Design of informed Metropolis-Hastings proposal distributions

Giacomo Zanella

Bocconi University, Milan, Italy

LMS-EPSRC Durham Symposium
26 July to 5 August 2017

Informed Proposals

Aim: sampling from a probability measure π defined on Ω

Metropolis-Hastings (MH) kernel

1. Sample $y \sim Q(x, \cdot)$
2. Accept y with probability $1 \wedge \alpha(x, y)$ where $\alpha(x, y) = \frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)}$

Informed Proposals

Aim: sampling from a probability measure π defined on Ω

Metropolis-Hastings (MH) kernel

1. Sample $y \sim Q(x, \cdot)$
2. Accept y with probability $1 \wedge \alpha(x, y)$ where $\alpha(x, y) = \frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)}$

Uninformed proposals	vs	Informed proposals
"blind" proposal : $Q(x, y) = Q(y, x)$		Q incorporates info about the target π
$\Downarrow$		$\Downarrow$
small moves and slow mixing		longer moves and better mixing

Question: *How* should we design an *informed* proposal Q?

Ideal choice would be $Q(x, y) = \pi(y)$ but that's typically unfeasible...

Example: gradient-based MCMC

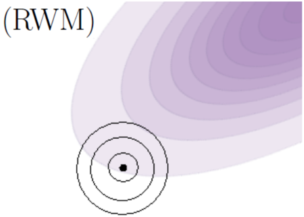
Framework: $\Omega = \mathbb{R}^n$, target $\pi(x)dx$

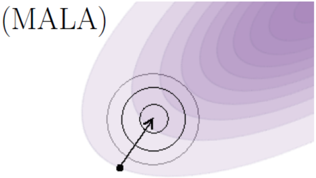
Typical uninformed proposal

$\rightarrow$ (RWM) $\quad Q_\sigma(x,\cdot) = N(x, \sigma^2 \mathbb{I}_n)$

How to design informed Q? Discretize π-rev. diffusion $dX_t = \frac{\nabla \log \pi(X_t)}{2} dt + dW_t$

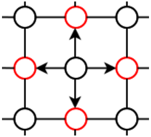
$\rightarrow$ (MALA) $\quad Q_\sigma(x,\cdot) = N(x + \sigma^2 \frac{\nabla \log \pi(x)}{2}, \sigma^2 \mathbb{I}_n)$





NB: by construction the bias towards high-probability regions is calibrated so that Q is approximately π-reversible.

Informed proposals in discrete spaces?



Ω finite state space

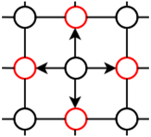
$\pi(x)$ target measure

$N(x)$ neighbourhood of x (e.g. $N_\sigma(x) = \{y \in \Omega : d(x, y) \leq \sigma\}$)

$K_\sigma(x, \cdot) = \text{Unif}(N_\sigma(x))$ natural uninformed proposal

Informed proposals in discrete spaces?

Ω finite state space
$\pi(x)$ target measure
$N(x)$ neighbourhood of x (e.g. $N_\sigma(x) = \{y \in \Omega : d(x,y) \leq \sigma\}$)
$K_\sigma(x,\cdot) = \mathrm{Unif}(N_\sigma(x))$ natural uninformed proposal

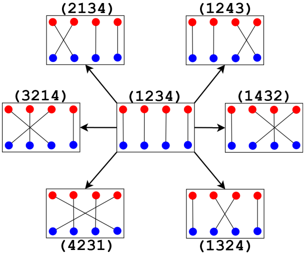


Example: sampling matchings

$$\Omega = \{\text{perfect matchings of } n+n \text{ bipartite graph}\}$$

$$\pi(x) \propto \prod_{e \in x} w_e$$

$$N(x) = \{y\text{'s obtained by swapping two edges of } x\}$$



Informed proposal $Q(x,\cdot) \rightsquigarrow$ non-uniform probability distribution on $N(x)$.
How should we design such a distribution?

Is the localized version of π, i.e. $Q_\pi(x,y) \propto \pi(y)\mathbb{1}_{N(x)}(y)$, a good choice?

