COMP 790-124: Goals for today

- Computing expectations
- Detailed balance
- Metropolis-Hastings,Gibbs
- How to turn a heuristic into an MCMC kernel
- Detour into Dirichlet-multinomial
- Finite mixture models and Gibbs sampling

Computing integrals

Crucial part of Bayesian inference is computation of integrals such as

$$p(\beta|\mathcal{D}) = \int_{\alpha \in A} p(\mathcal{D}|\alpha, \beta) p(\alpha) d\alpha$$

In general, integrals can be computed using a number of methods

- Numerical integration
- Monte Carlo methods
- Markov Chain Monte Carlo

If α is of even moderate dimension $\sim$5 the numerical integration has difficulties.

Notation

To simplify things, we will note that the integrals we wish to compute can be seen as expectations

$$p(\beta|\mathcal{D}) = \int_{\alpha \in A} p(\mathcal{D}|\alpha, \beta) p(\alpha) d\alpha = \mathrm{E}[p(\mathcal{D}|\alpha, \beta]$$

and hence we can focus on a general case of computing

$$\mathrm{E}[f(x)] = \int_{x} f(x) p(x) dx$$

Monte Carlo approximation

Let us assume that we are able to sample from $p(x)$.

Then using L samples from $p(x)$ we can approximate the desired expectation[1]

$$\mathrm{E}[f(x)] = \int_x f(x)p(x)dx \approx \frac{1}{L}\sum_{i=1}^{L} f(x_i)$$

[1] Strong law of large numbers gives us $1/n \sum A_i \rightarrow E[A]$ when $n \rightarrow \infty$

Importance sampling

Let us assume now that we are able to sample from some $q(x)$ efficiently.

$$
\begin{aligned}
\mathrm{E}[f(x)] &= \int_x f(x)p(x)dx = \int_x \underbrace{f(x)\frac{p(x)}{q(x)}}_{g(x)} q(x)dx \\
&= \int_x g(x)q(x)dx \\
&\approx \frac{1}{L}\sum_{i=1}^{L} g(x_i) \\
&= \frac{1}{L}\sum_{i=1}^{L} f(x_i)\underbrace{\frac{p(x_i)}{q(x_i)}}_{w_i}
\end{aligned}
$$

The values w_i are called importance weights.

Importance sampling for computing Bayes factor

Suppose we specified two models (MRFs for example)

$$\tilde{p}(x) \text{ and } \tilde{q}(x)$$

upto a normalization factor (we do not know $Z_p = \int_x \tilde{p}(x)dx$ or $Z_q = \int_x \tilde{q}(x)dx$).

Note that we can say $Z_q = \frac{\tilde{q}(x)}{q(x)}$, for any x such that $\tilde{q}(x) > 0$.

As long as we can sample from q, for example, we can compute the Bayes factor

$$\begin{aligned} \frac{Z_p}{Z_q} &= \frac{1}{Z_q}\int \tilde{p}(x)dx = \int \frac{\tilde{p}(x)}{\tilde{q}(x)} q(x)dx \\ &\approx \frac{1}{L}\sum_{l=1}^{L} \frac{\tilde{p}(x_l)}{\tilde{q}(x_l)} \end{aligned}$$

The state of the art techniques for computing such ratios is Annealed Importance Sampling.

Markov Chain Monte Carlo

If we can sample from $q(x)$ then expectations are easy:

$$\mathrm{E}[f(x)] = \int_x f(x)q(x)dx \approx \frac{1}{L}\sum_{i=1}^{L} f(x_i)$$

But producing independent samples from anything but standard distributions is quite hard.

Instead of producing independent samples, how about producing samples that are eventually independent from the target distribution?

Markov Chain Monte Carlo – Intuition

We construct a *random walker* that moves in the domain of x and guarantees that the number of times it visits a particular state x is proportional to $q(x)$. This is the Markov Chain of MCMC.

The samples obtained from the markov chain $x_1, \ldots, x_N$

$$q(x = v) \approx \frac{1}{N} \sum_{i=1}^{N} [x_i = v]$$

and hence

$$\mathrm{E}[f(x)] = \sum_x q(x) f(x) \approx \frac{1}{N} \sum_i f(x_i)$$

This is the Monte Carlo of MCMC.

