Comparison Theorems for Practical Slice Sampling

Slice Sampling is a popular algorithm for approximate sampling from intractable probability distributions, used in {JAGS, Matlab, ...}.

Its popularity stems from its wide applicability, intuitive geometric formulation and robustness, in both practice and theory.

An outstanding theoretical challenge has been that while the ideal slice sampler ~~has excellent~~ admits a robust and elegant convergence theory, practical implementations typically involve additional approximations, which prevents the existing theory from holding as is.

In recent work, we develop a theoretical framework for the analysis of such "hybrid" slice samplers, facilitating novel convergence results for slice sampling as implemented in practice.

We provide a number of concrete examples which illustrate the flexibility and practicality of our approach, including i) stepping-out and shrinkage procedures, ii) hit-and-run "on the slice",

No prior knowledge of the slice sampling algorithms will be assumed, and relevant notions of Markov chain convergence ~~with be~~ will be ~~developed~~ introduced as appropriate in the talk.

Optimality of MLE

in what sense? $\rightarrow$ MLE is typically consistent asymptotically unbiased, good variance.

$\downarrow$

want a finite-n comparison.

simplify. think about unbiased estimators; what is possible?
(good estimators might be close to unbiased)
(c.f. Gauss-Markov).
mathematically, convenient family.

Start with
$$\begin{cases} \int P_\theta(x)\,dx = 1 \\ \int P_\theta(x)\,\hat{\theta}(x)\,dx = \theta \end{cases}$$
$\Rightarrow$ θ-independent relation.

~~general~~ $\cancel{\int P_\theta(x)\,F(x) = G(\theta).}$

define $F(\theta) = \int P_\theta(x)\,f(x)\,dx \in \mathbb{R}$

$$\frac{\partial F}{\partial \theta_k} = \frac{\partial}{\partial \theta_k}\int P_\theta(x)\,f(x)\,dx$$
$$= \int P_\theta(x)\,\frac{\partial}{\partial \theta_k}\log P_\theta(x)\cdot f(x)\,dx$$

$$\nabla_\theta F(\underline{\theta}) = \mathrm{Cov}_\theta\left(\nabla_\theta \log P_\theta(x),\, f(x)\right) \in \mathbb{R}^p.$$

$f(x) = 1 \Rightarrow \cancel{0 = \mathrm{Cov}_\theta\,\nabla}$

① MCMC

$\pi \to [x_1 \to x_2 \to \cdots$

$\mathrm{Law}(X_t) \to \pi$

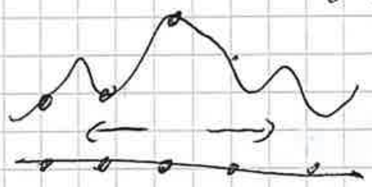
$\frac{1}{T}\sum_{t=1}^{T} f(x_t) \to \int \pi(dx)\, f(x).$

oracle $\log \pi$ (up to a constant), [More]

② Slice Sampling

$\pi(dx) = \varpi(x)\, \lambda^d(dx)$

$\Pi(d(x,u)) = \cancel{\varpi(x)}\; \mathbb{I}[0 \le u \le \varpi(x)]\, \lambda^{d+1}(d(x,u))$



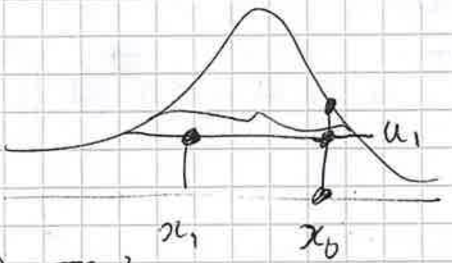
"iterated conditional simulation".

