PMR: Sampling

Probabilistic Modelling and Reasoning

Amos Storkey

School of Informatics, University of Edinburgh

Outline

Outline

Outline

The problem

- Bayesian methods involve doing integrals wrt distributions which can be hard to do
- Bayesian methods involve representing intractable distributions
- Markov Chain Monte-Carlo

The problem

- Bayesian methods involve doing integrals wrt distributions which can be hard to do
- Bayesian methods involve representing intractable distributions
- Markov Chain Monte-Carlo

Monte Carlo approximation

- Suppose we have an expectation we wish to compute: i.e. an integral

$$A = \langle f(\boldsymbol{\theta}) \rangle_P = \int d\boldsymbol{\theta} P(\boldsymbol{\theta}) f(\boldsymbol{\theta})$$

  This occurs often: compute mean of distribution. Compute error for distribution. Compute best prediction for a distribution etc.
- Cannot compute it. But can sample (i.e. draw instance from distribution) from $P(\boldsymbol{\theta})$.
- Use

$$A \approx \tilde{A} = \frac{1}{N_S} \sum_{i=1}^{N_S} f(\boldsymbol{\theta}_i)$$

  where $\boldsymbol{\theta}_i$ are samples from $P(\boldsymbol{\theta})$, and N_S is the number of samples.
- This is a Monte-Carlo approximation.

Monte Carlo approximation

- Suppose we have an expectation we wish to compute: i.e. an integral
$$A = \langle f(\boldsymbol{\theta}) \rangle_P = \int d\boldsymbol{\theta} P(\boldsymbol{\theta}) f(\boldsymbol{\theta})$$
This occurs often: compute mean of distribution. Compute error for distribution. Compute best prediction for a distribution etc.
- Cannot compute it. But can sample (i.e. draw instance from distribution) from $P(\theta)$.
- Use
$$A \approx \tilde{A} = \frac{1}{N_S} \sum_{i=1}^{N_S} f(\boldsymbol{\theta}_i)$$
where θ_i are samples from $P(\theta)$, and N_S is the number of samples.
- This is a Monte-Carlo approximation.

Monte Carlo approximation

- Suppose we have an expectation we wish to compute: i.e. an integral

$$A = \langle f(\boldsymbol{\theta}) \rangle_P = \int d\boldsymbol{\theta} P(\boldsymbol{\theta}) f(\boldsymbol{\theta})$$

  This occurs often: compute mean of distribution. Compute error for distribution. Compute best prediction for a distribution etc.
- Cannot compute it. But can sample (i.e. draw instance from distribution) from $P(\boldsymbol{\theta})$.
- Use

$$A \approx \tilde{A} = \frac{1}{N_S} \sum_{i=1}^{N_S} f(\boldsymbol{\theta}_i)$$

  where $\boldsymbol{\theta}_i$ are samples from $P(\boldsymbol{\theta})$, and N_S is the number of samples.
- This is a Monte-Carlo approximation.

Monte Carlo approximation

- Suppose we have an expectation we wish to compute: i.e. an integral

$$A = \langle f(\boldsymbol{\theta}) \rangle_P = \int d\boldsymbol{\theta} P(\boldsymbol{\theta}) f(\boldsymbol{\theta})$$

  This occurs often: compute mean of distribution. Compute error for distribution. Compute best prediction for a distribution etc.
- Cannot compute it. But can sample (i.e. draw instance from distribution) from $P(\boldsymbol{\theta})$.
- Use

$$A \approx \tilde{A} = \frac{1}{N_S} \sum_{i=1}^{N_S} f(\boldsymbol{\theta}_i)$$

  where $\boldsymbol{\theta}_i$ are samples from $P(\boldsymbol{\theta})$, and N_S is the number of samples.
- This is a Monte-Carlo approximation.

Monte Carlo approximation

- Suppose we have an expectation we wish to compute: i.e. an integral

$$A = \langle f(\boldsymbol{\theta}) \rangle_P = \int d\boldsymbol{\theta} P(\boldsymbol{\theta}) f(\boldsymbol{\theta})$$

  This occurs often: compute mean of distribution. Compute error for distribution. Compute best prediction for a distribution etc.
- Cannot compute it. But can sample (i.e. draw instance from distribution) from $P(\boldsymbol{\theta})$.
- Use

$$A \approx \tilde{A} = \frac{1}{N_S} \sum_{i=1}^{N_S} f(\boldsymbol{\theta}_i)$$

  where $\boldsymbol{\theta}_i$ are samples from $P(\boldsymbol{\theta})$, and N_S is the number of samples.
- This is a Monte-Carlo approximation.