Markov Chains

A sequence of random variables $\{x^1, \ldots, x^m\}$ forms a first order Markov chain if

$$p(x^{m+1}|x^1, \ldots, x^m) = p(x^{m+1}|x^m).$$

We specify a Markov chain by specifying initial variable probability

$$p(x^0)$$

and transition kernel

$$T_m(x^{m+1}|x^m) = p(x^{m+1}|x^m).$$

If the transition matrix is the same across all m ($T_m = T$) then the chain is homogenous.

A distribution q is invariant with respect to a Markov chain if

$$q(x) = \sum_y T(x|y)q(y)$$

Requirements for an MCMC

Three requirements

- Irreducibility: from any state there is a positive probability of reaching any other state in $k \geq 1$ steps
- Aperiodicity: chain does not get trapped in cycles (a single self transition is sufficient for this)
- Invariant distribution: chain leaves the target distribution invariant

Detailed balance

$$p(x) = \sum_y T(x|y)p(y)$$

We say that p is an invariant distribution for kernel T, also that p is an eigenvector of T (if discrete).

A stronger condition that implies invariance of p wrt to T

$$p(x)T(y|x) = p(y)T(x|y)$$

is called detailed balance[2]

[2] take sum over y on both sides of equation to show the implication

Metropolis-Hastings

Suppose we came up with a proposal distribution that generates a new state x^* from an old state x

$$q(x^*|x)$$

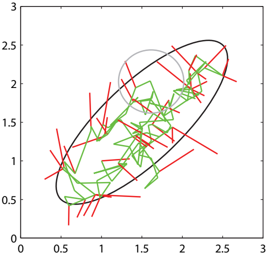
then we can design a 2-step algorithm that honors detailed balance

1. given current state x sample a proposed new state x^* from $q(x^*|x)$
2. accept this new state with prob

$$\alpha(x^*, x) = \min\left\{1, \frac{\overbrace{q(x|x^*)p(x^*)}^{\text{backward from } x^* \text{ to } x}}{\underbrace{q(x^*|x)p(x)}_{\text{forward from } x \text{ to } x^*}}\right\}$$

otherwise stay in x

Illustration of MH



Target distribution $p(x)$ is a 2D Gaussian. The 2 stddev contour is in black.

Proposal distribution is an isotropic/spherical Gaussian. The 2 stddev contour is in gray for a proposal distribution in one step.

Accepted steps are shown as green lines, rejected as red.

Showing detailed balance for Metropolis-Hastings

The kernel is

$$T(x^*|x) = q(x^*|x)\alpha(x^*, x)$$

then

$$\begin{aligned}
p(x)T(x^*|x) &= p(x)q(x^*|x)\alpha(x^*, x) \\
&= p(x)q(x^*|x)\min\left\{1, \frac{q(x|x^*)p(x^*)}{q(x^*|x)p(x)}\right\} \\
&= \min\left\{p(x)q(x^*|x), q(x|x^*)p(x^*)\right\} \\
&= \min\left\{\frac{p(x)q(x^*|x)}{q(x|x^*)p(x^*)}, 1\right\} q(x|x^*)p(x^*) \\
&= \min\left\{\frac{q(x^*|x)p(x)}{q(x|x^*)p(x^*)}, 1\right\} q(x|x^*)p(x^*) \\
&= T(x|x^*)p(x^*)
\end{aligned}$$

MH can work with unnormalized distributions

The acceptance probability

$$\alpha(x^*, x) = \min\left\{1, \frac{q(x|x^*)p(x^*)}{q(x^*|x)p(x)}\right\}$$

also works with unnormalized distributions since normalization constant cancels in the ratio

$$\frac{p(x^*)}{p(x)} = \frac{\frac{1}{Z}\tilde{p}(x^*)}{\frac{1}{Z}\tilde{p}(x)} = \frac{\tilde{p}(x^*)}{\tilde{p}(x)}$$

Gibbs sampler

Let us denote $x_{[-i]}$ the set of all but i^{th} variable in x.

A fairly straightforward sampler can be constructed that updates one variable at a time – like coordinate ascent but probabilistic.

$$q^{(i)}(x^*|x) \propto \begin{cases} p(x^*), & x^*_{[-i]} = x_{[-i]} \\ 0, & \text{otherwise} \end{cases}$$

Using Metropolis-Hastings the acceptance probability[3] is 1.

