CPSC 535
Computing Normalizing Constants

AD

March 2007

Problem Statement

- We have discussed methods to sample from

$$\pi(\theta) = \frac{\gamma(\theta)}{Z}$$

where $\gamma(\theta)$ is known pointwise whereas

$$Z = \int \gamma(\theta)\, d\theta$$

is unknown.

Problem Statement

- We have discussed methods to sample from

$$\pi(\theta) = \frac{\gamma(\theta)}{Z}$$

where $\gamma(\theta)$ is known pointwise whereas

$$Z = \int \gamma(\theta)\, d\theta$$

is unknown.

- In many problems, we need to compute Z; e.g.

$$\pi(\theta) = \frac{p(\theta, y)}{p(y)}.$$

- We will first discuss methods which relies on the output of an MCMC algorithm generating samples $\theta^{(i)} \sim \pi$.

- We will first discuss methods which relies on the output of an MCMC algorithm generating samples $\theta^{(i)} \sim \pi$.
- Perhaps surprisingly there is no simple way to estimate Z from these samples.

- We will first discuss methods which relies on the output of an MCMC algorithm generating samples $\theta^{(i)} \sim \pi$.
- Perhaps surprisingly there is no simple way to estimate Z from these samples.
- Estimating Z is actually a problem typically more complex to solve than sampling from $\pi(\theta)$.

Chib's Identity

- For any $\theta \in \Theta$, we have
$$Z = \frac{\gamma(\theta)}{\pi(\theta)}.$$

Chib's Identity

- For any $\theta \in \Theta$, we have

$$Z = \frac{\gamma(\theta)}{\pi(\theta)}.$$

- Assume we can come up with a pointwise estimate of $\pi(\theta)$, say

$$\widehat{\pi}(\theta) = \frac{1}{N} \sum_{i=1}^{N} K\left(\theta - \theta^{(i)}\right)$$

where $K(\cdot)$ is a smoothing kernel, e.g. Gaussian.

Chib's Identity

- For any $\theta \in \Theta$, we have
$$Z = \frac{\gamma(\theta)}{\pi(\theta)}.$$
- Assume we can come up with a pointwise estimate of $\pi(\theta)$, say
$$\widehat{\pi}(\theta) = \frac{1}{N}\sum_{i=1}^{N} K\left(\theta - \theta^{(i)}\right)$$
where $K(\cdot)$ is a smoothing kernel, e.g. Gaussian.
- Then we can obtain
$$\widehat{Z}(\theta) = \frac{\gamma(\theta)}{\widehat{\pi}(\theta)}$$

Chib's Identity

- For any $\theta \in \Theta$, we have
$$Z = \frac{\gamma(\theta)}{\pi(\theta)}.$$
- Assume we can come up with a pointwise estimate of $\pi(\theta)$, say
$$\widehat{\pi}(\theta) = \frac{1}{N}\sum_{i=1}^{N} K\left(\theta - \theta^{(i)}\right)$$
where $K(\cdot)$ is a smoothing kernel, e.g. Gaussian.
- Then we can obtain
$$\widehat{Z}(\theta) = \frac{\gamma(\theta)}{\widehat{\pi}(\theta)}$$
- Typically, we will pick for θ the conditional mean estimate
$$\theta = \frac{1}{N}\sum_{i=1}^{N} \theta^{(i)}.$$

- Coming up with a good smoothing kernel is difficult in practice even if there exists a huge literature on the subject.

- Coming up with a good smoothing kernel is difficult in practice even if there exists a huge literature on the subject.
- Consider the special case where we have

$$\begin{aligned}\pi(\theta) &= \pi(\theta_1,\theta_2)=\pi(\theta_2|\,\theta_1)\,\pi(\theta_1)\\ &= \pi(\theta_2|\,\theta_1)\,\frac{\gamma(\theta_1)}{Z}\end{aligned}$$

where $\pi(\theta_1|\,\theta_2)$ is standard and $\gamma(\theta_1)$ is known.

- Coming up with a good smoothing kernel is difficult in practice even if there exists a huge literature on the subject.
- Consider the special case where we have

$$\begin{aligned} \pi(\theta) &= \pi(\theta_1, \theta_2) = \pi(\theta_2 | \theta_1) \pi(\theta_1) \\ &= \pi(\theta_2 | \theta_1) \frac{\gamma(\theta_1)}{Z} \end{aligned}$$

  where $\pi(\theta_1 | \theta_2)$ is standard and $\gamma(\theta_1)$ is known.
- We have for any θ_1 the identity

$$Z = \frac{\gamma(\theta_1)}{\pi(\theta_1)}$$

- Coming up with a good smoothing kernel is difficult in practice even if there exists a huge literature on the subject.
- Consider the special case where we have

$$\begin{aligned} \pi(\theta) &= \pi(\theta_1, \theta_2) = \pi(\theta_2 | \theta_1)\, \pi(\theta_1) \\ &= \pi(\theta_2 | \theta_1) \frac{\gamma(\theta_1)}{Z} \end{aligned}$$

  where $\pi(\theta_1 | \theta_2)$ is standard and $\gamma(\theta_1)$ is known.
- We have for any θ_1 the identity

$$Z = \frac{\gamma(\theta_1)}{\pi(\theta_1)}$$

- To approximate $\pi(\theta_1)$, we use the identity

$$\begin{aligned} \pi(\theta_1) &= \int \pi(\theta_1 | \theta_2)\, \pi(\theta_2)\, d\theta_2 \\ &\approx \frac{1}{N} \sum_{i=1}^{N} \pi\left(\theta_1 \middle| \theta_2^{(i)}\right) \end{aligned}$$

  where $\left(\theta_1^{(i)}, \theta_2^{(i)}\right)$ might have been generated using the Gibbs sampler.