Monte Carlo properties

- Subject to some conditions, the approximation is asymptotically exact: as $N_S \to \infty$, $\tilde{A} \to A$ (Law of large numbers).
- The approximation error (s.d.) scales with $\sqrt{N_S}$ (Central Limit Theorem).
- The approximation depends on the smoothness of the function to be evaluated:
- More specifically the approximation error scales with the variance of the function value f over the distribution $P(\theta)$.
- The approximation error is independent of the size of the space that θ resides in.
- The same set of samples can be used for evaluating expectations of many different functions.
- Hence sampling procedure is independent of the expectation to be computed.

Monte Carlo properties

- Subject to some conditions, the approximation is asymptotically exact: as $N_S \to \infty$, $\tilde{A} \to A$ (Law of large numbers).
- The approximation error (s.d.) scales with $\sqrt{N_S}$ (Central Limit Theorem).
- The approximation depends on the smoothness of the function to be evaluated:
- More specifically the approximation error scales with the variance of the function value f over the distribution $P(\theta)$.
- The approximation error is independent of the size of the space that θ resides in.
- The same set of samples can be used for evaluating expectations of many different functions.
- Hence sampling procedure is independent of the expectation to be computed.

Monte Carlo properties

- Subject to some conditions, the approximation is asymptotically exact: as $N_S \to \infty$, $\tilde{A} \to A$ (Law of large numbers).
- The approximation error (s.d.) scales with $\sqrt{N_S}$ (Central Limit Theorem).
- The approximation depends on the smoothness of the function to be evaluated:
- More specifically the approximation error scales with the variance of the function value f over the distribution $P(\theta)$.
- The approximation error is independent of the size of the space that θ resides in.
- The same set of samples can be used for evaluating expectations of many different functions.
- Hence sampling procedure is independent of the expectation to be computed.

Monte Carlo properties

- Subject to some conditions, the approximation is asymptotically exact: as $N_S \to \infty$, $\tilde{A} \to A$ (Law of large numbers).
- The approximation error (s.d.) scales with $\sqrt{N_S}$ (Central Limit Theorem).
- The approximation depends on the smoothness of the function to be evaluated:
- More specifically the approximation error scales with the variance of the function value f over the distribution $P(\theta)$.
- The approximation error is independent of the size of the space that θ resides in.
- The same set of samples can be used for evaluating expectations of many different functions.
- Hence sampling procedure is independent of the expectation to be computed.

Monte Carlo properties

- Subject to some conditions, the approximation is asymptotically exact: as $N_S \to \infty$, $\tilde{A} \to A$ (Law of large numbers).
- The approximation error (s.d.) scales with $\sqrt{N_S}$ (Central Limit Theorem).
- The approximation depends on the smoothness of the function to be evaluated:
- More specifically the approximation error scales with the variance of the function value f over the distribution $P(\theta)$.
- The approximation error is independent of the size of the space that θ resides in.
- The same set of samples can be used for evaluating expectations of many different functions.
- Hence sampling procedure is independent of the expectation to be computed.

Monte Carlo properties

- Subject to some conditions, the approximation is asymptotically exact: as $N_S \to \infty$, $\tilde{A} \to A$ (Law of large numbers).
- The approximation error (s.d.) scales with $\sqrt{N_S}$ (Central Limit Theorem).
- The approximation depends on the smoothness of the function to be evaluated:
- More specifically the approximation error scales with the variance of the function value f over the distribution $P(\theta)$.
- The approximation error is independent of the size of the space that θ resides in.
- The same set of samples can be used for evaluating expectations of many different functions.
- Hence sampling procedure is independent of the expectation to be computed.

Monte Carlo properties

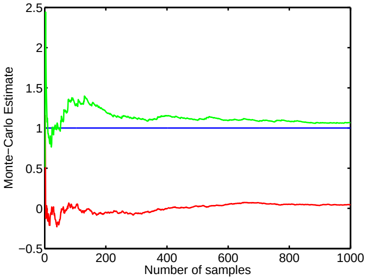
- Subject to some conditions, the approximation is asymptotically exact: as $N_S \to \infty$, $\tilde{A} \to A$ (Law of large numbers).
- The approximation error (s.d.) scales with $\sqrt{N_S}$ (Central Limit Theorem).
- The approximation depends on the smoothness of the function to be evaluated:
- More specifically the approximation error scales with the variance of the function value f over the distribution $P(\boldsymbol{\theta})$.
- The approximation error is independent of the size of the space that θ resides in.
- The same set of samples can be used for evaluating expectations of many different functions.
- Hence sampling procedure is independent of the expectation to be computed.