[3] exercise

Turning a heuristic climber into an MH kernel

Suppose you have constructed a heuristic hill climbing algorithm that has a set of moves and picks the best one based on a score S.

The score $S(x)$ induces a Boltzman distribution

$$p(x) \propto \exp\left\{-\frac{1}{\beta}S(x)\right\}$$

For each of the upward moves construct a reciprocal downward move that undoes it. Let f_i index all of the moves including $f_0(x) = x$ (gives aperiodicity)

$$q(x^*|x) = \frac{\sum_i [f_i(x) = x^*] \exp\{-\beta S(x^*)\}}{\sum_i \exp\{-\beta S(f_i(x))\}}$$

Still have to show irreducibility – transition graph is connected.

Convergence

MCMC algorithms in hands of novices do not converge, they get stuck.

The easiest diagnostic is restarting the sampler and seeing whether the expectations are the same regardless of the starting point.

Many cool tricks

- auxiliary variables
- annealing
- parallel chains
- split-and-merge

Dirichlet distribution

A distribution over distributions.

A sample from a Dirichlet distribution is a categorical distribution.[4]

For example, a sample from $\mathrm{Dir}(3, 10, 12)$ is a vector [0.2 0.3 0.5].

[4] altenatively parameters of a multinomial distribution

Dirichlet distribution

Dirichlet density is

$$p(\pi_1, \pi_2, \pi_3, \ldots, \pi_N | \alpha_1, ..., \alpha_N) = \frac{1}{Z} \prod_i \pi_i^{\alpha_i - 1}$$

where

$$Z = \int \prod_i \pi_i^{\alpha_i - 1} \mathrm{d}\pi = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)}$$

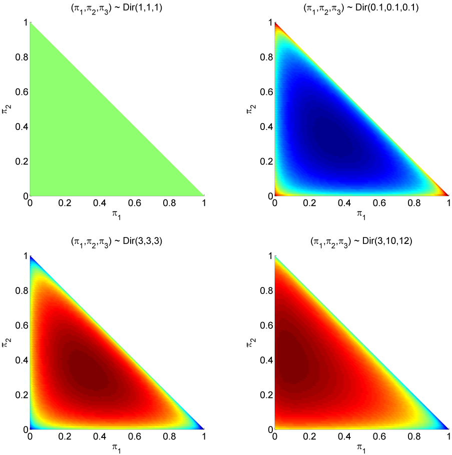
Γ denotes gamma function, a generalization of the factorial function

The parameters are constrained $\alpha_i > 0$ and density is defined on the interior of an n-dimensional simplex

$$\begin{aligned} \pi_i &> 0 \\ \sum_{i=1}^{n} \pi_i &= 1 \end{aligned}$$

Dirichlet distribution

We show log density across points $\pi = (\pi_1, \pi_2, \pi_3)$ lying in a simplex. Due to the constraint that $\sum_i \pi_i = 1$ we only have 2 free variables (here π_1 and π_2).



Dirichlet and multinomial distribution

We would be remiss in our comp. bio. duties if we did not mention the use of the Dirichlet-multinomial model for motifs.

We looked at position specific weighted matrices (PWM) – independent multinomial distributions.

$$p(\mathbf{x}|\theta) = \prod_{i} \prod_{v} \theta_{i,v}^{[\mathbf{x}_i = v]}$$

Dirichlet and multinomial distribution

ML estimation of θ can suffer from the same problems as coin fitting[5].

More importantly, we may want to incorporate some prior knowledge about likely patterns.

We can formulate a model

$$\begin{aligned} p(\theta_i) &= \mathrm{Dir}(\theta_{i,A}, \theta_{i,C}, \theta_{i,G}, \theta_{i,T} | \alpha_{i,A}, \alpha_{i,C}, \alpha_{i,G}, \alpha_{i,T}) \\ p(\mathbf{x}|\theta) &= \prod_i \prod_v \theta_{i,v}^{[\mathbf{x}_i = v]} \end{aligned}$$

where we can assume a noninformative prior

$$\alpha_{i,v} = c$$

or a prior on overall distribution of the letters in a sequence

$$\alpha_{i,v} = c_v$$

[5] e.g. if our training data has no occurrence of letter 'G' in position 3 then ML estimate $\hat{\theta}_{3,G} = 0$