- Similarly, we would usually pick for $\theta_1 = \frac{1}{N}\sum_{i=1}^{N}\theta_1^{(i)}$ and the final estimate is

$$\widehat{Z}\left(\theta_1\right) = \frac{\gamma\left(\theta_1\right)}{\frac{1}{N}\sum_{i=1}^{N}\pi\left(\theta_1 \middle| \theta_2^{(i)}\right)}.$$

- Similarly, we would usually pick for $\theta_1 = \frac{1}{N}\sum_{i=1}^{N}\theta_1^{(i)}$ and the final estimate is

$$\widehat{Z}(\theta_1) = \frac{\gamma(\theta_1)}{\frac{1}{N}\sum_{i=1}^{N}\pi\left(\theta_1 \middle| \theta_2^{(i)}\right)}.$$

- This choice performs much better than a standard smoothing estimate.

- Similarly, we would usually pick for $\theta_1 = \frac{1}{N}\sum_{i=1}^{N}\theta_1^{(i)}$ and the final estimate is

$$\widehat{Z}\left(\theta_1\right) = \frac{\gamma\left(\theta_1\right)}{\frac{1}{N}\sum_{i=1}^{N}\pi\left(\theta_1 \middle| \theta_2^{(i)}\right)}.$$

- This choice performs much better than a standard smoothing estimate.
- This approach remains however limited to low-dimensional problems.

Harmonic Mean Estimate

- Let us introduce the auxiliary probability distribution $q(\theta)$ then we have the following identity

$$\frac{1}{Z} = \int \frac{q(\theta)}{\gamma(\theta)} \pi(\theta) \, d\theta.$$

Harmonic Mean Estimate

- Let us introduce the auxiliary probability distribution $q(\theta)$ then we have the following identity

$$\frac{1}{Z} = \int \frac{q(\theta)}{\gamma(\theta)} \pi(\theta)\, d\theta.$$

- $q(\theta)$ is not an importance distribution here, $\pi(\theta)$ is.

Harmonic Mean Estimate

- Let us introduce the auxiliary probability distribution $q(\theta)$ then we have the following identity

$$\frac{1}{Z} = \int \frac{q(\theta)}{\gamma(\theta)} \pi(\theta)\, d\theta.$$

- $q(\theta)$ is not an importance distribution here, $\pi(\theta)$ is.
- It suggests the following Monte Carlo approximation

$$\widehat{\frac{1}{Z}} = \frac{1}{N} \sum_{i=1}^{N} \frac{q\left(\theta^{(i)}\right)}{\gamma\left(\theta^{(i)}\right)}, \text{ i.e. } \widehat{Z} = \left(\frac{1}{N} \sum_{i=1}^{N} \frac{q\left(\theta^{(i)}\right)}{\gamma\left(\theta^{(i)}\right)} \right)^{-1}.$$

Harmonic Mean Estimate

- Let us introduce the auxiliary probability distribution $q(\theta)$ then we have the following identity

$$\frac{1}{Z} = \int \frac{q(\theta)}{\gamma(\theta)} \pi(\theta)\, d\theta.$$

- $q(\theta)$ is not an importance distribution here, $\pi(\theta)$ is.
- It suggests the following Monte Carlo approximation

$$\widehat{\frac{1}{Z}} = \frac{1}{N} \sum_{i=1}^{N} \frac{q\left(\theta^{(i)}\right)}{\gamma\left(\theta^{(i)}\right)}, \text{ i.e. } \widehat{Z} = \left(\frac{1}{N} \sum_{i=1}^{N} \frac{q\left(\theta^{(i)}\right)}{\gamma\left(\theta^{(i)}\right)} \right)^{-1}.$$

- This algorithm requires selecting a distribution $q(\theta)$ such that for any $\theta \in \Theta$

$$\frac{q(\theta)}{\pi(\theta)} < C$$

- **Example**: If $\pi(\theta) = p(\theta | y)$ then it is tempting to select $q(\theta) = p(\theta)$ and

$$\widehat{\frac{1}{Z}} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{p\left(y | \theta^{(i)} \right)}.$$

- **Example**: If $\pi(\theta) = p(\theta | y)$ then it is tempting to select $q(\theta) = p(\theta)$ and
$$\widehat{\frac{1}{Z}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{p\left(y | \theta^{(i)}\right)}.$$
- However, this estimate will have an unbounded variance in most cases as $p(\theta | y)$ has typically thinner tails than $p(\theta)$.

- **Example**: If $\pi(\theta) = p(\theta|\, y)$ then it is tempting to select $q(\theta) = p(\theta)$ and
$$\widehat{\frac{1}{Z}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{p\left(y|\,\theta^{(i)}\right)}.$$
- However, this estimate will have an unbounded variance in most cases as $p(\theta|\, y)$ has typically thinner tails than $p(\theta)$.
- Even if we pick $\frac{q(\theta)}{\pi(\theta)} < C$, the variance of this estimate will typically be large.