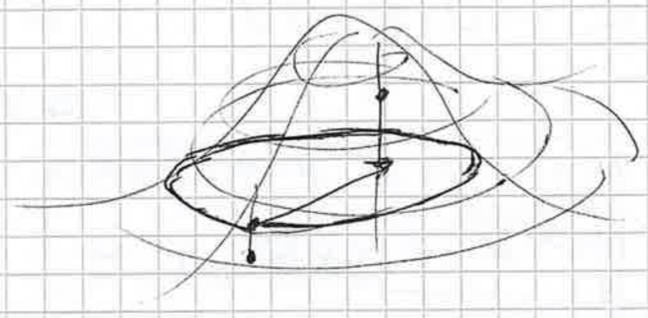
① at x, $u \sim \Pi(du \mid x) = \mathrm{Unif}([0, \varpi(x)])$

(sample a height)

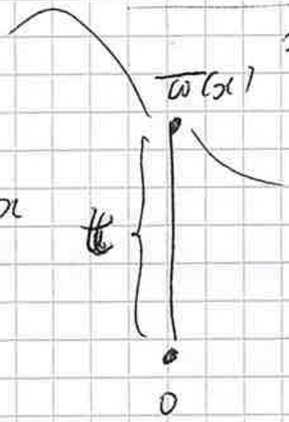
② at u, $x \sim \Pi(dx \mid u) = \mathrm{Unif}(\{x : \varpi(x) \ge u\})$.

(sample a point ~~[illegible]~~ at the height)

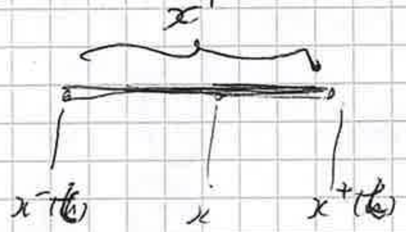




at x



at u



$x \longrightarrow x'$

$x' \sim U(x, dx').$ [Markov kernel]

$$\binom{a}{b}^T - \binom{a}{b} = a^T I(\theta) a + 2a^T b + b^T \mathrm{Cov}(\hat{\theta}(x)) b$$

$$\frac{d}{da}: \; I(\theta) a + b = 0, \quad a = -I(\theta)^{-1} b.$$

$$\Rightarrow b^T \left(\mathrm{Cov}\, \hat{\theta}(x) - I(\theta) \right) b \geq 0.$$

$$\begin{pmatrix} -I(\theta)^{-1} b \\ b \end{pmatrix}$$

Caveats let $\mathbb{E}\,\hat{\theta}(x) = (\cancel{1+\varepsilon})\theta$

$$\cancel{\mathbb{E}\,\|(1-\varepsilon)\hat{\theta}(x) - \theta\|^2 = (1-\varepsilon)^2\, \mathbb{E}\,\|\hat{\theta}(x)}$$

$$\cancel{\mathrm{Cov}\;(1-\varepsilon)\theta}$$

$$\mathrm{Cov}\,(A\hat{\theta}(x)) = A\, \mathrm{Cov}(\hat{\theta}(x)) A^T$$

$$\mathbb{E}\,(A\hat{\theta}(x) - \theta)^{\otimes 2} = \mathbb{E}(A(\hat{\theta}(x) - \theta) + (A-1)\theta)^{\otimes 2}$$

$$= ACA^T + \|(A-1)\theta\|^2$$

$$\boxed{\cancel{\frac{\partial}{\partial A} \quad A \to A + \delta A \quad (\delta A) C A^T +}}$$

Bias-Variance Tradeoff

One approach
1. come up with a sensible estimator
2. check for consistency
3. check that bias is sub-dominant
4. apply shrinkage/regularisation to control variance

(other stuff)

Thm$^{(M)}$ SS leaves π invariant $\Rightarrow$ valid for MCMC
Thm$^{(RR)}$ weak conditions $\Rightarrow$ SS is geometrically ergodic $\Rightarrow$ more thms
observation I know no example where SS is not geometrically ergodic

Thm (NES). π spherically symmetric, log-concave
$\Rightarrow$ convergence is like $e^{-t/d}$, t iterations
also: π itself doesn't matter, just mass of level sets
Thm (Sch) π = student-t $\Rightarrow$ convergence like e^{-t/d^2}
(heavy tails but still geometric)
$\exists$ other examples.

$\Rightarrow$ When implementable, SS is robust and handles fairly dimensionality well, for a 0^{th}-order method.

Issue "sample $x \sim \mathrm{Unif}(\{x : \pi(x) > t\})$.
$= \mathrm{Unif}(G(t))$

- If $G(t) \in$ ○, □, ---, sure.
- If $G(t)$ is more arbitrary .... work for it.

Let $\nu_t(dx) = \mathrm{Unif}(dx; G(t))$

SS: $t \sim \mathrm{Unif}([0, \pi(x)])$
$x' \sim \nu_t$

HSS: replace $x' \sim \nu_t$
with $x' \sim \mathrm{MCMC}(x \to x'; \text{target} = \nu_t)$

Usually easier to implement, slower Markov chain.
(examples!)