Framework

Target distribution

$\Pi(dx) = \pi(x)dx$ for some base measure dx

Natural uninformed kernel

A Markov transition kernel $K_\sigma(x, dy)$ satisfying:

1. K_σ is dx-reversible
2. $K_\sigma(x, \cdot) \Rightarrow \delta_x(\cdot)$ as $\sigma \downarrow 0$ and $K_\sigma(x, dy) \Rightarrow dy$ as $\sigma \uparrow \infty$

Examples: $K_\sigma(x, \cdot) = N(x, \sigma^2 \mathbb{I}_n)$ or $K_\sigma(x, \cdot) = \text{Unif}(N_\sigma(x))$

Aim

Incorporate information from π into K_σ to obtain a good proposal Q to target π.
Equivalently: bias K_σ towards high-prob. regions of π in an appropriate way.
NB: would like to be fairly general in terms of π and K_σ.

Heuristics

Naive informed proposal

$$Q_\pi(x,dy)=\frac{\pi(y)K_\sigma(x,dy)}{Z_\sigma(x)} \qquad e.g.\ Q_\pi(x,y)=\frac{\pi(y)\mathbb{1}_{N_\sigma(x)}(y)}{\pi(N_\sigma(x))} \text{ or } \frac{\pi(y)e^{-\frac{|x-y|^2}{2\sigma^2}}}{(K_\sigma*\pi)(x)}$$

Q_π looks reasonable for big σ because $Q_\pi(x,dy)\Rightarrow\Pi(dy)$ as $\sigma\uparrow\infty$. What happens for small σ?

K_σ dx-reversible implies Q_π reversible w.r.t. $\pi(x)Z_\sigma(x)$

$$\text{But } Z_\sigma=K_\sigma*\pi \text{ and thus } \pi(x)Z_\sigma(x)\Rightarrow\begin{cases}\pi(x) & \text{if } \sigma\uparrow\infty \text{ (Global move)}\\ \pi(x)^2 & \text{if } \sigma\downarrow 0 \text{ (Local move)}\end{cases}$$

$\rightsquigarrow$ Q_π is *not* appropriate to design *local* moves targeting π

Heuristics

Simple fix: introduce a balancing function g, $Q_{g(\pi)}(x, dy) \propto g(\pi(y))K_\sigma(x, dy)$

$$Q_{\sqrt{\pi}}(x, dy) = \frac{\sqrt{\pi(y)}K_\sigma(x, dy)}{(\sqrt{\pi} * K_\sigma)(x)} \text{ reversible w.r.t. } \sqrt{\pi}(x)(\sqrt{\pi} * k_\sigma)(x) \quad \xrightarrow{\sigma \downarrow 0} \pi(x)$$

$Q_{\sqrt{\pi}}$ produces local moves that are *asymptotically* π*-reversible as* $\sigma \downarrow 0$
$\rightsquigarrow$ $Q_{\sqrt{\pi}}$ is appropriate to design local moves targeting π

Locally balanced proposals

Class of proposals considered: "point-wise informed" proposals of the form

$$Q_{g,\sigma}(x, dy) \propto g\left(\frac{\pi(y)}{\pi(x)}\right) K_\sigma(x, dy) \qquad \text{for some } g : \mathbb{R}_+ \to \mathbb{R}_+$$

Definition: $\{Q_\sigma(x, dy)\}_{\sigma>0}$ is *locally balanced* if Q_σ is Π_σ-reversible and $\Pi_\sigma \Rightarrow \Pi$ as $\sigma \downarrow 0$.

Theorem: Let K_σ be dx-reversible and $K_\sigma(x, \cdot) \Rightarrow \delta_x(\cdot)$ as $\sigma \downarrow 0$. Then $\{Q_{g,\sigma}\}_{\sigma>0}$ is locally balanced iff

$$g(t) = t\, g(1/t) \qquad \forall t > 0$$

NB: some regularity assumptions on g and π needed to guarantee integrability.

Question: locally-balanced proposals are asymptotically π-reversible as $\sigma \downarrow 0$. Intuitively, this is a good features for a local Metropolis-Hastings proposal. Can we say something more explicit in terms of efficiency of the induced MCMC ?

"Answer": in high-dimensions locally-balanced proposals are maximal elements in terms of Peskun ordering.

Peskun Ordering

Lemma (Peskun ordering)

Let P_1 and P_2 be π-reversible Markov kernels on a finite Ω. If

$$P_1(x,y) \leq P_2(x,y) \qquad \forall x \neq y \tag{1}$$

then the Spectral Gaps and Asymptotic Variances of P_1 and P_2 satisfy

$$\begin{aligned} \mathrm{Gap}(P_1) \ &\leq \ \mathrm{Gap}(P_2) \\ \mathrm{Var}_\pi(h, P_1) \ &\geq \ \mathrm{Var}_\pi(h, P_2) \quad \forall h : \Omega \to \mathbb{R}\,. \end{aligned}$$

Intuition: if (1) holds, then P_2 is more efficient than P_1.