- **Example**: If $\pi(\theta) = p(\theta | y)$ then it is tempting to select $q(\theta) = p(\theta)$ and
$$\widehat{\frac{1}{Z}} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{p\left(y | \theta^{(i)}\right)}.$$
- However, this estimate will have an unbounded variance in most cases as $p(\theta | y)$ has typically thinner tails than $p(\theta)$.
- Even if we pick $\frac{q(\theta)}{\pi(\theta)} < C$, the variance of this estimate will typically be large.
- The harmonic mean estimate is restricted to low-dimensional problems.

(Brute Force) Importance Sampling

- Assume now that we will based our MC estimate of Z on samples from another distribution $q(\theta)$.

(Brute Force) Importance Sampling

- Assume now that we will based our MC estimate of Z on samples from another distribution $q(\theta)$.
- We have the identity

$$Z = \int \frac{\gamma(\theta)}{q(\theta)} q(\theta)\, d\theta$$

so by using samples $\theta^{(i)} \sim q(\theta)$

$$\widehat{Z} = \frac{1}{N} \sum_{i=1}^{N} \frac{\gamma\left(\theta^{(i)}\right)}{q\left(\theta^{(i)}\right)}.$$

(Brute Force) Importance Sampling

- Assume now that we will based our MC estimate of Z on samples from another distribution $q(\theta)$.
- We have the identity

$$Z = \int \frac{\gamma(\theta)}{q(\theta)} q(\theta)\, d\theta$$

so by using samples $\theta^{(i)} \sim q(\theta)$

$$\widehat{Z} = \frac{1}{N} \sum_{i=1}^{N} \frac{\gamma\left(\theta^{(i)}\right)}{q\left(\theta^{(i)}\right)}.$$

- For the algorithm to work properly, we need

$$\frac{\pi(\theta)}{q(\theta)} < C.$$

(Brute Force) Importance Sampling

- Assume now that we will based our MC estimate of Z on samples from another distribution $q(\theta)$.
- We have the identity

$$Z = \int \frac{\gamma(\theta)}{q(\theta)} q(\theta)\, d\theta$$

so by using samples $\theta^{(i)} \sim q(\theta)$

$$\widehat{Z} = \frac{1}{N} \sum_{i=1}^{N} \frac{\gamma\left(\theta^{(i)}\right)}{q\left(\theta^{(i)}\right)}.$$

- For the algorithm to work properly, we need

$$\frac{\pi(\theta)}{q(\theta)} < C.$$

- Once more, this is nothing but Importance Sampling and will fail for high-dimensional problem.

Bridge Sampling

- Assume you have two distributions $\pi_1(\theta)$ and $\pi_0(\theta)$, we are interested in computing the ratio

$$\frac{Z_1}{Z_0} = \frac{\int \gamma_1(\theta)\, d\theta}{\int \gamma_0(\theta)\, d\theta}.$$

Bridge Sampling

- Assume you have two distributions $\pi_1(\theta)$ and $\pi_0(\theta)$, we are interested in computing the ratio

$$\frac{Z_1}{Z_0} = \frac{\int \gamma_1(\theta)\, d\theta}{\int \gamma_0(\theta)\, d\theta}.$$

- The Bridge sampling identity is

$$\frac{Z_1}{Z_0} = \frac{\int \gamma_1(\theta)\, \alpha(\theta)\, \pi_0(\theta)\, d\theta}{\int \gamma_0(\theta)\, \alpha(\theta)\, \pi_1(\theta)\, d\theta}$$

where $\alpha(\theta)$ is an arbitrary function satisfying

$$\int \alpha(\theta)\, \pi_0(\theta)\, \pi_1(\theta)\, d\theta < \infty$$

Bridge Sampling

- This suggests the following MC estimate given N_0 samples $\theta_0^{(i)}$ from $\pi_0(\theta)$ and N_1 samples $\theta_1^{(i)}$ from $\pi_1(\theta)$

$$\widehat{\frac{Z_1}{Z_0}} = \frac{\frac{1}{N_0}\sum_{i=1}^{N} \gamma_1\left(\theta_0^{(i)}\right) \alpha\left(\theta_0^{(i)}\right)}{\frac{1}{N_1}\sum_{i=1}^{N} \gamma_0\left(\theta_1^{(i)}\right) \alpha\left(\theta_1^{(i)}\right)}.$$

- This suggests the following MC estimate given N_0 samples $\theta_0^{(i)}$ from $\pi_0(\theta)$ and N_1 samples $\theta_1^{(i)}$ from $\pi_1(\theta)$

$$\widehat{\frac{Z_1}{Z_0}} = \frac{\frac{1}{N_0}\sum_{i=1}^{N} \gamma_1\left(\theta_0^{(i)}\right)\alpha\left(\theta_0^{(i)}\right)}{\frac{1}{N_1}\sum_{i=1}^{N} \gamma_0\left(\theta_1^{(i)}\right)\alpha\left(\theta_1^{(i)}\right)}.$$

- Taking for example $\alpha(\theta) = \gamma_0^{-1}(\theta)$, we have

$$\frac{Z_1}{Z_0} = \int \frac{\gamma_1(\theta)}{\gamma_0(\theta)}\pi_0(\theta)\, d\theta$$

which is the harmonic mean estimate.

- Bridge sampling is thus a generalization of what we have discussed before.

