Scaling issues w/ rejection sampling

eg from 'Learning in Graphical Models'

We need an envelope (or 'proposal') q to match the entire jt. distribution f.

Even if we have q s.t. $\sup\{h/q\} \le K < \infty$

Finding K may be v. hard

K may be v. large $\Rightarrow$ alg. may be very inefficient

Toy E.g. Sample M-dim. Gaussian w/ mean $\underset{\sim}{0}$.

Want $X \sim N_M(\underset{\sim}{0}, \sigma_f^2 I)$ $\leftarrow f$ (target)

Say q is M-dim. Gaussian w/ mean $\underset{\sim}{0}$.

'Propose' $Y \sim N_M(\underset{\sim}{0}, \sigma_q^2 I)$ $\leftarrow q$ (proposal)

Suppose $\sigma_q = 1.01\,\sigma_f$ so proposal(q) is heavier-tailed than target (f).

Smallest K satisfying $\sup_x \frac{f(x)}{q(x)} \le K$

$$\sup_x \frac{f(x)}{q(x)} \overset{\text{maximized at origin}}{=} \frac{(2\pi\sigma_f^2)^{-M/2}}{(2\pi\sigma_q^2)^{-M/2}} = \left(\frac{\sigma_q}{\sigma_f}\right)^M = \exp\{M\log(1.01)\}$$

If $M = 1000$, $K > 20{,}000$

$\Rightarrow$ acc. rate $< 1/20{,}000$

K grows exponentially w/ M.

[Of course, we would sample these r.v.s individually in practice: they are independent.]

Aside: generating multivariate normals $N_p(\underset{\sim}{\mu}, \Sigma)$.

Recall: generating univariate normal: $x \sim N(0,1)$

Then, $\sigma x + \mu \sim N(\mu, \sigma^2)$

Multivariate normal: $\underset{\sim}{x} \sim N_p(\underset{\sim}{0}, I_{N \times N})$ (p i.i.d. $N(0,1)$)

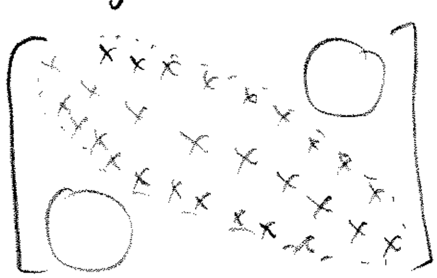
Find C, 'Choleski factor' of Σ s.t. $CC^T = \Sigma$.

$C \underset{\sim}{x} \sim N_p(0, C I C^T) = N_p(0, \Sigma)$

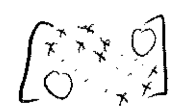
$C \underset{\sim}{x} + \underset{\sim}{\mu} \sim N_p(\underset{\sim}{\mu}, \Sigma)$.

Choleski factoring/decomposition is computationally expensive, of order $p^3/3$ flops (floating point operations).

For large p: if matrix has special structure, use it. For e.g. band matrix



Common in inverse cov. of Markov random field models

E.g. AR-1, $\Sigma^{-1} =$ 

Choleski takes $p b_w^2$ flops where b_w = bandwidth

= upper bw (# non-zero diags above diagonal)

+ lower " (" " " below)

+ 1

mvrnorm

M.C. s.errors

Assessing accuracy of estimates

$X_1, \ldots, X_n \overset{iid}{\sim} f \qquad \hat{\mu}_n = \frac{\sum_{i=1}^n g(X_i)}{n}$

From CLT before:

$$\hat{\mu}_n \approx N(\mu, \sigma_g^2/n) \quad \text{for large } n.$$

where $\sigma_g^2 = Var_f\, g(x)$

Easy to estimate σ_g^2: compute sample variance based on $g(X_1), \ldots, g(X_n)$.

$$\hat{\sigma}_g^2 = \frac{1}{n-1} \sum_{i=1}^n \left(g(X_i) - \hat{\mu}_n\right)^2$$

Along w/ our estimates, report our estimated M.C. s.error

$= \hat{\sigma}_g/\sqrt{n}$.