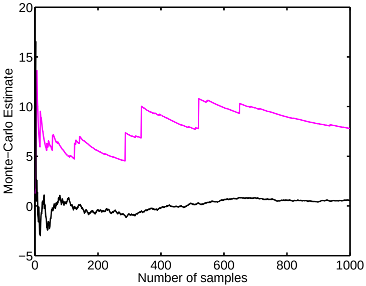
Monte Carlo properties

- Subject to some conditions, the approximation is asymptotically exact: as $N_S \to \infty$, $\tilde{A} \to A$ (Law of large numbers).
- The approximation error (s.d.) scales with $\sqrt{N_S}$ (Central Limit Theorem).
- The approximation depends on the smoothness of the function to be evaluated:
- More specifically the approximation error scales with the variance of the function value f over the distribution $P(\boldsymbol{\theta})$.
- The approximation error is independent of the size of the space that θ resides in.
- The same set of samples can be used for evaluating expectations of many different functions.
- Hence sampling procedure is independent of the expectation to be computed.

Monte-Carlo in action

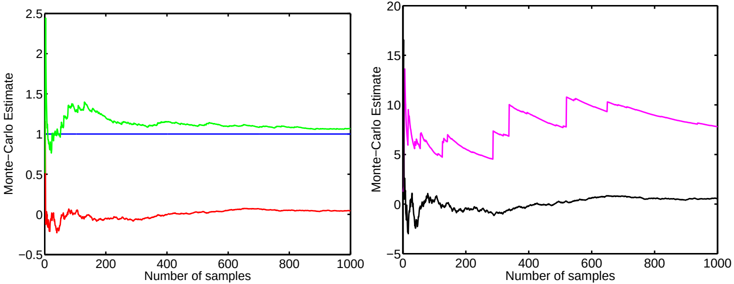
- Compute expectations with respect to a $N(0,1)$ Gaussian Distribution of $f_1(x) = 1$, $f_2(x) = x$, $f_3(x) = x^2$, $f_4(x) = 20\sin(x)$, $f_5(x) = \exp(0.6x^2)$ (!!)
- Some Graphs:





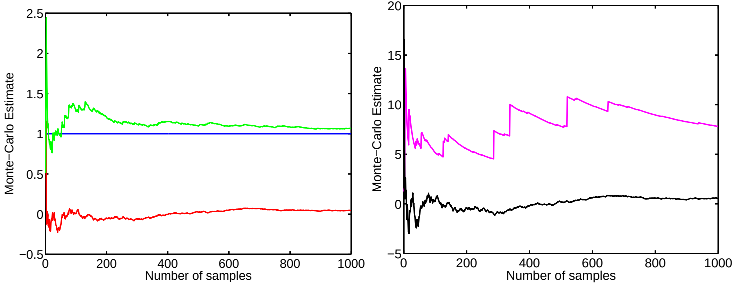
Monte-Carlo in action

- Compute expectations with respect to a $N(0,1)$ Gaussian Distribution of $f_1(x) = 1$, $f_2(x) = x$, $f_3(x) = x^2$, $f_4(x) = 20\sin(x)$, $f_5(x) = \exp(0.6x^2)$ (!!)
- Some Graphs:



Monte-Carlo in action

- Compute expectations with respect to a $N(0,1)$ Gaussian Distribution of $f_1(x) = 1$, $f_2(x) = x$, $f_3(x) = x^2$, $f_4(x) = 20\sin(x)$, $f_5(x) = \exp(0.6x^2)$ (!!)
- Some Graphs:



What if samples are not independent?

- Presuming the marginal distributions of the samples are correct, and
- Various other conditions (forgetfulness).
- This still works, but rate of convergence is reduced.

What if samples are not independent?

- Presuming the marginal distributions of the samples are correct, and
- Various other conditions (forgetfulness).
- This still works, but rate of convergence is reduced.

What if samples are not independent?

- Presuming the marginal distributions of the samples are correct, and
- Various other conditions (forgetfulness).
- This still works, but rate of convergence is reduced.

What if samples are not independent?

- Presuming the marginal distributions of the samples are correct, and
- Various other conditions (forgetfulness).
- This still works, but rate of convergence is reduced.

What if we don't know how to sample?

- One dimensional distributions are easy to sample from if we can evaluate the inverse of the cumulative distribution function $F(\theta)$:

```
s=rand;
sample = Finv(s);
```

- Otherwise may need another approach: e.g. Importance Sampling. Rejection Sampling.
- Will presume that we can evaluate the distribution we are interested in (up to a multiplicative constant).