Dirichlet and multinomial distribution

Given a set of sequences $\{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ posterior distribution for θ is

$$\begin{aligned} p(\theta|\mathbf{x}) &\propto p(\mathbf{x}|\theta)p(\theta) \propto \theta_{i,v}^{N_{i,v}} \theta_{i,v}^{\alpha_{i,v}-1} \\ &= \theta_{i,v}^{N_{i,v}+\alpha_{i,v}-1} \end{aligned}$$

where $N_{i,v} = \sum_t [\mathbf{x}_{t,i} = v]$ count of how many times we saw letter v in position i across all of the sequences.

Dirichlet and multinomial distribution – conjugacy

We note that

$$\begin{aligned} p(\theta_{i,v}) &\propto \theta_{i,v}^{\alpha_{i,v}-1} \\ p(\theta_{i,v}|\mathbf{x}) &\propto \theta_{i,v}^{N_{i,v}+\alpha_{i,v}-1} \end{aligned}$$

The form is the same – the posterior is also a Dirichlet distribution but with data-adjusted parameters.

In general, priors are conjugate to likelihoods if the posterior is of the same shape as the prior.

Dirichlet and multinomial distribution – MAP

The MAP estimates for θ are

$$\theta_{i,v}^{\mathrm{MAP}} = \frac{N_{i,v} + \alpha_{i,v} - 1}{\sum_b (N_{i,b} + \alpha_{i,b} - 1)} = \frac{N_{i,v} + \alpha_{i,v} - 1}{T + \sum_b (\alpha_{i,b} - 1)}$$

and again we see the role of the pseudo-counts: they either smooth out uneven empirical counts or provide push for parameters in the direction of prior knowledge.

Dirichlet-multinomial distribution

Again we are in luck in this set-up that we can compute marginal likelihood in closed form.

$$
\begin{aligned}
p(\mathbf{x}|\alpha) &= \int p(\mathbf{x}|\theta)p(\theta|\alpha)\mathrm{d}\theta \\
&= \int \left[\prod_t \prod_i \prod_v \theta_{i,v}^{[\mathbf{x}_{t,i}=v]}\right] \frac{1}{Z_\alpha} \prod_i \prod_v \theta_{i,v}^{\alpha_{i,v}-1}\mathrm{d}\theta \\
&= \frac{1}{Z(\alpha)} \int \left[\prod_i \prod_v \theta_{i,v}^{N_{i,v}}\right] \prod_i \prod_v \theta_{i,v}^{\alpha_{i,v}-1}\mathrm{d}\theta \\
&= \frac{1}{Z(\alpha)} \int \prod_i \prod_v \theta_{i,v}^{N_{i,v}+\alpha_{i,v}-1}\mathrm{d}\theta \\
&= \frac{Z(N+\alpha)}{Z(\alpha)}
\end{aligned}
$$

where

$$
Z(\beta) = \prod_i \frac{\prod_v \Gamma(\beta_{i,v})}{\Gamma(\sum_v \beta_{i,v})}
$$

Finite mixture models with Dirichlet prior

We will start with a finite mixture model with K classes

$$\begin{aligned} p(h = c|\pi) &= \pi_c \\ p(\mathbf{x}|h = c, \theta) &= f_c(\mathbf{x}; \theta_c) \end{aligned}$$

It has two types of parameters: mixing proportions π_i and component specific parameters θ_c (e.g. mean and covariance of Gaussian).

With our Bayesian hats on we want to place a prior on the parameters π.

The Dirichlet distribution is an ideal candidate

$$p(\pi) = \mathrm{Dir}(\pi|\alpha)$$

Mixture models - posterior

Our model now looks like

$$\begin{aligned} p(\pi) &= \mathrm{Dir}(\pi|\alpha) \\ p(h = c|\pi) &= \pi_c \\ p(\theta_c) &= G(\theta_c) \\ p(\mathbf{x}|h = c, \theta) &= f_c(\mathbf{x}; \theta_c) \end{aligned}$$

where we still need to specify a prior on parameters θ denoted by G and that choice will depend on choice of mixture components $f_1, \ldots, f_K$.

Unless f_c are also multinomials we won't be able to avail ourselves of conjugacy so we have to compute posteriors differently.