Peskun Ordering

Lemma (Peskun ordering with constant)

Let P_1 and P_2 be π-reversible Markov kernels on a finite Ω. If

$$P_1(x,y) \leq c\, P_2(x,y) \qquad \forall x \neq y \tag{2}$$

for some $c > 0$, then the Spectral Gaps and Asymptotic Variances of P_1 and P_2 satisfy

$$\begin{aligned} \mathrm{Gap}(P_1) &\leq c\, \mathrm{Gap}(P_2) \\ \mathrm{Var}_\pi(h, P_1) &\geq c\, \mathrm{Var}_\pi(h, P_2) + (1-c)\mathrm{Var}_\pi(h) \quad \forall h : \Omega \to \mathbb{R}\,. \end{aligned}$$

Intuition: if (2) holds, then P_2 is $\frac{1}{c}$-times more efficient than P_1.

Asymptotic Peskun ordering of locally balanced proposals

Consider Ω finite and $K(x,\cdot) = \text{Unif}(N(x))$. Given $Z_g(x) = \sum_{y \in N(x)} g\left(\frac{\pi(y)}{\pi(x)}\right)$ define

$$c_g \quad = \quad \sup_{x \in \Omega,\ y \in N(x)} \frac{Z_g(y)}{Z_g(x)} \ \geq \ 1\ .$$

Theorem

Let $g : \mathbb{R}_+ \to \mathbb{R}_+$ and $\tilde{g}(t) = \min\{g(t), t\, g(1/t)\}$. Then the MH kernels obtained from the proposals Q_g and $Q_{\tilde{g}}$ respectively satisfy

$$P_g(x,y) \ \leq \ c_g^2\ P_{\tilde{g}}(x,y) \qquad \forall\, x \neq y\, .$$

Intuition: for every g there is a locally-bal. $\tilde{g}$ which is more efficient modulo c_g^2.

Asymptotic regime

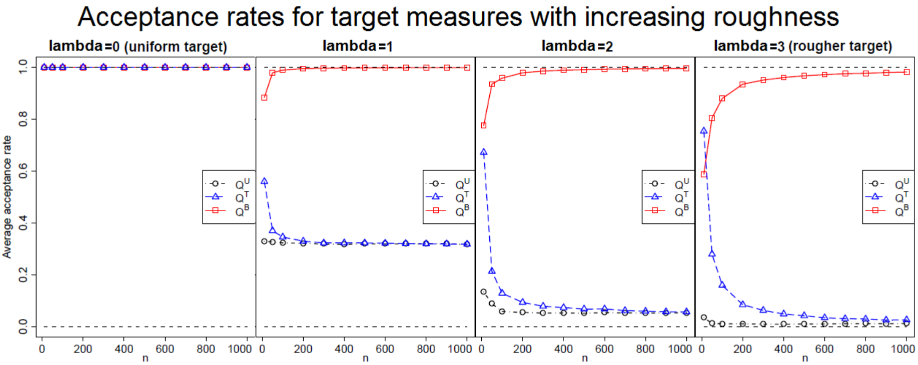
In many contexts $c_g \to 1$ as the dimension of Ω goes to infinity. In these cases locally balanced proposals are **asymptotically optimal in the Peskun sense**

Example: sampling matchings

$\Omega_n = \{\text{perfect matchings of } n+n \text{ bipartite graph}\}$

$\pi(x) \propto \prod_{e \in x} w_e$ with $w_e \overset{iid}{\sim} \text{LogNormal}(0, \lambda^2)$ $\quad N(x) = \{\text{switching two edges}\}$

$Q_U(x,y) \propto \mathbb{1}_{N(x)}(y)$ $\quad Q_\pi(x,y) \propto \pi(y)\mathbb{1}_{N(x)}(y)$ $\quad Q_{\sqrt{\pi}}(x,y) \propto \sqrt{\pi(y)}\mathbb{1}_{N(x)}(y)$