- Bridge sampling is thus a generalization of what we have discussed before.
- Assuming that we can obtain iid samples from $\pi_0(\theta)$ and $\pi_1(\theta)$ then the optimal $\alpha(\theta)$ in the sense of minimizing the asymptotic variance of $\log\left(\widehat{\frac{Z_1}{Z_0}}\right)$ is given by

$$\begin{aligned}\alpha(\theta) &\propto \frac{1}{s_0\pi_0(\theta)+s_1\pi_1(\theta)} \\ &\propto \frac{1}{s_0 Z_0^{-1}\gamma_0(\theta)+s_1 Z_1^{-1}\gamma_1(\theta)}\end{aligned}$$

where

$$s_0=\frac{N_0}{N_0+N_1},\ s_1=\frac{N_1}{N_0+N_1}.$$

- Bridge sampling is thus a generalization of what we have discussed before.
- Assuming that we can obtain iid samples from $\pi_0(\theta)$ and $\pi_1(\theta)$ then the optimal $\alpha(\theta)$ in the sense of minimizing the asymptotic variance of $\log\left(\widehat{\frac{Z_1}{Z_0}}\right)$ is given by

$$\begin{aligned}\alpha(\theta) &\propto \frac{1}{s_0\pi_0(\theta)+s_1\pi_1(\theta)}\\ &\propto \frac{1}{s_0Z_0^{-1}\gamma_0(\theta)+s_1Z_1^{-1}\gamma_1(\theta)}\end{aligned}$$

  where

$$s_0=\frac{N_0}{N_0+N_1},\ s_1=\frac{N_1}{N_0+N_1}.$$

- Clearly, this optimal choice cannot be selected but it suggests using an iterative procedure.

- Bridge sampling is thus a generalization of what we have discussed before.
- Assuming that we can obtain iid samples from $\pi_0(\theta)$ and $\pi_1(\theta)$ then the optimal $\alpha(\theta)$ in the sense of minimizing the asymptotic variance of $\log\left(\widehat{\frac{Z_1}{Z_0}}\right)$ is given by

$$\begin{aligned}\alpha(\theta) &\propto \frac{1}{s_0\pi_0(\theta)+s_1\pi_1(\theta)}\\ &\propto \frac{1}{s_0 Z_0^{-1}\gamma_0(\theta)+s_1 Z_1^{-1}\gamma_1(\theta)}\end{aligned}$$

  where

$$s_0=\frac{N_0}{N_0+N_1},\ s_1=\frac{N_1}{N_0+N_1}.$$

- Clearly, this optimal choice cannot be selected but it suggests using an iterative procedure.
- Such a procedure can considerably improve performance of 'naive' techniques but is still limited.

From Bridge Sampling to Path Sampling

- We can rewrite

$$\alpha(\theta) = \frac{\gamma_{1/2}(\theta)}{\gamma_0(\theta) . \gamma_1(\theta)}$$

where $\gamma_{1/2}(\theta)$ is an intermediate unnormalized density and

$$\begin{aligned} \frac{Z_1}{Z_0} &= \frac{\int \gamma_1(\theta)\, \alpha(\theta)\, \pi_0(\theta)\, d\theta}{\int \gamma_0(\theta)\, \alpha(\theta)\, \pi_1(\theta)\, d\theta} = \frac{\int \frac{\gamma_{1/2}(\theta)}{\gamma_0(\theta)} \pi_0(\theta)\, d\theta}{\int \frac{\gamma_{1/2}(\theta)}{\gamma_1(\theta)} \pi_1(\theta)\, d\theta} \\ &= \frac{Z_{1/2} / Z_0}{Z_{1/2} / Z_1} \end{aligned}$$

From Bridge Sampling to Path Sampling

- We can rewrite

$$\alpha(\theta) = \frac{\gamma_{1/2}(\theta)}{\gamma_0(\theta) . \gamma_1(\theta)}$$

where $\gamma_{1/2}(\theta)$ is an intermediate unnormalized density and

$$\begin{aligned} \frac{Z_1}{Z_0} &= \frac{\int \gamma_1(\theta) \alpha(\theta) \pi_0(\theta) d\theta}{\int \gamma_0(\theta) \alpha(\theta) \pi_1(\theta) d\theta} = \frac{\int \frac{\gamma_{1/2}(\theta)}{\gamma_0(\theta)} \pi_0(\theta) d\theta}{\int \frac{\gamma_{1/2}(\theta)}{\gamma_1(\theta)} \pi_1(\theta) d\theta} \\ &= \frac{Z_{1/2} / Z_0}{Z_{1/2} / Z_1} \end{aligned}$$

- So we can think of bridge sampling as moving from γ_0 to γ_1 by introducing $\gamma_{1/2}$ and the optimal intermediate (unnormalized) distribution is

$$\gamma_{1/2}(\theta) = \frac{\pi_0(\theta) \pi_1(\theta)}{s_0 \pi_0(\theta) + s_1 \pi_1(\theta)}.$$

- We can push this bridge idea further by introducing $L-1$ intermediate distributions; say $\gamma\left(\theta \mid \alpha_l\right)$ where $l=0,\ldots,L$ with $\gamma\left(\theta \mid \alpha_0\right)=\gamma_0\left(\theta\right)$ and $\gamma\left(\theta \mid \alpha_L\right)=\gamma_1\left(\theta\right)$ and $Z\left(\alpha_l\right)=\int \gamma\left(\theta \mid \alpha_l\right) d\theta$ using

$$\frac{Z_1}{Z_0}=\prod_{l=1}^{L}\frac{Z\left(\alpha_l\right)}{Z\left(\alpha_{l-1}\right)}.$$

- We can push this bridge idea further by introducing $L-1$ intermediate distributions; say $\gamma\left(\theta|\,\alpha_l\right)$ where $l=0,...,L$ with $\gamma\left(\theta|\,\alpha_0\right)=\gamma_0\left(\theta\right)$ and $\gamma\left(\theta|\,\alpha_L\right)=\gamma_1\left(\theta\right)$ and $Z\left(\alpha_l\right)=\int\gamma\left(\theta|\,\alpha_l\right)d\theta$ using

$$\frac{Z_1}{Z_0}=\prod_{l=1}^{L}\frac{Z\left(\alpha_l\right)}{Z\left(\alpha_{l-1}\right)}.$$

- Using a sequence of intermediate distributions to move from $\gamma_0\left(\theta\right)$ to $\gamma_1\left(\theta\right)$ is a crucial and ubiquitous idea in Monte Carlo.

