

Replica Exchange: acceptance criterion (informal motivation)

Plain exchanges disturb equilibrium distributions
$\rightarrow$ correct with acceptance criterion

$$\begin{tikzcd}
p_i(x): & x^i_{k-1} \arrow[r] & x^i_k & x^i_{k+1} \arrow[r] & x^i_{k+2} \\
p_j(x): & x^j_{k-1} \arrow[r] & x^j_k \arrow[ur, "p_{j\to i}"] & x^j_{k+1} \arrow[r] & x^j_{k+2}
\end{tikzcd}$$

Probability of accepting x^i_k as $(k+1)$st state of $p_j(x)$ chain:

$$p_{i\to j} = \frac{p_j(x^i_k)}{p_j(x^j_k)}$$

Probability of accepting x^j_k as $(k+1)$st state of $p_i(x)$ chain:

$$p_{j\to i} = \frac{p_i(x^j_k)}{p_i(x^i_k)}$$

Replica Exchange: acceptance criterion (informal motivation)

Multiply $p_{i\to j}$ and $p_{j\to i}$:

General acceptance criterion

$$p_{\text{acc}} = \min\left\{1, \frac{p_i(x^j)}{p_i(x^i)} \times \frac{p_j(x^i)}{p_j(x^j)}\right\}$$

Common case with $p_i(x) \propto e^{-\beta_i E(x)}$:

$$p_{\text{acc}} = \min\left\{1, e^{(\beta_i-\beta_j)(E(x^i)-E(x^j))}\right\}$$

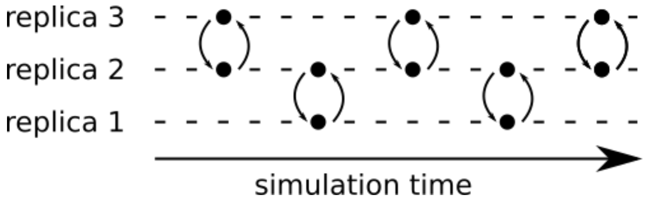
Physics terms

β: “inverse temperature”
$E(x) = -\log p(x)$: “energy”

The more different β_i and β_j, the lower p_{acc}!

Replica Exchange: choosing swap partners

Choose swap partners such that all replicas are connected, for example:

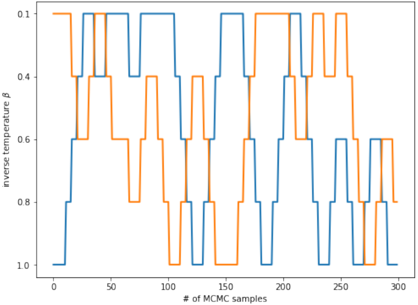


Deterministic even-odd swapping: try to swap $(\mathbf{0}, 1), (\mathbf{2}, 3), \ldots$ (even), then $(\mathbf{1}, 2), (\mathbf{3}, 4), \ldots$ (odd)

Stochastic even-odd swapping: decide randomly whether to swap even or odd

Replica Exchange: life of a state

States have to traverse the “temperature ladder” to be useful:



Stochastic swap scheme: traversal time $\propto$ $(\#\text{ replicas})^2$ (diffusive)
Deterministic swap scheme: better scaling (non-diffusive)

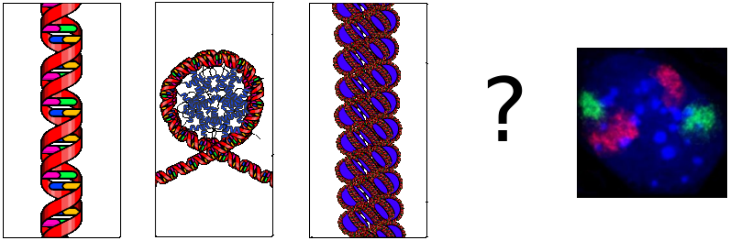
In general: more replicas $\not\Rightarrow$ better sampling

– demo time –

Knobs to tune

- choice of distribution family $p_i(x)$:
  - informed by structure of the sampling problem
  - in Bayesian inference often morphes posterior into prior
- number of interpolating distributions:
  - set target acceptance rate ($\approx 23\%$)
  - heuristics (e.g., constant KL divergence between neighbors)
  - hardware constraints
- frequency of swap attempts:
  "often"
- choice of swap partners

Application: Bayesian determination of chromosome structures



DNA double helix
2 nm (width)

nucleosomes
11 nm

30 nm fiber?

chromosome territories
~micrometer

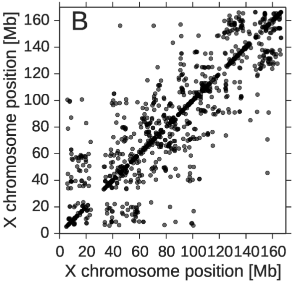
adapted from Richard Wheeler at en.wikipedia,
http://commons.wikimedia.org/w/index.php?curid=4017531

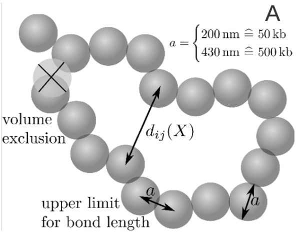
from Cremer & Cremer, 2001

Structure in "?" regime:

- active / passive chromatin compartments
- globular domains of several ≈ 100 nanometers in size

Bayesian chromosome structure determination





Data:
Binary contacts between loci; model with "logistic regression":

$$p(\text{contact } i \leftrightarrow j | \vec{x}_i, \vec{x}_j) = \frac{1}{1 + \mathrm{e}^{-\alpha(d_c - |\vec{x}_i - \vec{x}_j|)}}$$

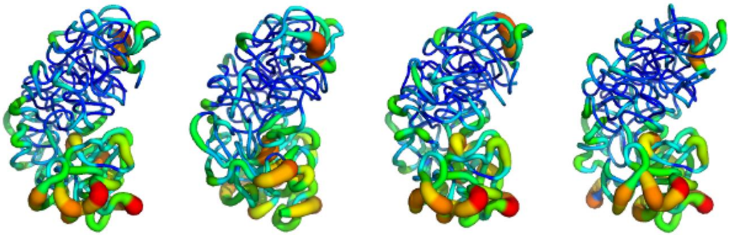
Prior:
beads-on-a-chain with volume exclusion $E_{\text{ve}}(\vec{X})$ and chain connectivity $E_{\text{nb}}(\vec{X})$

Tempered family of posteriors:

$$p_i(\vec{X}|D) \propto [p(D|\vec{X})]^{\beta_i} \times \mathrm{e}^{-E_{\text{ve}}(\vec{X})} \times \mathrm{e}^{-E_{\text{nb}}(\vec{X})}$$

Application: Bayesian determination of chromatin structures

Structures obtained from posterior sampling show several clusters:



Largest clusters: $\approx$ partial mirror images of each other

Keywords for further reading

Replica Exchange with Non-equilibrium Switches (RENS):

use non-equilibrium switching trajectories to increase acceptance rates

Recycle RE samples:

use histogram reweighting to calculate evidences and automatically tune interpolating distributions

Multidimensional Replica Exchange:

vary not one, but several “temperatures”