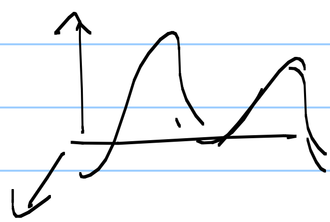
Lecture 6

Note Title 4/2/2006

Example: From Book.

$$f(x,y) = 0.5\, e^{-90(x-0.5)^2 - 45(y+0.1)^4} + e^{-45(x+0.4)^2 - 60(y-0.5)^2}$$

in domain $[-1,1] \times [-1,1]$



Two peaks $(0.5, -0.1)$ & $(-0.4, 0.5)$

Messy (multimodal) distribution & Truncated domain.

Approximate by $g(x,y)$ a truncated mixture of Gaussians Centered on the two peaks.

$$g(x,y) \propto I_{(x,y)\in[-1,1]\times[-1,1]} \cdot \{ (0.46)\, N(\underline{\mu}_1, \underline{\underline{\Sigma}}_1) + (0.54)\, N(\underline{\mu}_2, \underline{\underline{\Sigma}}_2) \}$$

$\underline{\mu}_1 = (0.5, -0.1)$

$\underline{\mu}_2 = (-0.4, 0.5)$

$$\underline{\underline{\Sigma}}_1 = \begin{bmatrix} 1/180 & 0 \\ 0 & 1/90 \end{bmatrix} \quad \underline{\underline{\Sigma}}_2 = \begin{bmatrix} 1/90 & 0 \\ 0 & 1/120 \end{bmatrix}$$

Sample as follows:

Draw sample from $N(\underline{\mu}_1, \underline{\underline{\Sigma}}_1)$ with prob 0.46

Otherwise sample from $N(\underline{\mu}_2, \underline{\underline{\Sigma}}_2)$ (with prob 0.54.

Reject sample $(\omega = 0)$ if it falls outside range $[-1,1] \times [-1,1]$.

(Page 2)

Importance Sampling can be super-efficient.
The variance of $\hat{\mu}_m$ can be smaller than that obtained by sampling from $\pi(x)$

i.e. variance $= 0$ if $g(x) \propto \pi(x)h(x)$ and normalization constant known — but, if so, no need to sample.

The Rao-Blackwellization method can be generalized to importance sampling.

Thm. Let $f(z_1, z_2)$ & $g(z_1, z_2)$ be two distributions where the support of f is a subset of the support of g. Then.

$$\operatorname{Var}_g \left\{ \frac{f(z_1, z_2)}{g(z_1, z_2)} \right\} \geq \operatorname{Var}_g \left\{ \frac{f_1(z_1)}{g_1(z_1)} \right\},$$

where $f_1(z_1) = \int f(z_1, z_2)\, dz_2$ & $g_1(z_1) = \int g(z_1, z_2)\, dz_2$ are marginals

Moral, in Monte Carlo computation you should do as much as possible analytically (e.g. marginalization)

(Page 3)

Rule of Thumb for Importance Sampling.

Effective Sample Size

$$ESS(m) = \frac{m}{1 + var_g\{\omega(x)\}}$$

$\omega(x) = \frac{\pi(x)}{g(x)}$

If $\pi(x) = g(x)$
$ESS(m) = 1.$

Claim: $$\frac{var_\pi\{h(x)\}}{var_g\{h(x)\omega(x)\}} \simeq \frac{1}{1 + var_g\{\omega(x)\}}$$

only an approximation for $\pi(x) \approx g(x)$.

This claim, when true, implies ~~that~~ you can estimate the effectiveness of a sampling distribution $g(x)$ independent of $h(x)$. (Useful, but limited).

$ESS(m)$ gives a measure of how different the sampling distribution is from the target.

$var_g\{\omega(x)\}$ can be measured by the coefficient of variation (c.v.)

$$cv^2(\omega) = \sum_{j=1}^{m} \frac{(\omega^j - \bar{\omega})^2}{(m-1)\bar{\omega}^2}.$$

$\bar{\omega} = \frac{1}{m} \sum_{j=1}^{m} \omega_j$

$\omega(x) = \frac{\ell(x)}{g(x)}$

Because if we only know $c\pi(x) = \ell(x)$ unnormalized then $\frac{1}{(m-1)} \sum_{j=1}^{m} (\omega^j - \bar{\omega})^2 \sim c^2 Var_\pi\{\omega\}$

$\bar{\omega}^2 \sim c^2$

(Page 4)

Adaptive Importance Sampling.