Standard, sensible frequentist statistics!

Importance Sampling

Not a method for generating samples from f.
Instead estimate expectations w.r.t. f by reweighting samples from another distr. q (importance function).

Basic idea: $\mu = \int g(x) f(x)\, dx = E_f\{g(x)\}$

Before: $\hat{\mu}_n = \dfrac{\sum_{i=1}^n g(X_i)}{n}$ when $X_1, \ldots, X_n \overset{iid}{\sim} f$

Now if $\mu = \int \left\{ g(x) \dfrac{f(x)}{q(x)} \right\} q(x)\, dx$ ~~(Assume $q(x)=0 \Rightarrow f(x)=0$)~~ if $q(x) > 0$ whenever $g(x) f(x) \neq 0$

$= E_q\left\{ g(x) \dfrac{f(x)}{q(x)} \right\}$

Imp. sampling estimate, $\hat{\mu}_n = \dfrac{\sum_{i=1}^n g(Y_i) \frac{f(Y_i)}{q(Y_i)}}{n}$ when $Y_1, \ldots, Y_n \overset{iid}{\sim} q$

$\hat{\mu}_n \to \mu$ (as before by SLLN if $E_f|g| < \infty$)

Problem: we are assuming normalizing constant is known.

More general <u>ratio imp. sampling</u> estimate: assuming $q(x) > 0$ wherever $f(x) > 0$

Note that $E_q\left\{ \dfrac{f(x)}{q(x)} \right\} = \int \dfrac{f(x)}{q(x)} q(x)\, dx = 1.$

So, $\mu = E_q\left\{ g(x) \dfrac{f(x)}{q(x)} \right\} \Big/ E_q\left\{ \dfrac{f(x)}{q(x)} \right\}$

$= E_q\left\{ g(x) \dfrac{h(x)}{q(x)} \right\} \Big/ E_q\left\{ \dfrac{h(x)}{q(x)} \right\}$ where $\dfrac{h(x)}{c} = f(x)$

Suggesting the estimator: $Y_i \overset{iid}{\sim} q$

$$\tilde{\mu}_n = \frac{\sum_{i=1}^n g(Y_i) \frac{h(Y_i)}{q(Y_i)}}{\sum_{i=1}^n \frac{h(Y_i)}{q(Y_i)}} \longrightarrow \mu$$

SLLN + a version of Slutsky's Thm.

$\tilde{\mu}_n$ has a small bias (unlike $\hat{\mu}_n$) but it has a lower MSE (J Liu pg. 33)

Note that $\tilde{\mu}_n = \sum_{i=1}^{n} g(Y_i) \underbrace{\left[\frac{h(Y_i)/q(Y_i)}{\sum_{i=1}^{n} h(Y_i)/q(Y_i)} \right]}_{\text{'normalized' weights}}$

$= \sum_{i=1}^{n} g(Y_i)\, \tilde{w}(Y_i)$

where $\tilde{w}(Y_i) = \frac{h(Y_i)/q(Y_i)}{\sum_{i=1}^{n} h(Y_i)/q(Y_i)}$

In practice, $\tilde{w}(Y_i)$ is numerically more stable than unnormalized version.

Note that numerator and denominator can be estimated using different sets of samples.

$$\mu = \frac{E_{q_1}\left(g(x)\frac{h(x)}{q_1(x)}\right)}{E_{q_2}\left(\frac{h(x)}{q_2(x)}\right)}$$

as long as $q_1(x) > 0$ whenever $g(x)h(x) \neq 0$
and $q_2(x) > 0$ " $h(x) > 0$

Can estimate μ from $Y_1, \ldots, Y_n \overset{iid}{\sim} q_1$
$W_1, \ldots, W_n \overset{iid}{\sim} q_2$

Useful when $g(x)h(x) \neq 0$ on a very different region than where $h(x) > 0$. e.g. when estimating tail probabilities.