- We can push this bridge idea further by introducing $L-1$ intermediate distributions; say $\gamma\left(\theta|\,\alpha_l\right)$ where $l=0,...,L$ with $\gamma\left(\theta|\,\alpha_0\right)=\gamma_0\left(\theta\right)$ and $\gamma\left(\theta|\,\alpha_L\right)=\gamma_1\left(\theta\right)$ and $Z\left(\alpha_l\right)=\int\gamma\left(\theta|\,\alpha_l\right)d\theta$ using

$$\frac{Z_1}{Z_0}=\prod_{l=1}^{L}\frac{Z\left(\alpha_l\right)}{Z\left(\alpha_{l-1}\right)}.$$

- Using a sequence of intermediate distributions to move from $\gamma_0\left(\theta\right)$ to $\gamma_1\left(\theta\right)$ is a crucial and ubiquitous idea in Monte Carlo.
- In the case where $\gamma_0\left(\theta\right)=p\left(\theta\right)$ and $\gamma_1\left(\theta\right)=p\left(\theta,y\right)$ then we can pick

$$\gamma\left(\theta|\,\alpha\right)=p\left(\theta\right)\left[p\left(y|\,\theta\right)\right]^{\alpha}$$

to move smoothly from the prior to the posterior.

Path Sampling

- The path sampling identity is a limiting case of bridge sampling as $L \longrightarrow \infty$.

Path Sampling

- The path sampling identity is a limiting case of bridge sampling as $L \longrightarrow \infty$.
- It starts from

$$\begin{aligned}
\frac{d \log Z(\alpha)}{d\alpha} &= \frac{d}{d\alpha} \log \int \gamma(\theta|\,\alpha)\, d\theta \\
&= \frac{1}{Z(\alpha)} \int \frac{d}{d\alpha} \gamma(\theta|\,\alpha)\, d\theta \\
&= \frac{1}{Z(\alpha)} \int \frac{d \log \gamma(\theta|\,\alpha)}{d\alpha} \gamma(\theta|\,\alpha)\, d\theta \\
&= \int \frac{d \log \gamma(\theta|\,\alpha)}{d\alpha} \pi(\theta|\,\alpha)\, d\theta
\end{aligned}$$

Path Sampling

- The path sampling identity is a limiting case of bridge sampling as $L \longrightarrow \infty$.
- It starts from

$$\begin{aligned}
\frac{d \log Z(\alpha)}{d\alpha} &= \frac{d}{d\alpha} \log \int \gamma(\theta|\alpha)\, d\theta \\
&= \frac{1}{Z(\alpha)} \int \frac{d}{d\alpha} \gamma(\theta|\alpha)\, d\theta \\
&= \frac{1}{Z(\alpha)} \int \frac{d \log \gamma(\theta|\alpha)}{d\alpha} \gamma(\theta|\alpha)\, d\theta \\
&= \int \frac{d \log \gamma(\theta|\alpha)}{d\alpha} \pi(\theta|\alpha)\, d\theta
\end{aligned}$$

- Integrating from $\alpha = 0$ to 1 then

$$\log \frac{Z(1)}{Z(0)} = \int_0^1 \int \frac{d \log \gamma(\theta|\alpha)}{d\alpha} \pi(\theta|\alpha)\, d\theta d\alpha$$

- Note that this identity is nothing but the famous score identity in statistics; i.e. if we have

$$p\left(y|\,\alpha\right)=\int p\left(x,y|\,\alpha\right)dx$$

then

$$\frac{d\log p\left(y|\,\alpha\right)}{d\alpha}=\int\frac{d\log p\left(x,y|\,\alpha\right)}{d\alpha}p\left(x|\,y,\alpha\right)dx.$$

- Note that this identity is nothing but the famous score identity in statistics; i.e. if we have

$$p(y|\alpha) = \int p(x, y|\alpha)\, dx$$

then

$$\frac{d \log p(y|\alpha)}{d\alpha} = \int \frac{d \log p(x, y|\alpha)}{d\alpha} p(x|y, \alpha)\, dx.$$

- Extension to a multivariate parameter α is straightforward. We introduce

$$\alpha(t) = (\alpha_1(t), ..., \alpha_k(t))$$

where $\gamma(\theta|\alpha(0)) = \gamma_0(\theta)$ and $\gamma(\theta|\alpha(1)) = \gamma_1(\theta)$ then

$$\frac{d \log Z(\alpha(t))}{dt} = \int \frac{d \log \gamma(\theta|\alpha(t))}{dt} \pi(\theta|\alpha(t))\, d\theta$$

where

$$\frac{d \log \gamma(\theta|\alpha(t))}{dt} = \sum_{i=1}^{k} \int \frac{d\alpha_i(t)}{dt} \frac{\partial \log \gamma(\theta|\alpha(t))}{\partial \alpha_i(t)} \pi(\theta|\alpha(t))\, d\theta$$

