Lecture 3
• Importance sampling:
\[ \hat{\theta} = \sum_{j=1}^{J} \frac{1}{\sum_{j' \neq j} \pi(X_{j'})} \]
• The set \( \{X, \omega \} \) is said “properly weighted with respect to \( \pi \) if for any \( h(X) \), \( \theta = E_{\pi} [h(X)] \) can be approximated as
\[ \hat{\theta} = \sum_{j=1}^{J} \frac{1}{\sum_{j'} \pi(X_{j'})} \]

Example: inferring beta distribution
• Observations \( x_1, \ldots, x_n \) from \( \text{Beta}(\alpha, \beta) \). How to estimate the unknown parameters?
  – Likelihood function:
  \[ p(\text{data} | \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} \prod_{i=1}^{n} \alpha^x_i \beta^{1-x_i} \]
  \[ = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} \exp \left[ \alpha \sum_{i=1}^{n} \log x_i + \beta \sum_{i=1}^{n} \log(1-x_i) \right] \]
  – Bayesian approach?

Many problems can be turned into missing data problems
• Hierarchical Model/Random Effects Models
  \[ Y_1 \sim \tau_1 \]
  \[ Y_2 \sim \tau_2 \]
  \[ \vdots \]
  \[ Y_n \sim \tau_n \]
  \[ - [y_1] \sim N(\tau_1, 1); \]
  \[ - \tau_2, \ldots, \tau_n \text{ (missing data)} \]
  \[ \text{are iid } N(\alpha, \sigma^2) \]
• Latent Class/Mixture Models
  Example: \( [y_1 | \lambda = 1] \sim N(\theta_1, \sigma^2) \)
  \[ \text{and } [y_1 | \lambda = 0] \sim N(\theta_2, \sigma^2) \]
  \[ P(\lambda = 1) = \pi \]
  Can’t observe the random effects of the indicators!!

A missing data problem
• We observe \( \{X, y \} \sim N(\theta, \Sigma) \)
  – Interested in the posterior of \( \Sigma \);
  – “Easy” with complete data
  \[ p(\Sigma | \text{data}) = N(\Sigma | \text{data}) \]
  \[ \text{inverse wishart}. \] Use Odell & Feiveson (1966)
  – Imcomplete data?
  \[ \left( \begin{array}{c} 1 \quad 1 \\ 1 \quad 2 \end{array} \right) \left( \begin{array}{c} 1 \quad 1 \\ 1 \quad 2 \end{array} \right) \]

More About Missing Data Problems
• Complete-Data Model: \( [\theta | y] = f(\theta | y) \)
  – Example 1: \( x \sim \text{multivariate normal}, \theta \sim \text{covariance} \);
  – Example 2: \( x \sim \text{multinomial}, \theta \sim \text{unknown frequency} \);
  – Example 3: \( x \sim \text{Markov chain}, \theta \sim \text{transition probabilities} \);
• Sometimes only part of \( x \), say \( y \), can be observed.
  Write the decomposition as \( x \sim (y, z) \),
• Observed-Data Model: \( [\theta | y] = f(\theta | y) = \int f(y, z | \theta) dz \)
• Question: How to find the MLE?

Importance Sampling for MDP
• Sampling from some “trial distribution”
  \[ g(\theta, y_{\text{new}}) = g_1(\theta) g_2(y_{\text{new}} | \theta) \]
  e.g. \( g_2(y_{\text{new}} | \theta) = f(y_{\text{new}} | y_{\text{old}}, \theta) \)
  – The hierarchical model example:
  \[ Y_1 \sim \tau_1 \]
  \[ Y_2 \sim \tau_2 \]
  \[ \vdots \]
  \[ Y_n \sim \tau_n \]
  Prior: \( p(\alpha, \sigma) = 1 \)
  \[ \text{Draw } (\alpha, \sigma) \text{ from a “trial”} \]
  \[ \text{Then draw } \tau_i \ text{s}. \]
  \[ \text{Computing importance wgt} \]
  \[ \hat{p} = \prod_{i=1}^{n} \left[ \frac{1}{\alpha} \right] \sim \exp \left[ \sum_{i=1}^{n} \left( \frac{\alpha - \tau_i}{2\sigma^2} \right) \right] \]
Adaptive Importance Sampling

For example, one can start with trial $q_i(x) \sim \pi(\mu_i, \Sigma_i)$ and use IS to estimate $E(X)$ and $\text{var}(X)$, called $\mu_1$ and $\Sigma_1$; then get $q_i(x) \sim t(\mu_i, \Sigma_i)$; iterate till convergence.