Importance Sampling Summary

- Sample from a distribution that we can sample from.
- Reweight sample to adjust to the distribution we should have sampled from.
- Sample $\boldsymbol{\theta}_i$ from $Q(\boldsymbol{\theta})$. Compute weight $w_i \propto P(\boldsymbol{\theta}_i)/Q(\boldsymbol{\theta}_i)$.
- Represent expectation using:

$$\tilde{A} = \frac{1}{\sum_{i=1}^{N_S} w_i} \sum_{i=1}^{N_S} w_i f(\boldsymbol{\theta}_i)$$

Importance Sampling Summary

- Sample from a distribution that we can sample from.
- Reweight sample to adjust to the distribution we should have sampled from.
- Sample $\boldsymbol{\theta}_i$ from $Q(\boldsymbol{\theta})$. Compute weight $w_i \propto P(\boldsymbol{\theta}_i)/Q(\boldsymbol{\theta}_i)$.
- Represent expectation using:

$$\tilde{A} = \frac{1}{\sum_{i=1}^{N_S} w_i} \sum_{i=1}^{N_S} w_i f(\boldsymbol{\theta}_i)$$

Importance Sampling Summary

- Sample from a distribution that we can sample from.
- Reweight sample to adjust to the distribution we should have sampled from.
- Sample $\boldsymbol{\theta}_i$ from $Q(\boldsymbol{\theta})$. Compute weight $w_i \propto P(\boldsymbol{\theta}_i)/Q(\boldsymbol{\theta}_i)$.
- Represent expectation using:

$$\tilde{A} = \frac{1}{\sum_{i=1}^{N_S} w_i} \sum_{i=1}^{N_S} w_i f(\boldsymbol{\theta}_i)$$

Importance Representation

- Density $P(\mathbf{x}) = \frac{1}{Z}\Phi(\mathbf{x})$.
- Want

$$E_P(f) = \int d\mathbf{x}\, P(\mathbf{x}) f(\mathbf{x})$$

- Can approximate with (given supp(Q) $\supset$ supp(P))

$$E_P(f) = \int d\mathbf{x}\, Q(\mathbf{x}) w(\mathbf{x}) f(\mathbf{x})$$

  using $w(\mathbf{x}) = P(\mathbf{x})/Q(\mathbf{x})$
- What if cannot compute P, just Φ? Can use

$$E_P(f) = \frac{1}{Z} \int d\mathbf{x}\, Q(\mathbf{x}) w(\mathbf{x}) f(\mathbf{x})$$

  using $w(\mathbf{x}) = \Phi(\mathbf{x})/Q(\mathbf{x})$. and $Z = \int d\mathbf{x}\, w(\mathbf{x}) Q(\mathbf{x})$.

Importance Sampling

- Suppose we have sample set $\{\mathbf{x}_i | i = 1, 2, \ldots, N_S\}$ from $Q(\mathbf{x})$, and $w_i = \Phi(\mathbf{x})/Q(\mathbf{x})$.
- Let $Z_S = \sum_{i=1}^{N_S} w_i$. Then

$$\sum_{i=1}^{N_S} f(\mathbf{x}_i) \frac{w_i}{Z_S} \xrightarrow[N_S \to \infty]{a.s.} E_P(f).$$

- But Why?

Importance Sampling

- Stage 1:

$$E\left(\sum_{i=1}^{N} f(\mathbf{x}_i)\frac{w_i}{Z}\right) = E_P(f)$$

- Law of large numbers implies (given conditions) sum converges to $E_P(f)$.
- Stage 2: Note that also $Z_S \to Z$. Hence

$$\sum_{i=1}^{N} f(\mathbf{x}_i)\frac{w_i}{Z_S} = \left(\sum_{i=1}^{N} f(\mathbf{x}_i)\frac{w_i}{Z}\right)\left(\frac{Z}{Z_S}\right)$$

  tends to $E_P(f)$ almost surely.
- Note importance sampling is not an *unbiased* sampling technique, due to Z/Z_S.

Importance Sampling

- Sample from a distribution that we can sample from.
- Reweight sample to adjust to the distribution we should have sampled from.
- Sample $\boldsymbol{\theta}_i$ from $Q(\boldsymbol{\theta})$. Compute weight $w_i = P(\boldsymbol{\theta}_i)/Q(\boldsymbol{\theta}_i)$.
- Represent expectation using:

$$\tilde{A} = \frac{1}{\sum_{i=1}^{N_S} w_i} \sum_{i=1}^{N_S} w_i f(\boldsymbol{\theta}_i)$$

Importance Sampling

- Sample from a distribution that we can sample from.
- Reweight sample to adjust to the distribution we should have sampled from.
- Sample $\boldsymbol{\theta}_i$ from $Q(\boldsymbol{\theta})$. Compute weight $w_i = P(\boldsymbol{\theta}_i)/Q(\boldsymbol{\theta}_i)$.
- Represent expectation using:

$$\tilde{A} = \frac{1}{\sum_{i=1}^{N_S} w_i} \sum_{i=1}^{N_S} w_i f(\boldsymbol{\theta}_i)$$

Importance Sampling

- Sample from a distribution that we can sample from.
- Reweight sample to adjust to the distribution we should have sampled from.
- Sample $\boldsymbol{\theta}_i$ from $Q(\boldsymbol{\theta})$. Compute weight $w_i = P(\boldsymbol{\theta}_i)/Q(\boldsymbol{\theta}_i)$.
- Represent expectation using:

$$\tilde{A} = \frac{1}{\sum_{i=1}^{N_S} w_i} \sum_{i=1}^{N_S} w_i f(\boldsymbol{\theta}_i)$$