Another use of imp. sampling:

Can use single sample (set of samples) to compute expectations w.r.t. many different distr. by reweighting samples differently each time.

Suppose we have a parametric family $\{f_\theta : \theta \in \Theta\}$ and we want

$$\mu(\theta) = E_{f_\theta}(g(x)) = \int g(x) f_\theta(x)\,dx \quad \text{for } \theta \in \Theta$$

E.g. $f_\theta(x)$ is normal density and $\theta = (\mu, \sigma^2)$

And, We want $E(X^3)$ for different combinations of μ, σ^2.

How this may be useful: If we want to find θ that maximizes $E_{f_\theta}(g(x))$.

Note: also useful for Monte Carlo maximum likelihood.

Naive M.C. assuming we can sample from f_θ $\forall \theta \in \Theta$.

$X_1^{(\theta_1)}, \ldots, X_{n_1}^{(\theta_1)} \sim f_{\theta_1}$ to estimate $E_{f_{\theta_1}}(g(x))$

$\vdots$

$X_1^{(\theta_k)}, \ldots, X_{n_k}^{(\theta_k)} \sim f_{\theta_k}$ $\quad$ " $\quad$ " $\quad$ $E_{f_{\theta_k}}(g(x))$

Too many samples to generate especially if k is large, Θ is highly multivariate.

However, suppose we find q s.t.

$$f_\theta(x) > 0 \Rightarrow q(x) > 0 \quad \forall x, \ \forall \theta \in \Theta$$

Then, simulate one set of samples

$X_1, \ldots, X_n \sim q$.

Use importance sampling

Estimate of $E_{f_\theta}(g(x)) = \sum_{i=1}^{n} g(X_i)\, w_\theta(X_i)$

where $w_\theta(X_i) = \dfrac{h_\theta(X_i)/q(X_i)}{\sum_{i=1}^{n} \frac{h_\theta(X_i)}{q(X_i)}}$

~~where~~ and $\dfrac{h_\theta(x)}{c(\theta)} = f_\theta(x)$

Note: $c(\theta)$ is a function of θ so really a normalizing <u>function</u>.

Estimate $E_{f_{\theta_i}}(g(x))$ for any $\theta_i \in \Theta$ using single set of samples.

Very efficient (but not due to variance reduction.)

q ~~need~~ should not be optimal for a single f_{θ_i}, but rather should be adequate $\forall \theta \in \Theta$.

Difficult to find good q in general. (may not be able to get away w/ single q in practice for hard problems.)

Observation:

(1) Importance sampling is useful for estimating normalizing constants. (This is implicitly used in ratio imp. sampling.)

We know $h(x)/c = f(x)$

Given $h(x)$, how can we estimate c?

$\because \quad E_q \left\{ \frac{f(x)}{q(x)} \right\} = 1$

We have $E_q \left\{ \frac{h(x)}{q(x)} \right\} = c$ normalizing constant is written here as an expectation

Can use Monte Carlo to estimate expectation.
Useful for Monte Carlo max. likel.

(2) General purpose tool for integration.

Suppose you want to find $\int_a^b \psi(x)\,dx$

If you have q s.t. $q(x) > 0$ when $\psi(x) \neq 0$,

$$\int_a^b \psi(x)\,dx = \int_a^b \left[\frac{\psi(x)}{q(x)}\right] q(x)\,dx = E_q \left\{ \frac{\psi(x)}{q(x)} \right\}.$$

Rare event problems: importance sampling can be very useful.

E.g. want $P(Z > 4.5)$ where $Z \sim N(0,1)$.