Practical Implementation

- We first discretize $\alpha \in [0,1]$ using Monte Carlo or a simple grid

$$\log \frac{Z(1)}{Z(0)} \approx L \sum_{i=1}^{L} \int \left. \frac{d \log \gamma(\theta | \alpha)}{d\alpha} \right|_{\alpha = \frac{i}{L}} \pi\left(\theta \middle| \frac{i}{L}\right) d\theta.$$

Practical Implementation

- We first discretize $\alpha \in [0, 1]$ using Monte Carlo or a simple grid

$$\log \frac{Z(1)}{Z(0)} \approx L \sum_{i=1}^{L} \int \left. \frac{d \log \gamma(\theta|\, \alpha)}{d\alpha} \right|_{\alpha = \frac{i}{L}} \pi\left(\theta|\, \frac{i}{L}\right) d\theta.$$

- We typically use MCMC to obtain N samples $\theta_{\frac{i}{L}}^{(j)}$ from $\pi\left(\theta|\, \frac{i}{L}\right)$ for each $i = 1, ..., L$.

Practical Implementation

- We first discretize $\alpha \in [0,1]$ using Monte Carlo or a simple grid

$$\log \frac{Z(1)}{Z(0)} \approx L \sum_{i=1}^{L} \int \left. \frac{d \log \gamma(\theta|\, \alpha)}{d\alpha} \right|_{\alpha=\frac{i}{L}} \pi\left(\theta|\, \frac{i}{L}\right) d\theta.$$

- We typically use MCMC to obtain N samples $\theta_{\frac{i}{L}}^{(j)}$ from $\pi\left(\theta|\, \frac{i}{L}\right)$ for each $i = 1, ..., L$.
- We construct the estimate

$$\widehat{\log \frac{Z(1)}{Z(0)}} = \frac{L}{N} \sum_{i=1}^{L} \sum_{j=1}^{N} \left. \frac{d \log \gamma\left(\theta_{\frac{i}{L}}^{(j)} \middle|\, \frac{i}{L}\right)}{d\alpha} \right|_{\alpha=\frac{i}{L}} .$$

Jarzinsky's identity

- Consider a sequence of distributions to $\pi_n(\theta)$ such that

$$\pi_n(\theta) = \frac{\gamma_n(\theta)}{Z_n}$$

with $\pi_0(\theta)$ a simple distribution (Z_0 known) and $\pi_L(\theta) = \frac{\gamma_L(\theta)}{Z_L}$ is the target.

Jarzinsky's identity

- Consider a sequence of distributions to $\pi_n(\theta)$ such that

$$\pi_n(\theta) = \frac{\gamma_n(\theta)}{Z_n}$$

  with $\pi_0(\theta)$ a simple distribution (Z_0 known) and $\pi_L(\theta) = \frac{\gamma_L(\theta)}{Z_L}$ is the target.

- Introduce a sequence of MCMC transition kernels such that

$$\int \pi_n(\theta) K_n(\theta, \theta') \, d\theta = \pi_n(\theta').$$

Jarzinsky's identity

- Consider a sequence of distributions to $\pi_n(\theta)$ such that

$$\pi_n(\theta) = \frac{\gamma_n(\theta)}{Z_n}$$

with $\pi_0(\theta)$ a simple distribution (Z_0 known) and $\pi_L(\theta) = \frac{\gamma_L(\theta)}{Z_L}$ is the target.

- Introduce a sequence of MCMC transition kernels such that

$$\int \pi_n(\theta) K_n(\theta, \theta') \, d\theta = \pi_n(\theta').$$

- Jarzinsky's identity states that

$$\frac{Z_L}{Z_0} = \int \left(\prod_{n=1}^{L} \frac{\gamma_n(\theta_{n-1})}{\gamma_{n-1}(\theta_{n-1})} \right) \pi_0(\theta_0) \prod_{n=1}^{L} K_n(\theta_{n-1}, \theta_n) \, d\theta_{0:n}$$

- So if $\theta_{0:n}^{(i)} \sim \pi_0(\theta_0) \prod_{n=1}^{L} K_n(\theta_{n-1}, \theta_n)$ then

$$\frac{\widehat{Z_L}}{Z_0} = \frac{1}{N} \sum_{i=1}^{N} \prod_{n=1}^{L} \frac{\gamma_n\left(\theta_{n-1}^{(i)}\right)}{\gamma_{n-1}\left(\theta_{n-1}^{(i)}\right)}.$$

- So if $\theta_{0:n}^{(i)} \sim \pi_0(\theta_0) \prod_{n=1}^{L} K_n(\theta_{n-1}, \theta_n)$ then

$$\frac{\widehat{Z_L}}{Z_0} = \frac{1}{N} \sum_{i=1}^{N} \prod_{n=1}^{L} \frac{\gamma_n\left(\theta_{n-1}^{(i)}\right)}{\gamma_{n-1}\left(\theta_{n-1}^{(i)}\right)}.$$

- This equality is very powerful and shows that it is possible to estimate unbiasedly Z_L / Z_0 using non-homogeneous Markov chain simulation.