Or one can start with a mixture of t-distributions. See Oh and Berger (1993, JASA).

The sequential importance sampling

- Suppose $X=(x_1, \ldots, x_d)$ of interest is $\pi(x)=\pi(x_1|x_2, \ldots, x_d)\cdots\pi(x_d)$.
- Similarly, the trial density can also be decomposed $q(X)=q_1(x_1)q(x_2|x_1)\cdots q(x_d|x_1, \ldots, x_{d-1})$.

- The weight can be computed sequentially.
- At any stage we may want to reject a partial sample because the “temporary” weight is already very small.
- In practice, we create a sequence of approximations, $\pi_1(x_1), \pi_2(x_1,x_2), \ldots$ to guide the IS at each stage.

Bias correction: importance sampling

- Sampling distribution of $X_N=(x_1, x_1, \ldots, x_N)$
  
  $p(X_N)=\frac{1}{k_1(x_1)}\frac{1}{k_2(x_0,x_1)}\cdots\frac{1}{k_d(x_{d-1},x_d)}\nonumber$
  
  $=\psi(x_1|x_0)\psi(x_2|x_1)\psi(x_3|x_2)\cdots\psi(x_N|x_{N-1})$

  Where $k_j$ is the number available neighbor of $x_j$.

- Target distribution: $\pi_N(X_N)$ is constant

- Weight function:

  $w(X_N)=k_1\times\cdots\times k_N=Z_N\pi_N(X_N)$

  Total number of SAWs

Reconsidering Rejection Sampling

- What can we do if the envelop constant is not well chosen?

  - We can accept and then correct the bias

  
  $p(x|\text{accept})=\begin{cases} \pi_N(x)/q(x) & \text{if } x \in A \\ q(x)/\pi_N(x) & \text{if } x \notin A \end{cases}$

  So the accepted $x$’s should be weighted by

  $w(x)=c\pi_N(x)/q(x)$

  The accepted sample has a smaller cv$^2$ than using $q(x)$ as a trial density.

  However, this by itself is not too useful unless we are estimating $E_{\pi_N}h(x)$, where evaluating $h(\cdot)$ is expensive. But this becomes more useful in a sequential setting.

Polymer Simulation

- Self-Avoid Random Walk

  Interested in under uniform distribution

  $E_x[R^2]=0.917N^\beta, Z_N=N^\beta\times2.638^N$

- Add one new position at a time to one of the $k$ (IS) available positions.

  \begin{itemize}
  \item Question: sampling distribution of the chain?
  \item $p(x)$ (straight chain) $=\frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2}$
  \end{itemize}

  \begin{itemize}
  \item $p(x)$ (diagonal chain) $=\frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2}$
  \end{itemize}

  Where $k$ is the number available neighbor.

  For $d=3$, $\nu=0.588$; $d=4$, $\nu=0.5$.

  Hammersley & Morton (1954)

  Rosenbluth & Rosenbluth (1955)

Summary and Generalization

- Target: simulating from $\pi_N(X_N)$ is constant

- Strategy: growing the molecule sequentially and modifying via importance weight:

  $p(X_N)=\psi(x_1|x_0)\psi(x_2|x_1)\times\cdots\psi(x_N|x_{N-1})$

  $w(X_N)=\psi(x_1|x_0)\psi(x_2|x_1)\times\cdots\psi(x_N|x_{N-1})^{-1}$

  $E[w(X_N)]=Z_N^{-1}$

  Partition function

- But the attrition problem is still serious and the weight distribution is too skewed.
More generally ...

The polymer may be made of different types of monomers and these monomers may interact, which gives rise to an “energy function” $U(X_N)$

Boltzmann distribution:

$$\pi(X_N) \propto \exp\left\{-\frac{U(X_N)}{T}\right\}$$

Improvements

- Enrichment method (1957): using partially generated chains multiple times.
- Better way to add monomer (1970s) --- more forward looking.
- Pruning (1997).

Example: Energy minimization

- Generalize the SAW model to include an energy function:
  $$\pi(X_N) \propto \exp\left\{-\frac{U(X_N)}{T}\right\}$$
- Apply the SIS to the problem:

  $U(X_N) = \sum_{i=1}^{N-1} c[i(i), i(j)]$

  e.g., $c[a, b] = 1$ if and only if $a = b = B$