It is good to try to learn as much as possible about the target distribution

Simple way, assume a t-distribution a trial distribution

$$g_0(x) = t_\alpha(x;\mu,\sigma^2) = \frac{1}{(2\sigma^2)^{1/2}} \frac{\Gamma((\alpha+1)/2)}{(\alpha\pi)^{1/2}\Gamma(\alpha/2)} \left\{1+\frac{1}{2\alpha\sigma^2}(y-\mu)^2\right\}^{-\frac{(\alpha+1)}{2}}$$

t-distribution falls off slower than the Gaussan.

Use weighted sampling to estimate the mean & covariance of the target distribution μ_1, σ_1^2

Then use a new trial

$$g_1(x) = t_\alpha(x, \mu_1, \sigma_1^2)$$

For any statistic $\phi(x)$ (eg. $\phi(x) = x, x^2, ..$)

Estimate ϕ for π by

$$\frac{\frac{1}{m}\sum_{i=1}^{m}\phi(x^i)\,\omega(x^i)}{\frac{1}{m}\sum_{i=1}^{m}\omega(x^i)}.$$

Assumed that $h(x)$ is complicated function, could to evaluate it as rarely as possible.

(Page 5)

Adaptive Importance Sampling.

Parametric form. $g_0(x)$ — need $g_1(x) = g(x, \lambda)$
for some λ

Pick $g_1(x)$ to minimize
$$var_{g_1}(\omega) = \sum_x \frac{\pi^2(x)}{g_1(x) g_0(x)} g_0(x) - 1.$$
as function of λ.

Estimate $var_{g_1}(\omega)$ by the coeff of variation
$$c\hat{v}^2(\lambda) = \hat{H}(\lambda) - 1, \quad \hat{H}(\lambda) = \frac{1}{m} \sum_{i=1}^{m} \frac{\pi^2(x^{(i)})}{g_1(x^{(i)}) g_0(x^{i})}$$
samples x^i from $g_0(x)$.

If normalization of $\pi(x)$ is unknown.
$$c\hat{v}^2(\lambda) = \frac{\hat{H}(\lambda)}{\bar{w}_0^2} - 1, \quad \bar{w}_0 = \frac{1}{m} \sum_{i=1}^{m} \frac{\pi(x^{(i)})}{g_0(x^{(i)})}.$$

Problem: Adaptive methods risk being unstable....

(page 6)

Justification of Claim:

$$\mathrm{var}_\pi\{h\}\{1+\mathrm{var}_g(\omega)\} \simeq \mathrm{var}_g\{h\omega\}$$

$g(x)$, $h(x)$ $\quad \omega(x) = \frac{\pi(x)}{g(x)}$

$$\underline{\text{L.H.S}} = \sum_x g(x)\, h^2(x) \left(\frac{\pi(x)}{g(x)}\right)^2 - \left(\sum_x g(x)\, h(x) \frac{\pi(x)}{g(x)}\right)^2$$

$$\underline{\text{LHS}} = \sum_x \left(\frac{\pi(x)^2}{g(x)}\right) h^2(x) - \left\{\sum_x \pi(x)\, h(x)\right\}^2 .$$

$$\underline{\text{R.H.S}} = \left(\sum_x \pi(x)\, h^2(x) - \left\{\sum_x \pi(x) h(x)\right\}^2\right) \times \left\{1 + \sum_x g(x) \frac{\pi^2(x)}{g^2(x)} - \left(\sum_x g(x) \frac{\pi(x)}{g(x)}\right)^2\right\}$$

$$\underline{\text{R.H.S}} = \left(\sum_x \pi(x) h^2(x) - \left\{\sum_x \pi(x)\, h(x)\right\}^2\right) \sum_x \frac{\pi^2(x)}{g(x)} .$$

For equality $\quad$ LHS = RHS is equivalent to

$$\sum_x \frac{\pi^2(x) h^2(x)}{g(x)} - \left(\sum_x \pi(x)\, h(x)\right)^2 \sum_x \frac{\pi^2(x)}{g(x)} = \sum_x \pi(x)\, h(x) \left\{1 - \sum_x \frac{\pi^2(x)}{g(x)}\right\}$$

Equality holds, if $\pi(x) = g(x)$.

Approximate holds if we do expansion in $g(x)/\pi(x)$. $\underline{\text{But}}$ can be violated easily