Question: how much slower?
how do we trade off ease / efficiency in practice?
what price are we currently paying?
"comparison theory"

Fixed p, $n \to \infty$: bias not always an issue.
$p \asymp n$, $(p,n) \to (\infty,\infty)$: more subtle.

$$MSE(\hat\theta(X)) = \|\mathbb{E}\,\hat\theta(X) - \theta\|_2^2 + Tr(Cov(\hat\theta(X)))$$

statistical "efficiency" : $Var \asymp Var_{CRLB}$
(asymptotic)

(often only asymptotically unbiased)

Proof Elements

(in practice, want "bias is not an issue")

① $\int P_\theta(x)\, \nabla_\theta \log P_\theta(x)\, dx = 0.$

② $\int P_\theta(x)\, \nabla_\theta^2 - \log P_\theta(x)\, dx$

$$= \int P_\theta(x) \left(\nabla_\theta \log P_\theta(x)\right)\left(\nabla_\theta \log P_\theta(x)\right)^T dx$$

③ $\int P_{\theta_*}(x) \log \frac{P_{\theta_*}(x)}{P_\theta(x)}\, dx \geq 0.$

④ $\ell(\theta; X) = \log P_\theta(X)$
$\ell(\theta) = \mathbb{E}_{\theta_*}[\log P_\theta(X)] \leq \ell(\theta_*).$

Proof let $\hat\theta(x)$ be unbiased for θ.

$$\Rightarrow \forall \theta \in \Theta, \quad \int P_\theta(x)\, \hat\theta(x)\, dx = \theta.$$

$$\frac{\partial}{\partial\theta_i} \Rightarrow \int P_\theta(x) \left(\frac{\partial}{\partial\theta_i} \log P_\theta(x)\right) \hat\theta_j(x)\, dx = \delta_{ij}$$

$$Cov(\nabla_\theta \log P_\theta(X), \hat\theta(X)) = I.$$

$$Cov\begin{pmatrix} \nabla_\theta \log P_\theta(X) \\ \hat\theta(X) \end{pmatrix} = \begin{pmatrix} \mathcal{I}(\theta) & I_p \\ I_p & Cov(\hat\theta(X)) \end{pmatrix} \succeq 0.$$

$\cancel{\begin{pmatrix} v_1 \\ v_2 \end{pmatrix}^T (\;)\begin{pmatrix} v \\ v \end{pmatrix} = \left\langle \begin{pmatrix} 0 \\ v \end{pmatrix}, \begin{pmatrix} v \\ Cov(\hat\theta(X))\, v \end{pmatrix} \right\rangle =}$

Convergence of Markov chains

$$\mu_n = \mu P^n.$$

$$d(\mu, \pi) = \int \pi(dx) \left(\frac{d\mu}{d\pi}(x) - 1 \right)^2 = \left\| \frac{d\mu}{d\pi} - 1 \right\|^2_{L^2(\pi)}.$$

$$\|f\|^2_{L^2(\pi)} = \int \pi(dx)\, f(x)^2$$

$$Pf(x) = \int P(x, dy)\, f(y).$$

$$P^n f(x) \longrightarrow \pi(f) \quad \text{(ergodicity)}$$

$$\| P^n f - \pi(f) \|^2_{L^2(\pi)} \longrightarrow 0 \qquad \text{"}L^2\text{ convergence"}$$

(focus on $\pi(f) = 0$)

≈ <u>Best case</u> : $\| P f \|_2^2 \leq (1 - \lambda) \|f\|_2^2 \quad \forall f \in L_0^2(\pi)$

⟹ exponential rate.

actually, equivalent to

$$\overbrace{\mathcal{E}(P^*P, f)}^{} \geq \lambda \|f\|_2^2$$

remaining energy

$$\underbrace{\frac{1}{2} \int \pi(dx)\, (P^*P)(x, dy)\, (f(x) - f(y))^2 \qquad (I - P^2) = (I + P)(I - P) \geq (I - P)}_{\text{energy dissipation}}$$