- So if $\theta_{0:n}^{(i)} \sim \pi_0(\theta_0) \prod_{n=1}^{L} K_n(\theta_{n-1}, \theta_n)$ then

$$\widehat{\frac{Z_L}{Z_0}} = \frac{1}{N} \sum_{i=1}^{N} \prod_{n=1}^{L} \frac{\gamma_n\left(\theta_{n-1}^{(i)}\right)}{\gamma_{n-1}\left(\theta_{n-1}^{(i)}\right)}.$$

- This equality is very powerful and shows that it is possible to estimate unbiasedly Z_L / Z_0 using non-homogeneous Markov chain simulation.
- This has had a major impact in statistical physics since its introduction in 1997.

- Proof of Jarzinsky's inequality: We introduce a probability distribution

$$\pi_L(\theta_L)\prod_{n=0}^{L-1}L_n(\theta_{n+1},\theta_n)$$

then

$$\frac{Z_L}{Z_0}=\int\frac{\gamma_L(\theta_L)\prod_{n=0}^{L-1}L_n(\theta_{n+1},\theta_n)}{\gamma_0(\theta_0)\prod_{n=1}^{L}K_n(\theta_{n-1},\theta_n)}.\pi_0(\theta_0)\prod_{n=1}^{L}K_n(\theta_{n-1},\theta_n)\,d\theta_{0:n}$$

- Proof of Jarzinsky's inequality: We introduce a probability distribution

$$\pi_L(\theta_L) \prod_{n=0}^{L-1} L_n(\theta_{n+1}, \theta_n)$$

then

$$\frac{Z_L}{Z_0} = \int \frac{\gamma_L(\theta_L) \prod_{n=0}^{L-1} L_n(\theta_{n+1}, \theta_n)}{\gamma_0(\theta_0) \prod_{n=1}^{L} K_n(\theta_{n-1}, \theta_n)} . \pi_0(\theta_0) \prod_{n=1}^{L} K_n(\theta_{n-1}, \theta_n) \, d\theta_{0:n}$$

- If

$$L_{n-1}(\theta_n, \theta_{n-1}) = \frac{\pi_n(\theta_{n-1}) K_n(\theta_{n-1}, \theta_n)}{\pi_n(\theta_n)}$$

then Jarzynski's equality follows. $L_{n-1}(\theta_n, \theta_{n-1})$ is the time-reversal kernel associated to $K_n(\theta_{n-1}, \theta_n)$.

- Proof of Jarzinsky's inequality: We introduce a probability distribution

$$\pi_L(\theta_L)\prod_{n=0}^{L-1}L_n(\theta_{n+1},\theta_n)$$

then

$$\frac{Z_L}{Z_0}=\int\frac{\gamma_L(\theta_L)\prod_{n=0}^{L-1}L_n(\theta_{n+1},\theta_n)}{\gamma_0(\theta_0)\prod_{n=1}^{L}K_n(\theta_{n-1},\theta_n)}.\pi_0(\theta_0)\prod_{n=1}^{L}K_n(\theta_{n-1},\theta_n)\,d\theta_{0:n}$$

- If

$$L_{n-1}(\theta_n,\theta_{n-1})=\frac{\pi_n(\theta_{n-1})K_n(\theta_{n-1},\theta_n)}{\pi_n(\theta_n)}$$

then Jarzynski's equality follows. $L_{n-1}(\theta_n,\theta_{n-1})$ is the time-reversal kernel associated to $K_n(\theta_{n-1},\theta_n)$.

- Note that if K_n is π_n-reversible then

$$\frac{\pi_n(\theta_{n-1})K_n(\theta_{n-1},\theta_n)}{\pi_n(\theta_n)}=K_n(\theta_n,\theta_{n-1}).$$

- Advantages of Jarzynski's inequality over path sampling

- Advantages of Jarzynski's inequality over path sampling
  - No need to run a lot of MCMC chains until equilibrium.

- Advantages of Jarzynski's inequality over path sampling
  - No need to run a lot of MCMC chains until equilibrium.
  - Simple importance sampling method which can be parallelized.

- Advantages of Jarzynski's inequality over path sampling
    - No need to run a lot of MCMC chains until equilibrium.
    - Simple importance sampling method which can be parallelized.
- Drawbacks

- Advantages of Jarzynski's inequality over path sampling
    - No need to run a lot of MCMC chains until equilibrium.
    - Simple importance sampling method which can be parallelized.
- Drawbacks
    - It is just importance sampling and the variance will be huge if the sequence of distributions is not carefully selected.

- Advantages of Jarzynski's inequality over path sampling
    - No need to run a lot of MCMC chains until equilibrium.
    - Simple importance sampling method which can be parallelized.
- Drawbacks
    - It is just importance sampling and the variance will be huge if the sequence of distributions is not carefully selected.
    - Selecting $L_{n-1}(\theta_n, \theta_{n-1})$ as the time reversal kernel is computationally convenient but far from optimal.

Application to Mixture Models

- Consider 100 data

$$Y_i \sim \sum_{k=1\omega i}^{4} \mathcal{N}\left(\mu_i, \sigma_i^2\right)$$

Application to Mixture Models

- Consider 100 data

$$Y_i \sim \sum_{k=1\omega i}^{4} \mathcal{N}\left(\mu_i, \sigma_i^2\right)$$

- We set (conditionally) conjugate priors on $\omega_{1:4}, \mu_{1:4}, \sigma_{1:4}^2$

$$\begin{aligned} \omega_{1:4} &\sim \mathcal{D}\left(1,1,1,1\right), \\ \mu_j &\sim \mathcal{N}\left(\xi, \kappa^{-1}\right),\ \lambda_j \sim \mathcal{G}a(\nu, \chi). \end{aligned}$$