Gibbs sampling

In order to use Gibbs sampling we need to be able to compute conditional probabilities for all variables in our model:

- class membership for data instances $\mathbf{h}$
- mixing proportions π
- parameters of each mixture component θ.

Gibbs sampling – recap

Suppose all variables in the model are in the vector $\mathbf{z} = \{z_1, \ldots z_V\}$.

Gibbs sampler repeatedly sweeps through variables $i = 1, \ldots V$ and updates each by sampling its value from

$$p(z_i | \mathbf{z}_{[-i]})$$

An important observation when it comes to Gibbs is that

$$p(z_i = c | \mathbf{z}_{[-i]}) = \frac{1}{Z} p(z_i = c, \mathbf{z}_{[-i]})$$

where

$$Z = \sum_b p(z_i = b, \mathbf{z}_{[-i]})$$

Gibbs sampling – class membership

Conditional distribution for class membership for t^{th} instance given the rest of variables[6]

$$\begin{aligned} p(h_t|\mathbf{h}_{[-t]}, \alpha, \pi, \mathbf{x}) &= p(h_t|\pi, \mathbf{x}_t) \\ &\propto p(\mathbf{x}_t|h_t)p(h_t|\pi) \\ &= p(\mathbf{x}_t|h_t)\pi_{h_t} \\ &= f_{h_t}(\mathbf{x}_t)\pi_{h_t} \end{aligned}$$

so then in order to update h_t we would sample a class c from

$$q(c) = \frac{f_c(\mathbf{x}_t)\pi_c}{\sum_b f_b(\mathbf{x}_t)\pi_b}$$

[6] notation $\mathbf{h}_{[-t]}$ stands for $h_1, \ldots, h_{t-1}, h_{t+1}, \ldots, h_T$

Gibbs sampling - mixing proportions sampling

We need to compute the conditional probability for mixing proportions given the rest of variables[7]

$$p(\pi|\mathbf{h}, \alpha, \mathbf{x}) = p(\pi|\mathbf{h}, \alpha)$$

but this is easy

$$p(\pi|\mathbf{h}, \alpha) = \mathrm{Dir}(\pi|\alpha + N)$$

where $N_c = \sum_t [\mathbf{h}_t = c]$.

[7] we can't do this one coordinate at a time because $\sum_i \pi_i = 1$, so $\pi_{[-c]}$ fully determines π_c – sampler would never budge!

Gaussian sampling – component parameters

Up to this point all considerations of the model are quite generic – it does not depend what we choose for the mixture components (Gaussian, multinomials, Poissons or some assortment of distributions).

We will make a particular choice, a spherical Gaussian with a fixed covariance

$$f_c(\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_t; \mu_c, \sigma^2 \mathbf{I})$$

and hence $\theta = \{\mu_1, \ldots \mu_K\}$ and we place a Gaussian prior on θ

$$G(\theta) = \mathcal{N}(\theta; 0, \sigma_0^2 \mathbf{I})$$

Gibbs sampling – component parameters

$$p(\theta_c|\mathbf{x},\mathbf{h},\pi,\theta_{[-c]}) = p(\theta_c|\mathbf{x},\mathbf{h})$$

and in particular θ_c depends only on n_c points in class c and prior

$$\begin{aligned} p(\theta_c|\mathbf{x},\mathbf{h}) &\propto \mathcal{N}(\theta_c; 0, \sigma_0^2\mathbf{I}) \prod_{t:\mathbf{h}_t=c} \mathcal{N}(\mathbf{x}_t; \theta_c, \sigma^2\mathbf{I}) \\ &= \mathcal{N}\left(\theta_c; \frac{\sigma_0^2 \sum_{t:\mathbf{h}_t=c} \mathbf{x}_t}{\sigma^2 + n_c\sigma_0^2}, \frac{\sigma^2\sigma_0^2}{\sigma^2 + n_c\sigma_0^2}\mathbf{I}\right) \end{aligned}$$

and we are done. We have three types of updates and we can proceed to sample our distribution.

Demo and code staring

We did ...

- Computing expectations
- Detailed balance
- Metropolis-Hastings,Gibbs
- How to turn a heuristic into an MCMC kernel
- Dirichlet-multinomial and Bayesian mixture modeling