<u>Fact</u> P "positive", reversible $\mathcal{E}(P, f) \geq \gamma \|f\|^2$

$$\Rightarrow \quad \|P^n f\|^2 \leq (1 - \gamma)^{n} \|f\|$$

"good energy dissipation ⟹ good convergence"

CRAMÉR-RAO LOWER BOUND

Gauss-Markov: Estimator $\wedge$ Linear $\wedge$ Unbiased
$\implies$ OLS is minimum variance

Same for MLE?

$$\ell_n(\theta; \cancel{y_{1:n}}) = \log P_\theta(x_{1:n})$$

$$\hat{\theta}_n = \operatorname{argmax}_\theta \ell(\theta; \cancel{y_{1:n}})$$

$$X_1, \ldots, X_n \overset{iid}{\sim} P_{\theta_*}$$

assumptions $\rightarrow$

$$\hat{\theta}_n \overset{d}{\approx} N(\theta_*, (n\mathcal{I}_n)^{-1})$$

$$\mathcal{I}_* = \mathbb{E}_{\theta_*}[\nabla_\theta^2 - \log P_\theta(X)]$$

$$= \mathrm{Cov}_{\theta_*}[\nabla_\theta^2 \log P_\theta(X)]_{\theta=\theta_*}$$

$$\approx \mathrm{Cov}_{\hat{\theta}_n}[\nabla_\theta^2 \log P_\theta(X)|_{\theta=\hat{\theta}_n}]$$

don't do it with n?

$$\implies \mathrm{Var}(\hat{\theta}_{MLE}) = 1/n \cdot \mathcal{I}_*(1+o(1))$$

can we find $\hat{\theta}$ st $\mathrm{Var}(\hat{\theta}_{MLE}) \ll 1/n I_*$?

Cramér-Rao: Let $\hat{\theta}$ be unbiased for θ, ie

$$\forall \theta \in \Theta, \quad \int P_\theta(x)\hat{\theta}(x)\,dx = \theta.$$

Then $\underbrace{\mathrm{Cov}_\theta(\hat{\theta}(X))}$ $\geq \mathcal{I}(\theta)^{-1}$.

so, $\forall v$, $\mathrm{Cov}_\theta(\langle\hat{\theta}(x), v\rangle) \geq v^T I(\theta)^{-1} v \in \mathbb{R}$

Intuition: If $\mathcal{E}(P, f) \geq \gamma \|f\|^2$,
then $P \approx \gamma$ independent samples from π.
~~[illegible]~~ decorrelation time $\approx \gamma^{-1}$ at equilibrium

Comparison: If $\mathcal{E}(P_1, f) \geq K\, \mathcal{E}(P_2, f)$
then $P_1 \approx K$ steps of P_2

We will follow in this direction.

P_1 = implementable algorithm
P_2 = ideal slice sampling

(actually: more generally, $\|f\|^2 \leq \mathcal{E}(P, f) + \beta(\omega)\|f\|_\infty^2$

even if ~~$\beta > 0$ $\forall (\alpha,\beta)$ [illegible]~~
$\beta(\omega) > 0$, (non-exp) convergence, eg

$$\|P^n f\|_2^2 \leq \gamma(n) \cdot \|f\|_\infty^2$$ Examples

so, we might instead prove

$$\mathcal{E}(P_2, f) \leq s\, \mathcal{E}(P_1, f) + \beta(s) \cdot \|f\|_\infty^2$$

$\approx P_1$ not much worse than P_2 (depending on β)

Let H_t be ν_t-rev, pos. $\quad m(t) = \nu(G(t))$

$$\forall t, \forall s \quad \|f\|_{0,\nu_t}^2 \leq s \cdot \mathcal{E}(H_t, f) + \beta(t, s)\|f\|_{osc}^2$$

Define $\quad \beta(s) = \int_0^{\|\varpi\|_\infty} \beta(s,t)\, m(t)\, dt$

$$\Longrightarrow \boxed{\cancel{[illegible]}} \leq s \cdot \mathcal{E}(H, f) + \beta(s)\|f\|_{osc}^2$$

$\mathcal{E}(U, f)$

$$\Longrightarrow \forall v \in \mathbb{R}^p, \quad v^T(\mathcal{I}(\theta) - C(\theta)^{-1})v \geq 0$$

$$\mathcal{I}(\theta) \succeq C(\theta)^{-1}$$

$$\underbrace{C(\theta)}_{\text{variance of unbiased estimator}} \succeq \underbrace{\mathcal{I}(\theta)^{-1}}_{\text{"Cramér-Rao lower bound"}}$$

Current reaction: is this lower bound achievable?
<u>roughly</u>: asymptotically (large-n), for reasonable models, yes, and the MLE does so.