Application to Mixture Models

- Consider 100 data

$$Y_i \sim \sum_{k=1\omega i}^{4} \mathcal{N}\left(\mu_i, \sigma_i^2\right)$$

- We set (conditionally) conjugate priors on $\omega_{1:4}, \mu_{1:4}, \sigma_{1:4}^2$

$$\begin{aligned} \omega_{1:4} &\sim \mathcal{D}\left(1,1,1,1\right), \\ \mu_j &\sim \mathcal{N}\left(\xi, \kappa^{-1}\right),\ \lambda_j \sim \mathcal{G}a(\nu, \chi). \end{aligned}$$

- We consider

$$\begin{aligned} \pi_n\left(\omega_{1:4}, \mu_{1:4}, \sigma_{1:4}^2 \middle| y_{1:100}\right) &\propto \left[f\left(y_{1:100} \middle| \omega_{1:4}, \mu_{1:4}, \sigma_{1:4}^2\right)\right]^{\phi_n} \\ &\quad \times \pi\left(\omega_{1:4}, \mu_{1:4}, \sigma_{1:4}^2\right). \end{aligned}$$

where $0 \leq \phi_1 < \cdots < \phi_p = 1$ are tempering parameters.

- We simulated 100 data with weights 0.25, means (-3,0,3,6) and standard deviations 0.55.

- We simulated 100 data with weights 0.25, means (-3,0,3,6) and standard deviations 0.55.
- The posterior admits 4! well-separated modes

- We simulated 100 data with weights 0.25, means (-3,0,3,6) and standard deviations 0.55.
- The posterior admits 4! well-separated modes
- We use a MCMC kernel K_n with invariant distribution π_n and the time reversal backward kernel.

- We simulated 100 data with weights 0.25, means (-3,0,3,6) and standard deviations 0.55.
- The posterior admits 4! well-separated modes
- We use a MCMC kernel K_n with invariant distribution π_n and the time reversal backward kernel.
- The MCMC kernel K_n is a composition of the following update steps:

- We simulated 100 data with weights 0.25, means (-3,0,3,6) and standard deviations 0.55.
- The posterior admits 4! well-separated modes
- We use a MCMC kernel K_n with invariant distribution π_n and the time reversal backward kernel.
- The MCMC kernel K_n is a composition of the following update steps:
    - Update $\mu_{1:r}$ via a MH kernel with additive normal random walk proposal.

- We simulated 100 data with weights 0.25, means (-3,0,3,6) and standard deviations 0.55.
- The posterior admits 4! well-separated modes
- We use a MCMC kernel K_n with invariant distribution π_n and the time reversal backward kernel.
- The MCMC kernel K_n is a composition of the following update steps:
  - Update $\mu_{1:r}$ via a MH kernel with additive normal random walk proposal.
  - Update $\lambda_{1:r}$ via a MH kernel with multiplicative log-normal random walk proposal.

- We simulated 100 data with weights 0.25, means (-3,0,3,6) and standard deviations 0.55.
- The posterior admits 4! well-separated modes
- We use a MCMC kernel K_n with invariant distribution π_n and the time reversal backward kernel.
- The MCMC kernel K_n is a composition of the following update steps:
  - Update $\mu_{1:r}$ via a MH kernel with additive normal random walk proposal.
  - Update $\lambda_{1:r}$ via a MH kernel with multiplicative log-normal random walk proposal.
  - Update $\omega_{1:r}$ via a MH kernel with additive normal random walk proposal on the logit scale.

- We ran the algorithm with $N = 1000$ particles for $p =$50, 100, 200, 500 and 1000 time steps with 1 and 10 MCMC iterations per time step.

- We ran the algorithm with $N = 1000$ particles for p =50, 100, 200, 500 and 1000 time steps with 1 and 10 MCMC iterations per time step.
- We selected a piecewise linear cooling schedule $\{\phi_n\}$. Over 1000 time steps, the sequence increased uniformly from 0 to 15/100 for the first 200 time points then from 15/100 to 40/100 for the next 400 and finally from 40/100 to 1 for the last 400 time points. The other time specifications had the same proportion of time attributed to the tempering parameter setting.

- We ran the algorithm with $N = 1000$ particles for p =50, 100, 200, 500 and 1000 time steps with 1 and 10 MCMC iterations per time step.
- We selected a piecewise linear cooling schedule $\{\phi_n\}$. Over 1000 time steps, the sequence increased uniformly from 0 to 15/100 for the first 200 time points then from 15/100 to 40/100 for the next 400 and finally from 40/100 to 1 for the last 400 time points. The other time specifications had the same proportion of time attributed to the tempering parameter setting.
- Additional simulations with resampling

Sampler Details	Iterations per time step	
AIS (50 time steps)	1	10
Avg. Log Posterior	-191.07	-166.73
Avg. Log Normalizing Constant	-249.04	-242.07
AIS (100 time steps)	1	10
Avg. Log Posterior	-180.76	-162.37
Avg. Log Normalizing Constant	-250.22	-244.17
AIS (200 time steps)	1	10
Avg. Log Posterior	-174.40	-160.00
Avg. Log Normalizing Constant	-247.45	-245.92

Sampler Details	Iterations per time step	
AIS (500 time steps)	1	10
Avg. Log Posterior	-167.67	-157.06
Avg. Log Normalizing Constant	-247.30	-247.94
AIS (1000 time steps)	1	10
Avg. Log Posterior	-163.14	-155.31
Avg. Log Normalizing Constant	-247.50	-247.36