Naive M.C.: $Z_1, \ldots, Z_n \overset{iid}{\sim} N(0,1)$

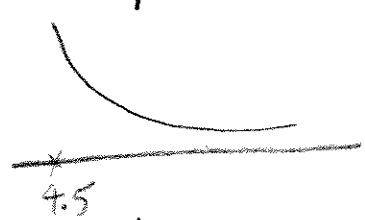
$$\hat{\mu}_n = \sum_{i=1}^{n} I(Z_i > 4.5)/n$$

Even for $n = 100{,}000$, $\hat{\mu}_n$ usually 0 ($\because P(Z>4.5)$ is small)

Instead use q = shifted Exp $(4.5, 1)$

$Y_1, \ldots, Y_n \overset{iid}{\sim}$ Shifted Exp$(4.5, 1)$

Shifted Exp: Expon(scale = 1) shifted right to 4.5



$$\text{pdf} = \frac{e^{-(x-4.5)}}{\int_{4.5}^{\infty} e^{-x}\,dx}$$

My Experiment

Importance sampling:
With $n = 10{,}000$

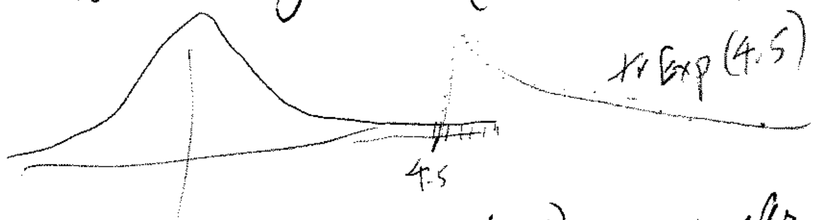
$$\hat{\mu}_n = \frac{\sum_{i=1}^{n} I(Y_i > 4.5)\frac{f(Y_i)}{q(Y_i)}}{n}$$

gives 3.35×10^{-6}

$n = 100{,}000$ gives 3.38×10^{-6}

True value (R's pnorm function): 3.3976×10^{-6}.

Case study: $P(\cancel{Z} > 4.5)$ $\quad Z \sim N(0,1)$



Simple imp. est. } $q = trExp(4.5)$ works well } support ≤ 4.5

$q = trExp(2)$ " "

$q = trE(5)$ does <u>not</u> work ! } support > 4.5

Ratio est. } $q = trE(4.5) \Rightarrow \hat{\mu} = 1$ does not work }

$q = trE(2) \Rightarrow \hat{\mu} = 0.001$

$q = trE(5) \Rightarrow \hat{\mu} = 1$ does not work }

Q.1. How can we get ratio est. to work ?

Q.2. What are conditions under which imp. est. work ?

(1) simple (2) ratio

<u>Ratio est.</u> works if $g = N(4.5, 1)$ (Need 1 million samples to get good estimate)

~~[illegible]~~ does not " " $g = \cancel{[illegible]}\, Exp(1)$ even 10 million samples

1 million samples $\Rightarrow \hat{\mu} = 6.9 \times 10^{-6}$

FOR MYSELF (NOT CLASS) MC 20 extra

Notes: Naive imp. sampling works as long as q is s.t. $g(x) f(x) \neq 0 \Rightarrow q(x) > 0$ a.s.

Less stringent than ratio imp. sampling condition: q is s.t. $f(x) > 0 \Rightarrow q(x) > 0$ a.s.

For e.g. for our toy problem $q = N(4.5, 1)$ works but $q = ShExp(4.5)$ does not work (when using ratio imp. sampling.)

Soln: Can use different importance fns. for numerator and denominator.

Tail probability problems are common: e.g. hypothesis testing. (p-values).

Estimating probabilities of rare events e.g. physics, astronomy

Importance sampling : M.C. standard errors

Estimate is useless w/o estimate of its variability

Simple imp. sampling : easy!

$$Var(\hat{\mu}_n) = Var\left\{\frac{1}{n}\sum_{i=1}^{n} g(x_i)\frac{f(x_i)}{q(x_i)}\right\}$$

$$= \frac{1}{n} Var_q\left\{g(x)\frac{f(x)}{q(x)}\right\} \quad (\because x_1,\dots,x_n \overset{iid}{\sim} q)$$

$\therefore$ we have sample $X_1,\dots,X_n \overset{iid}{\sim} q$,

estimate $Var_q\left\{g(x)\frac{f(x)}{q(x)}\right\}$ by sample variance

of $g(X_1)\frac{f(X_1)}{q(X_1)}, \dots, g(X_n)\frac{f(X_n)}{q(X_n)}$, say $\hat{\sigma}^2$

Est. M.C. s. error $= \frac{\hat{\sigma}}{\sqrt{n}}$.