The MLE is <u>almost unbiased</u> (as $n \to \infty$),
<u>almost minimal-variance</u> (w.r.t. CRLB)
$\Longrightarrow$ "optimal".

<u>Caveats</u>:
- At {finite n / large p / ...}, more to life than unbiasedness, variance.
- All relies on correct model specification. If $P_* \notin \{P_\theta : \theta \in \Theta\}$, no luck
  $\to$ doesn't say much about robustness...
- Sharp when $\eta(x)$ is parallel to $\nabla_\theta \log p_\theta(x)$.
  $\to$ exponential families (can develop)
- needs "regular" model (no unif. ...)

$\Rightarrow$ ① study $u^n \to \pi$

② study $H_t^n \to \nu_t$

$\Longrightarrow$ study $H \to \pi$.

Is the ideal slice sampler good?

Is the approximation good?

Metropolis (π, ν, Q), $\nu \otimes Q$ symmetric

$$\begin{cases} \text{RWM}: (\pi, \text{Leb}, \text{GRW}) \\ \text{IMH} \quad (\pi, q, q) \\ \text{pCN} \quad (\pi, \gamma_{m,C}, \text{OU}) \end{cases}$$

Metropolis $\subseteq$ HybridSlice

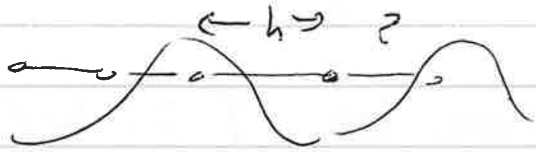
Metropolis $\preceq$ HybridSlice

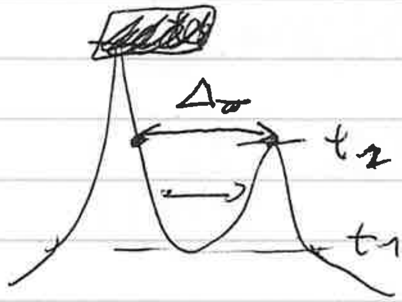
RWM $\preceq$ Slice (π, Leb)

IMH $\preceq$ Slice (π, q)

pCN $\preceq$ Slice $(\pi, \gamma_{m,C})$

Stepping-Out + Shrinkage





$$\mathcal{E}(H_t, f) \geq \lambda(t) \|f\|_{\nu_t}^2$$

$$\lambda(t) = \frac{h - \delta(t)}{h} \times \frac{m(t)}{m(t) + \delta(t)}$$

$$\geq \frac{h - \Delta}{h} \times \frac{m(t)}{m(t) + \Delta}$$

$$\Rightarrow \mathcal{E}(H, f) \geq \frac{h - \Delta}{h} \times \frac{m_-}{m_- + \Delta} \mathcal{E}(U, f)$$

$$h = 2\Delta \Rightarrow \geq \frac{1}{2} \frac{m_-}{m_- + \Delta}$$

So, $\mathbb{E}_\theta[\nabla_\theta \log P_\theta(X)] = 0$

$$I(\theta) = \mathbb{E}_\theta[\nabla_\theta \log P_\theta(X) \cdot \nabla_\theta \log P_\theta(X)^T]$$

$$= \mathrm{cov}_\theta[\nabla_\theta \log P_\theta(X)]$$

$$= \mathbb{E}_\theta[\nabla_\theta^2(-\log P_\theta(X))]$$

(do examples)

One more:

③ Since $\int P_\theta(x)\eta(x)\,dx = \theta$ (by assumption)

$\frac{d}{d\theta} \Rightarrow \int P_\theta(x)\eta(x)\nabla_\theta \log P_\theta(x)^T dx = I_p$

$$\mathbb{E}_\theta[\underbrace{\eta(X)}_{\text{mean } \theta}\ \underbrace{\nabla_\theta \log P_\theta(X)^T}_{\text{mean } 0}] = I_p$$

$$\mathrm{cov}_\theta[\eta(X), \nabla_\theta \log P_\theta(X)] = I_p.$$

Now, let $Y = \begin{pmatrix} \eta(X) \\ \nabla_\theta \log P_\theta(X) \end{pmatrix}$.