Rejection Sampling

- Sample from an upper bound to the distribution we want. Throw away samples to get the right shape distribution.
- Choose a distribution $Q(\boldsymbol{\theta})$ that we can sample from, s.t. $P(\boldsymbol{\theta}) < wQ(\boldsymbol{\theta})$
- Sample $\boldsymbol{\theta}_i$ from $Q(\boldsymbol{\theta})$. Sample u from uniform $U(0, 1)$.
- if $u < P(\boldsymbol{\theta}_i)/wQ(\boldsymbol{\theta}_i)$ accept sample $\boldsymbol{\theta}_i$ and move on to next i.
- Otherwise throw away $\boldsymbol{\theta}_i$ and try again.

Rejection Sampling

- Sample from an upper bound to the distribution we want. Throw away samples to get the right shape distribution.
- Choose a distribution $Q(\boldsymbol{\theta})$ that we can sample from, s.t. $P(\boldsymbol{\theta}) < wQ(\boldsymbol{\theta})$
- Sample θ_i from $Q(\theta)$. Sample u from uniform $U(0,1)$.
- if $u < P(\boldsymbol{\theta}_i)/wQ(\boldsymbol{\theta}_i)$ accept sample θ_i and move on to next i.
- Otherwise throw away θ_i and try again.

Rejection Sampling

- Sample from an upper bound to the distribution we want. Throw away samples to get the right shape distribution.
- Choose a distribution $Q(\theta)$ that we can sample from, s.t. $P(\theta) < wQ(\theta)$
- Sample θ_i from $Q(\theta)$. Sample u from uniform $U(0, 1)$.
- if $u < P(\theta_i)/wQ(\theta_i)$ accept sample θ_i and move on to next i.
- Otherwise throw away θ_i and try again.

Rejection Sampling

- Sample from an upper bound to the distribution we want. Throw away samples to get the right shape distribution.
- Choose a distribution $Q(\boldsymbol{\theta})$ that we can sample from, s.t. $P(\boldsymbol{\theta}) < wQ(\boldsymbol{\theta})$
- Sample $\boldsymbol{\theta}_i$ from $Q(\boldsymbol{\theta})$. Sample u from uniform $U(0, 1)$.
- if $u < P(\boldsymbol{\theta}_i)/wQ(\boldsymbol{\theta}_i)$ accept sample $\boldsymbol{\theta}_i$ and move on to next i.
- Otherwise throw away θ_i and try again.

Rejection Sampling

- Sample from an upper bound to the distribution we want. Throw away samples to get the right shape distribution.
- Choose a distribution $Q(\boldsymbol{\theta})$ that we can sample from, s.t. $P(\boldsymbol{\theta}) < wQ(\boldsymbol{\theta})$
- Sample $\boldsymbol{\theta}_i$ from $Q(\boldsymbol{\theta})$. Sample u from uniform $U(0, 1)$.
- if $u < P(\boldsymbol{\theta}_i)/wQ(\boldsymbol{\theta}_i)$ accept sample $\boldsymbol{\theta}_i$ and move on to next i.
- Otherwise throw away $\boldsymbol{\theta}_i$ and try again.

Slice sampling

- Sample from area of distribution by
- Initialise y.
- Sampling location x uniformly from the slice at current height: $(x|f(x) \geq y)$ (in practice various methods are used to do this).
- Sampling from the height uniformly at current location $(y|0 < y \leq f(x))$.
- Repeat
- Set of locations is sample.

Slice sampling

- Sample from area of distribution by
- Initialise y.
- Sampling location x uniformly from the slice at current height: $(x|f(x) \geq y)$ (in practice various methods are used to do this).
- Sampling from the height uniformly at current location $(y|0 < y \leq f(x))$.
- Repeat
- Set of locations is sample.

Slice sampling

- Sample from area of distribution by
- Initialise y.
- Sampling location x uniformly from the slice at current height: $(x|f(x) \geq y)$ (in practice various methods are used to do this).
- Sampling from the height uniformly at current location $(y|0 < y \leq f(x))$.
- Repeat
- Set of locations is sample.

Slice sampling

- Sample from area of distribution by
- Initialise y.
- Sampling location x uniformly from the slice at current height: $(x|f(x) \geq y)$ (in practice various methods are used to do this).
- Sampling from the height uniformly at current location $(y|0 < y \leq f(x))$.
- Repeat
- Set of locations is sample.