Note: No guarantee that ~~above~~ M.C. s. error is finite

If $E_q\left\{g^2(x)\frac{f^2(x)}{q^2(x)}\right\} < \infty$ and $E_q\left\{\frac{f^2(x)}{q^2(x)}\right\} < \infty$

then s. error is finite.

Implication: q w/ lighter tails than f are not appropriate since variance will be infinite for many fns. g.

(Second condition $E_q\left\{\frac{f^2(x)}{q^2(x)}\right\} = \int \frac{f^2(x)}{q(x)}dx = \int \frac{f(x)}{q(x)} f(x)\,dx \le \int K f(x)\,dx$

$= K < \infty$ if rej. sampling condition is satisfied)

However, more important to consider ratio imp. estimator.

To rigorously quantify M.C. s.error, need C.L.T.

Imp. sampling CLT: If $E_q\left\{\frac{f^2(x)}{q^2(x)}\right\} < \infty$ and $E_q\left\{g^2(x)\frac{f^2(x)}{q^2(x)}\right\} < \infty$ then (1) the CLT holds for the ratio imp. sampling estimate and (2) the method of moments estimate of its asymptotic variance is consistent.

Multivariate CLT: If $X_1, \ldots, X_n \overset{iid}{\sim} q$, $w(X_i) = \frac{f(X_i)}{q(X_i)}$,

$$\sqrt{n}\begin{pmatrix} \frac{1}{n}\sum g(X_i)w(X_i) - \mu \\ \frac{1}{n}\sum_{i=1}^{n} w(X_i) - 1 \end{pmatrix} \xrightarrow{\mathcal{L}} N(\underset{\sim}{0}, \Sigma_{2\times 2}).$$

To derive variance of ratio estimator, appeal to delta-method where if $Q(u,v)$ is real-valued, then

$$\sqrt{n}\left(Q\left(\frac{1}{n}\sum_{i=1}^{n} g(X_i)w(X_i), \frac{1}{n}\sum_{i=1}^{n} w(X_i)\right) - Q(\mu, 1)\right)$$
$$\xrightarrow{\mathcal{L}} N(0, \underbrace{Q'(\mu,1)^T \Sigma\, Q'(\mu,1)}_{\sigma^2\text{, asymptotic variance}})$$

where Q' is a 2×1 matrix of partial derivatives.

For ratio estimation $Q(a,b) = a/b$,

$$Q'(a,b) = \left(\frac{1}{b}, -\frac{a}{b^2}\right)^T$$

$$Q'(\mu,1) = (1, -\mu)^T$$

$$\Sigma = \begin{bmatrix} Var(a) & Cov(a,b) \\ Cov(a,b) & Var(b) \end{bmatrix}$$

where $a = \frac{1}{n}\sum_{i=1}^{n} g(x_i)\, w(x_i)$, $b = \frac{1}{n}\sum_{i=1}^{n} w(x_i)$

Thus, $\sqrt{n}\,(\tilde{\mu}_n - \mu) \xrightarrow{\mathcal{L}} N(0, \sigma^2)$

where $\sigma^2 = Q'(\mu,1)^T\, \Sigma\, Q'(\mu,1)$.

If we obtain consistent estimator of σ^2, such as method of moments est., can appeal to Slutsky's Thm.