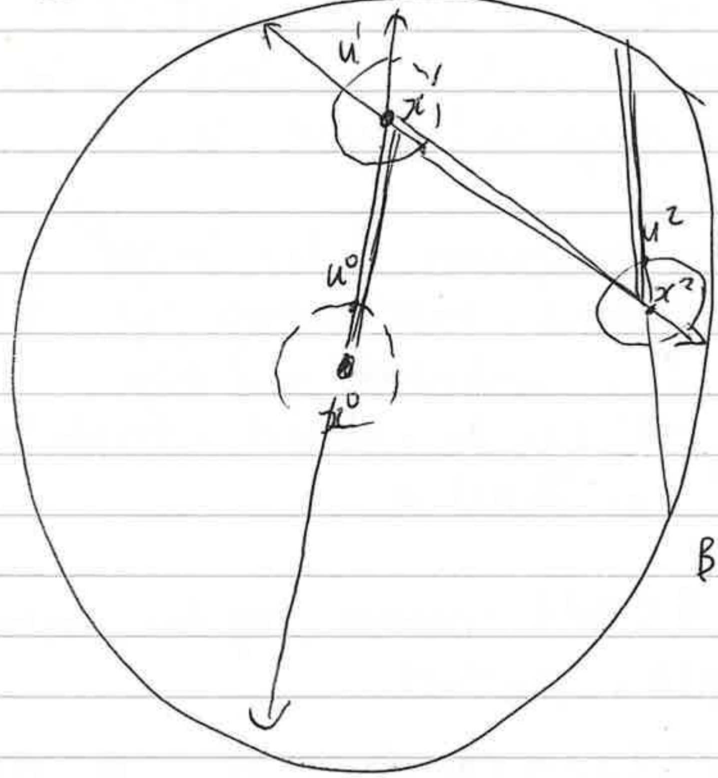
Compute $\mathrm{Cov}_\theta(Y) = \begin{pmatrix} C(\theta) & I_p \\ I_p & \mathcal{I}(\theta) \end{pmatrix} \succeq 0.$

$$\begin{pmatrix} u \\ v \end{pmatrix}^T \mathrm{cov}_\theta(Y) \begin{pmatrix} u \\ v \end{pmatrix} = u^T C(\theta) u + 2u^T v + v^T \mathcal{I}(\theta) v \geq 0$$

min wrt $u \Rightarrow u^* = -C(\theta)^{-1} v$

$$\Rightarrow \begin{pmatrix} -C(\theta)^{-1} v \\ v \end{pmatrix}^T \mathrm{cov}_\theta(Y) \begin{pmatrix} -C(\theta)^{-1} v \\ v \end{pmatrix} = v^T \{\mathcal{I}(\theta) - C(\theta)^{-1}\} v$$

Hit-and-Run $\mathrm{Unif}(K)$, also 0^{th}-order.



$B(x,R) \supseteq K \supseteq B(x,r)$

LV: Hit-and-Run has $\gamma \gtrsim 2^{-33} \cdot \frac{1}{d^2} \cdot \left(\frac{r_K}{R_K}\right)^2$

dimension

conditioning

$\Longrightarrow$ Slice Sampling w/ quasi-concave π.

$\underline{\Pi 38}$ $m \le V'' \le L \Rightarrow \mathcal{E}(U, f) \le 2^{33} \cdot d^2 \cdot \left(\frac{L}{m}\right) \mathcal{E}(H, f)$.

$\underline{\Pi 39}$ $\pi_{d,m} \Rightarrow \|f\|^2_{d,m} \le 2^{33} \cdot d^2 \cdot \frac{(d+1)(d+m-1)}{m-1} \mathcal{E}(H, f)$

$\underline{\cancel{X40}}$: $U(x) \in \|x\|^{p_1,p_2}$ at 0, $\|x\|^{q_1,q_2}$ at ∞

$\Rightarrow \kappa_G(t) \lesssim \left(\log\left(\frac{1}{t}\right)\right)^{\frac{1}{q_1} - \frac{1}{q_2}}$ 0^+