Slice sampling

- Sample from area of distribution by
- Initialise y.
- Sampling location x uniformly from the slice at current height: $(x|f(x) \geq y)$ (in practice various methods are used to do this).
- Sampling from the height uniformly at current location $(y|0 < y \leq f(x))$.
- Repeat
- Set of locations is sample.

Slice sampling

- Sample from area of distribution by
- Initialise y.
- Sampling location x uniformly from the slice at current height: $(x|f(x) \geq y)$ (in practice various methods are used to do this).
- Sampling from the height uniformly at current location $(y|0 < y \leq f(x))$.
- Repeat
- Set of locations is sample.

Higher dimensional systems

- Importance sampling and rejection sampling don't work well in higher dimensions.
- See discussion in 29.2, 29.3 of Mackay. Acceptance rate (or weight ratio) is exponentially decreasing with D.
- Other systems e.g. higher dimensional systems need another approach: Markov Chain Monte Carlo.

Higher dimensional systems

- Importance sampling and rejection sampling don't work well in higher dimensions.
- See discussion in 29.2, 29.3 of Mackay. Acceptance rate (or weight ratio) is exponentially decreasing with D.
- Other systems e.g. higher dimensional systems need another approach: Markov Chain Monte Carlo.

Higher dimensional systems

- Importance sampling and rejection sampling don't work well in higher dimensions.
- See discussion in 29.2, 29.3 of Mackay. Acceptance rate (or weight ratio) is exponentially decreasing with D.
- Other systems e.g. higher dimensional systems need another approach: Markov Chain Monte Carlo.

Higher dimensional systems

- Importance sampling and rejection sampling don't work well in higher dimensions.
- See discussion in 29.2, 29.3 of Mackay. Acceptance rate (or weight ratio) is exponentially decreasing with D.
- Other systems e.g. higher dimensional systems need another approach: Markov Chain Monte Carlo.

Markov Chains

- A Markov chain is a sequence model:

$$P(\theta_1, \theta_2, \ldots, \theta_N) = \prod_t P(\theta_t | \theta_{<t})$$

(where $\theta_{<t}$ denotes the set of all the values of $\theta_{t'}$ for $t' < t$) for which

$$P(\theta_t | \theta_{<t}) = P(\theta_t | \theta_{t-1}).$$

- This is called the Markov property.
- Typically we are interested in stationary Markov chains: $P(\theta_t | \theta_{t-1})$ is equal to some $P_T(\theta_1 | \theta_0)$ for all t.
- $P(\theta_t | \theta_{t-1})$ is called a transition probability.

Markov Chains

- A Markov chain is a sequence model:

$$P(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_N) = \prod_t P(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{<t})$$

(where $\boldsymbol{\theta}_{<t}$ denotes the set of all the values of $\theta_{t'}$ for $t' < t$) for which

$$P(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{<t}) = P(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1}).$$

- This is called the Markov property.
- Typically we are interested in stationary Markov chains: $P(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1})$ is equal to some $P_T(\boldsymbol{\theta}_1 | \boldsymbol{\theta}_0)$ for all t.
- $P(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1})$ is called a transition probability.

Markov Chains

- A Markov chain is a sequence model:

$$P(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_N) = \prod_t P(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{<t})$$

  (where $\boldsymbol{\theta}_{<t}$ denotes the set of all the values of $\theta_{t'}$ for $t' < t$) for which

$$P(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{<t}) = P(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1}).$$

- This is called the Markov property.
- Typically we are interested in stationary Markov chains: $P(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1})$ is equal to some $P_T(\boldsymbol{\theta}_1 | \boldsymbol{\theta}_0)$ for all t.
- $P(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1})$ is called a transition probability.

Properties of a Markov Chains

- Ergodicity: a Markov chain is ergodic if you would expect to get from each state to any other state in finite time, and if it is acyclic: its return time to any state is not always divisible by a number > 1.
- Reversibility: a Markov chain is reversible iff it satisfies detailed balance: for some distribution P_B: $P_B(\theta)P_T(\phi|\theta) = P_B(\phi)P_T(\theta|\phi)$
- Equilibrium Distribution: an ergodic Markov chain has a unique equilibrium distribution $P_\infty(\theta)$ such that

$$P_\infty(\theta) = \int d\theta' \; P_T(\theta|\theta')P_\infty(\theta')$$

- An ergodic reversible Markov chain satisfying detailed balance wrt P_B has P_B as its unique equilibrium distribution.

Properties of a Markov Chains

- Ergodicity: a Markov chain is ergodic if you would expect to get from each state to any other state in finite time, and if it is acyclic: its return time to any state is not always divisible by a number > 1.
- Reversibility: a Markov chain is reversible iff it satisfies detailed balance: for some distribution P_B: $P_B(\theta)P_T(\phi|\theta) = P_B(\phi)P_T(\theta|\phi)$
- Equilibrium Distribution: an ergodic Markov chain has a unique equilibrium distribution $P_\infty(\theta)$ such that

$$P_\infty(\theta) = \int d\theta' \; P_T(\theta|\theta')P_\infty(\theta')$$

- An ergodic reversible Markov chain satisfying detailed balance wrt P_B has P_B as its unique equilibrium distribution.