Simplifying, we get

$$\hat{\sigma}^2 = n \left(\frac{\sum_{i=1}^{n} g(x_i)\,\tilde{w}(x_i)}{\sum_{i=1}^{n} \tilde{w}(x_i)}\right)^2 \left[\frac{\sum_{i=1}^{n} g^2(x_i)\,\tilde{w}^2(x_i)}{\left(\sum_{i=1}^{n} g(x_i)\,\tilde{w}(x_i)\right)^2} + \frac{\sum_{i=1}^{n} \tilde{w}^2(x_i)}{\left(\sum_{i=1}^{n} \tilde{w}(x_i)\right)^2} - 2\,\frac{\sum_{i=1}^{n} g(x_i)\,\tilde{w}^2(x_i)}{\left(\sum_{i=1}^{n} g(x_i)\,\tilde{w}(x_i)\right)\sum_{i=1}^{n} \tilde{w}(x_i)}\right]$$

Comparison to rejection sampling

1. Is a generalization of rejection sampling, and is always at least as efficient (Y. Chen, 2006)
   when using χ^2 distances

2. Do not require
$$\sup_{x \in \Omega} \frac{f(x)}{g(x)} \le K \qquad \text{for some } K < \infty$$
   However, this is very desirable: if above holds and $E_g\{g^2(x)\} < \infty$ CLT holds and m.o.m. variance estimates are consistent.

3. Do not need to find K above.

4. May not want g to match f well; may be more important to match gf, for e.g.

5. Variance estimates are more unreliable.

6. Numerical stability issues: need to exponentiate $\log(h/g)$ to calculate weights. No need for this in rejection sampling (can do everything in log-scale).

7. Harder to estimate density. Use cdf at many pts.
$$F(x) = \int I(X \le x) f(x)\,dx = E_f\{I(X \le x)\}.$$

Imp. sampling estimates can be unstable

Suppose $q(x)$ is v. small for $x \in S$. where $g(x)h(x)$ is very large.

If $X^* \in S$ is obtained, it gets huge weight $\left(\frac{g(x)\,h(x)}{q(x)}\right)$

$\Rightarrow$ $\tilde{\mu}$ changes drastically. Improvement E. Ionides 2[illegible]; Truncated importance sampling

M.C. s.error: Even after many samples drawn, may not observe $X^* \in S$ ($\because$ $q(X^*)$ is small).

Hence, may never see X^* s.t. $g(X^*)\dfrac{h(X^*)}{q(X^*)}$ is huge,

so s.error will be heavily underestimated.

Worse/unstable M.C. estimates may imply worse/unreliable s.error estimates. (Worst possible situation: bad estimates, but we think we have good estimates!)

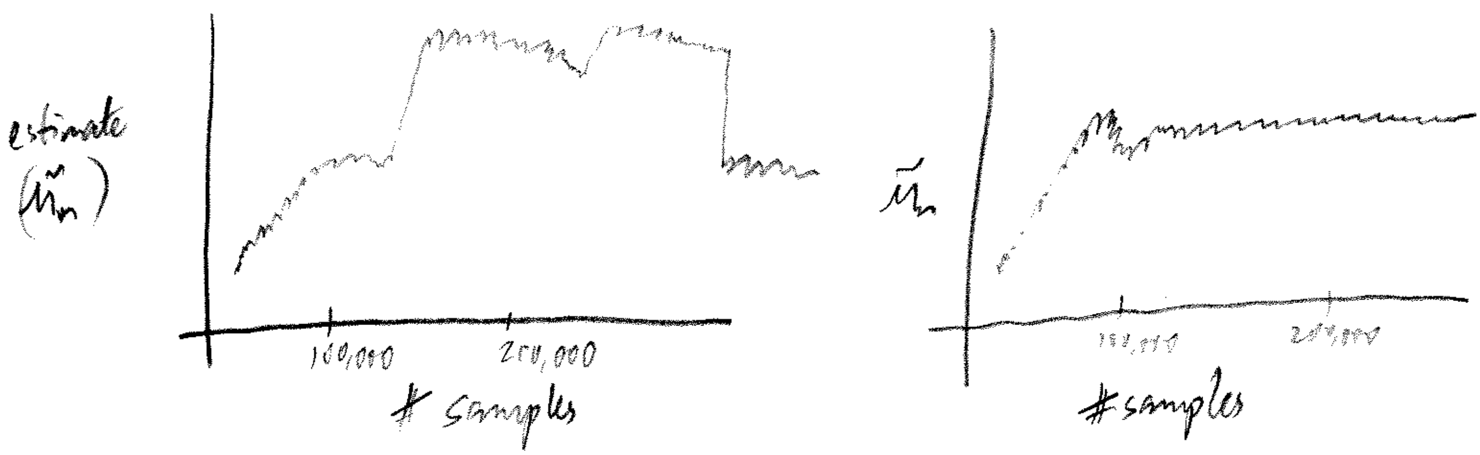
Importance sampling:

How many samples are necessary (what is N?) for good estimates?

Best approach: M.C. s.errors: when errors are small enough, stop sampling.

Problem: M.C. s.errors are not always reliable so it is best to use M.C. s.errors + some simple heuristics ('reality checks').

<u>Trace plots</u>:



(a) (b)

(a) Obviously poorly converging estimate

(b) Apparently more stable " (but cannot be 100% sure).

Plots: can tell you if estimate is obviously unstable but cannot tell you when estimator (and M.C.s.e.s) are reliable. (jump could happen after 1 million iteration for plot (b).)

MC 28

Techniques for finding importance functions

Some approaches (not mutually exclusive):

1. Exponential tilting see Ross
2. Defensive importance fns. (T. Hesterberg '95 Technometrics)

   Use a mixture: $q(x) = \rho\, q_1(x) + (1-\rho)\, q_2(x)$ $\rho \in (0,1)$

   w/ $q_1(x)$ matching $g(x)f(x)$ well but $q_2(x)$ is heavy-tailed. Set ρ close to 1; this ensures variances are finite.

3. Laplace approx: (Tierney '89, Tierney & Kadane '86)

   Assume $\log f$ admits Taylor expansion about its mode.

   Let $l(x) + a = \log f(x)$ for some constant a.

   Let $\hat{x}$ satisfy $l'(\hat{x}) = 0$ l' = vector of 1st derivatives of l.

   and let $l''(x)$ be matrix of 2nd derivatives of l.

Then, $\mu = \int g(x) f(x) dx = \int g(x) e^{l(x)+a} dx = e^a \int g(x) e^{l(x)} dx$

$$\approx e^a \int g(x) \exp\left(l(\hat{x}) + (x-\hat{x})^T l''(\hat{x}) (x-\hat{x})/2 \right) dx \quad (*)$$

($\because$ $l'(\hat{x}) = 0$)

Setting $g(x) = 1$ we obtain

$$1 = \mu_1 \simeq e^a \underbrace{\sqrt{2\pi}\, |l''(\hat{x})|^{-1/2}}_{\text{by integrating Kernel of normal density (this is normalizing constant)}} \exp(l(\hat{x})) \quad \text{--- (**)}$$

Hence we eliminate e^a by: dividing (*) by (**):

$$\mu \simeq \frac{1}{\sqrt{2\pi}} |l''(\hat{x})|^{+1/2} \int g(x) \exp\{(x-\hat{x})^T l''(\hat{x})(x-\hat{x})/2\}\, dx$$

Suggesting that f is approximately Normal w/ mean $\hat{x}$ and variance $-(l''(\hat{x}))^{-1}$ = negative inverse of of curvature at mode.

Rather than use this approximation, directly, may be good to use heavier-tailed (e.g. multivariate-t) w/ mean and variance estimates above

4. Lots of other approaches (c.f. J. Liu's book).

Important note: $\because$ we have to match entire distr. at once, imp sampling does not scale well.

Recent developments: sequential importance sampling

$$\underset{\text{easy}}{\pi_1} \xrightarrow{\text{reweight}} \pi_2 \xrightarrow{\text{reweight}} \cdots \rightarrow \underset{\text{target}}{\pi_n}$$

Mr 30