$\left(\log\left(\frac{1}{t}\right)\right)^{\frac{1}{p_1} - \frac{1}{p_2}}$ 1^-

$\rightsquigarrow$ $p_1 = p_2$, $\beta(s) \lesssim e^{-\Omega(s^{q_1 q_2/(q_2 - q_1)})}$

$p_1 \ne p_2$ $\beta(s) \lesssim s^{-\left(1+\frac{d}{q_1}\right)\frac{p_1 p_2}{p_1 - p_2}}$

Parametric Estimation and the CRLB

OLS : model $\rightarrow$ linear, unbiased estimators
Gauss-Markov $\rightarrow$ OLS is minimum variance
not whole story, but a good start

general parametric estimation : too broad!
· finite-dimensional, fixed sample size $\{P_\theta(x)\}$
· "linear" = ? , unbiased : still lossy
Q : how "good" can an unbiased estimator be?
A : CRLB : $Var(\hat{\theta}(X)) \geq \cdots$

setting $\{P_\theta(x) : \theta \in \Theta\}$ smooth, regular. may have $x = (x_{1:n})$
let $\hat{\theta}(X) = \eta(X)$, so that
$\forall \theta \in \Theta$, $\int P_\theta(x)\, \eta(x)\, dx = \theta$ as vectors

Measure uncertainty via

$$Cov_\theta(\hat{\theta}(X)) = \int P_\theta(x) (\eta(x) - \theta)(\eta(x) - \theta)^T = C(\theta).$$

What can we say about $C(\theta)$?

Some basic results

$\left(\frac{d}{d\theta} f(\theta) = f(\theta) \frac{d}{d\theta} \log f(\theta)\right)$

① Since $\int P_\theta(x)\, dx = 1$

$\frac{d}{d\theta} \Rightarrow \int P_\theta(x)\, \nabla_\theta \log P_\theta(x)\, dx = 0_p$

$\mathbb{E}_\theta[\nabla_\theta \log P_\theta(x)] = 0_p$

② Since $\int P_\theta(x)\, dx = 1$

$\left(\frac{d}{d\theta}\right)^2 \Rightarrow \int P_\theta(x) \{\nabla_\theta \log P_\theta(x) \nabla_\theta \log P_\theta(x)^T + \nabla_\theta^2 \log P_\theta(x)\} dx$
$= 0_{p \times p}$

$$\mathcal{I}(\theta) := \int P_\theta(x) (\nabla_\theta \log P_\theta(x))(\nabla_\theta \log P_\theta(x))^T dx$$
$$= \int P_\theta(x) (\nabla_\theta^2 (-\log P_\theta(x))) dx$$

Takeaways

- SS : good in theory and practice
- HSS : used in practice, gap in theory
- Comparison : how well does H approximate U ?
- Applications :

$$H \succeq U$$

If $H_t \succeq c$, then $H \succeq cU$

If H_t good, then H almost as good as U.

- We focus on H_t = Hit-and-Run, and similar. Easy to combine bounds.
- General comparison framework valid beyond SS

Maximum Likelihood Inference

Linear Model: interpretable, fast / direct / transparent, flexible, exact inference / pivots / conjugacy.

But, not always appropriate for problems / data / ...

Extension 1 Linear Mixed Models.

m subjects, n input-output pairs per subject

for $j = 1, \dots, m$, $\quad y_{ij} = \beta_0 + \beta_1 x_{ij} + \epsilon_{ij}$

or: subject-wise intercept $\Rightarrow y_{ij} = \beta_0 + \beta_1 x_{ij} + \sum I[j=k]\gamma_k + \epsilon_{ij}$

can't estimate γ_j well if n is small (not identified)

shrinkage / sharing of information: $(\gamma_1, \dots, \gamma_m) \sim N(0, \Sigma(\phi))$

($\approx$ identifiability, regularisation) $\quad$ e.g. ϕI.

Extension 2 Non-Gaussian Input-output Regression Models

count data: $y_i = \text{Poisson}(\lambda_i)$

$\log \lambda_i = \beta_0 + \beta_1 x_i \quad (\Rightarrow$ latent linear$)$

estimation by LS inappropriate $\rightarrow$ log nonlinear, $\text{Var}(y|x)$ nonconstant

- instead of forcing problems into LS format, identify salient features of LS problem which are structurally important, beyond "full tractability"

- what principles do we have? how do we come up with OLS?