Properties of a Markov Chains

- Ergodicity: a Markov chain is ergodic if you would expect to get from each state to any other state in finite time, and if it is acyclic: its return time to any state is not always divisible by a number > 1.
- Reversibility: a Markov chain is reversible iff it satisfies detailed balance: for some distribution P_B: $P_B(\theta)P_T(\phi|\theta) = P_B(\phi)P_T(\theta|\phi)$
- Equilibrium Distribution: an ergodic Markov chain has a unique equilibrium distribution $P_\infty(\theta)$ such that

$$P_\infty(\theta) = \int d\theta' \; P_T(\theta|\theta')P_\infty(\theta')$$

- An ergodic reversible Markov chain satisfying detailed balance wrt P_B has P_B as its unique equilibrium distribution.

Properties of a Markov Chains

- Ergodicity: a Markov chain is ergodic if you would expect to get from each state to any other state in finite time, and if it is acyclic: its return time to any state is not always divisible by a number > 1.
- Reversibility: a Markov chain is reversible iff it satisfies detailed balance: for some distribution P_B: $P_B(\theta)P_T(\phi|\theta) = P_B(\phi)P_T(\theta|\phi)$
- Equilibrium Distribution: an ergodic Markov chain has a unique equilibrium distribution $P_\infty(\theta)$ such that

$$P_\infty(\theta) = \int d\theta' \; P_T(\theta|\theta')P_\infty(\theta')$$

- An ergodic reversible Markov chain satisfying detailed balance wrt P_B has P_B as its unique equilibrium distribution.

Properties of a Markov Chains

- Ergodicity: a Markov chain is ergodic if you would expect to get from each state to any other state in finite time, and if it is acyclic: its return time to any state is not always divisible by a number > 1.
- Reversibility: a Markov chain is reversible iff it satisfies detailed balance: for some distribution P_B: $P_B(\boldsymbol{\theta})P_T(\boldsymbol{\phi}|\boldsymbol{\theta}) = P_B(\boldsymbol{\phi})P_T(\boldsymbol{\theta}|\boldsymbol{\phi})$
- Equilibrium Distribution: an ergodic Markov chain has a unique equilibrium distribution $P_\infty(\theta)$ such that

$$P_\infty(\boldsymbol{\theta}) = \int d\boldsymbol{\theta}' \; P_T(\boldsymbol{\theta}|\boldsymbol{\theta}')P_\infty(\boldsymbol{\theta}')$$

- An ergodic reversible Markov chain satisfying detailed balance wrt P_B has P_B as its unique equilibrium distribution.

So what?

- Did not know how to sample from a distribution $P(\theta)$.
- Idea: Use a Markov chain. Design so $P(\theta)$ is equilibrium distribution.
- Run Markov chain sampling 'for long enough' to get samples from equilibrium distribution.
- How to design Markov chain? Ensure satisfies detailed balance wrt. $P(\theta)$,
- Sampling from a chain:
- Initialise state θ_0. Compute $P_T(\theta_1|\theta_0)$. Sample from this to get θ_1. Repeat ad infinitum (or until you get bored).
- Markov Chain Monte-Carlo (MCMC)

So what?

- Did not know how to sample from a distribution $P(\theta)$.
- Idea: Use a Markov chain. Design so $P(\theta)$ is equilibrium distribution.
- Run Markov chain sampling 'for long enough' to get samples from equilibrium distribution.
- How to design Markov chain? Ensure satisfies detailed balance wrt. $P(\theta)$,
- Sampling from a chain:
- Initialise state θ_0. Compute $P_T(\theta_1|\theta_0)$. Sample from this to get θ_1. Repeat ad infinitum (or until you get bored).
- Markov Chain Monte-Carlo (MCMC)

So what?

- Did not know how to sample from a distribution $P(\theta)$.
- Idea: Use a Markov chain. Design so $P(\theta)$ is equilibrium distribution.
- Run Markov chain sampling 'for long enough' to get samples from equilibrium distribution.
- How to design Markov chain? Ensure satisfies detailed balance wrt. $P(\theta)$,
- Sampling from a chain:
- Initialise state θ_0. Compute $P_T(\theta_1|\theta_0)$. Sample from this to get θ_1. Repeat ad infinitum (or until you get bored).
- Markov Chain Monte-Carlo (MCMC)

So what?

- Did not know how to sample from a distribution $P(\theta)$.
- Idea: Use a Markov chain. Design so $P(\theta)$ is equilibrium distribution.
- Run Markov chain sampling ‘for long enough’ to get samples from equilibrium distribution.
- How to design Markov chain? Ensure satisfies detailed balance wrt. $P(\theta)$,
- Sampling from a chain:
- Initialise state θ_0. Compute $P_T(\theta_1|\theta_0)$. Sample from this to get θ_1. Repeat ad infinitum (or until you get bored).
- Markov Chain Monte-Carlo (MCMC)

So what?

- Did not know how to sample from a distribution $P(\theta)$.
- Idea: Use a Markov chain. Design so $P(\theta)$ is equilibrium distribution.
- Run Markov chain sampling 'for long enough' to get samples from equilibrium distribution.
- How to design Markov chain? Ensure satisfies detailed balance wrt. $P(\theta)$,
- Sampling from a chain:
- Initialise state θ_0. Compute $P_T(\theta_1|\theta_0)$. Sample from this to get θ_1. Repeat ad infinitum (or until you get bored).
- Markov Chain Monte-Carlo (MCMC)

So what?

- Did not know how to sample from a distribution $P(\theta)$.
- Idea: Use a Markov chain. Design so $P(\theta)$ is equilibrium distribution.
- Run Markov chain sampling 'for long enough' to get samples from equilibrium distribution.
- How to design Markov chain? Ensure satisfies detailed balance wrt. $P(\theta)$,
- Sampling from a chain:
- Initialise state θ_0. Compute $P_T(\theta_1|\theta_0)$. Sample from this to get θ_1. Repeat ad infinitum (or until you get bored).
- Markov Chain Monte-Carlo (MCMC)

So what?

- Did not know how to sample from a distribution $P(\theta)$.
- Idea: Use a Markov chain. Design so $P(\theta)$ is equilibrium distribution.
- Run Markov chain sampling 'for long enough' to get samples from equilibrium distribution.
- How to design Markov chain? Ensure satisfies detailed balance wrt. $P(\theta)$,
- Sampling from a chain:
- Initialise state θ_0. Compute $P_T(\theta_1|\theta_0)$. Sample from this to get θ_1. Repeat ad infinitum (or until you get bored).
- Markov Chain Monte-Carlo (MCMC)

So what?

- Did not know how to sample from a distribution $P(\boldsymbol{\theta})$.
- Idea: Use a Markov chain. Design so $P(\boldsymbol{\theta})$ is equilibrium distribution.
- Run Markov chain sampling 'for long enough' to get samples from equilibrium distribution.
- How to design Markov chain? Ensure satisfies detailed balance wrt. $P(\boldsymbol{\theta})$,
- Sampling from a chain:
- Initialise state $\boldsymbol{\theta}_0$. Compute $P_T(\boldsymbol{\theta}_1|\boldsymbol{\theta}_0)$. Sample from this to get $\boldsymbol{\theta}_1$. Repeat ad infinitum (or until you get bored).
- Markov Chain Monte-Carlo (MCMC)

MCMC - Metropolis-Hastings Sampler

- Markov chain: Propose $Q(\theta'|\theta_t)$.
- Accept with probability

$$P(Accept) = \min\left(1, \frac{P(\theta')Q(\theta_t|\theta')}{P(\theta_t)Q(\theta'|\theta_t)}\right)$$

- If accept, set $\theta_{t+1} = \theta'$, else set $\theta_{t+1} = \theta_t$.

MCMC - Metropolis-Hastings Sampler

- Markov chain: Propose $Q(\theta'|\theta_t)$.
- Accept with probability

$$P(Accept) = \min\left(1, \frac{P(\theta')Q(\theta_t|\theta')}{P(\theta_t)Q(\theta'|\theta_t)}\right)$$

- If accept, set $\theta_{t+1} = \theta'$, else set $\theta_{t+1} = \theta_t$.

MCMC - Metropolis-Hastings Sampler

- Markov chain: Propose $Q(\theta'|\theta_t)$.
- Accept with probability

$$P(Accept) = \min\left(1, \frac{P(\theta')Q(\theta_t|\theta')}{P(\theta_t)Q(\theta'|\theta_t)}\right)$$

- If accept, set $\theta_{t+1} = \theta'$, else set $\theta_{t+1} = \theta_t$.

MCMC - Metropolis-Hastings Sampler

- Markov chain: Propose $Q(\theta'|\theta_t)$.
- Accept with probability

$$P(Accept) = \min\left(1, \frac{P(\theta')Q(\theta_t|\theta')}{P(\theta_t)Q(\theta'|\theta_t)}\right)$$

- If accept, set $\theta_{t+1} = \theta'$, else set $\theta_{t+1} = \theta_t$.

To Do

Examinable Reading

Mackay Chapter 29, 30

Preparatory Reading

Mackay Chapter 45

Extra Reading

Any papers of Radford Neal that take your fancy.