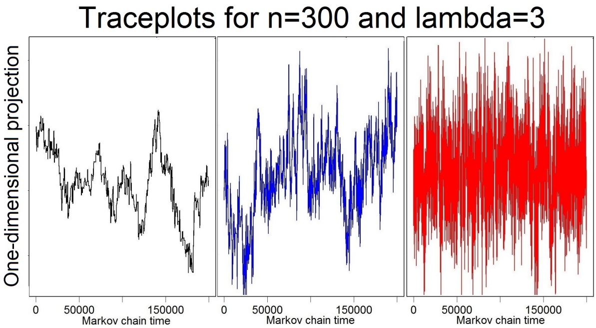


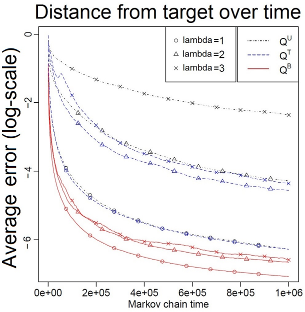
Example: sampling matchings

$\Omega_n = \{\text{perfect matchings of } n \times n \text{ bipartite graph}\}$

$\pi(x) \propto \prod_{e\in x} w_e$ with $w_e \overset{iid}{\sim} \text{LogNormal}(0, \lambda^2)$ $\quad N(x) = \{\text{swapping two edges}\}$

$Q_U(x,y) \propto \mathbb{1}_{N(x)}(y)$ $\quad Q_\pi(x,y) \propto \pi(y)\mathbb{1}_{N(x)}(y)$ $\quad Q_{\sqrt{\pi}}(x,y) \propto \sqrt{\pi(y)}\mathbb{1}_{N(x)}(y)$





Example: Ising model

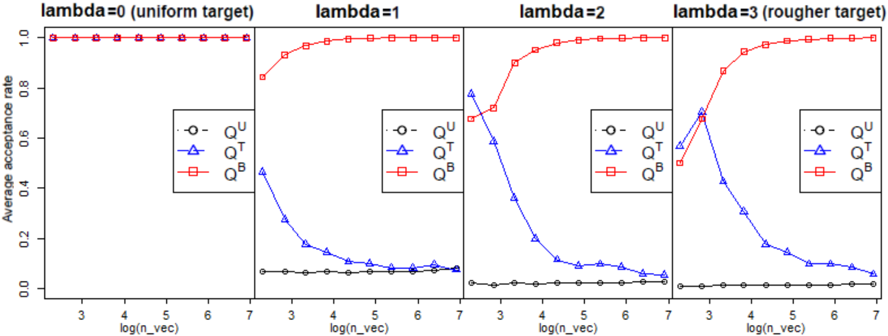
$\Omega_n = \{-1,1\}^V$ where V is the $n \times n$ lattice

$\pi(x) \propto \exp\{\lambda(\sum_{i\in V} \alpha_i x_i + \sum_{j\sim i} x_i x_j)\}$ with $\alpha_i \overset{iid}{\sim} \text{Unif}(-\sigma,\sigma)$

$N(x) = \{\text{flipping one bit}\}$

$Q_U(x,y) \propto \mathbb{1}_{N(x)}(y)$ $\quad Q_\pi(x,y) \propto \pi(y)\mathbb{1}_{N(x)}(y)$ $\quad Q_{\sqrt{\pi}}(x,y) \propto \sqrt{\pi(y)}\mathbb{1}_{N(x)}(y)$

Acceptance rates for target measures with increasing roughness



Optimal choice of locally-balanced proposal?

Question: is there an optimal choice of Q_g among the ones with $g(t) = tg(1/t)$?
Many different choices of g lead to locally-balanced proposals

	$g(t) = \sqrt{t}$	$g(t) = \frac{t}{1+t}$	$g(t) = 1 \wedge t$
$Q_g(x, y) \propto$	$\sqrt{\pi(y)}K(x, y)$	$\frac{\pi(y)}{\pi(x)+\pi(y)}K(x, y)$	$\left(1 \wedge \frac{\pi(y)}{\pi(x)}\right)K(x, y)$

Partial answers:

- Reminescent of choosing an expression for the acceptance probability in the accept/reject step. In that case the MH choice $1 \wedge \frac{\pi(y)}{\pi(x)}$ Peskun-dominates all others.
- In our case, there is no Peskun-ordering among couples of locally-balanced Q_g. Also, the restriction $g(t) \leq 1$, so the class of admissible g's in broader.
- In some simplified scenarios (e.g. $\{0, 1\}^n$ with product target) the optimal choice turned out to be $g(t) = \frac{t}{1+t}$, i.e. $\frac{\pi(y)}{\pi(x)+\pi(y)}K(x, y)$
- In simulations, different locally-balanced proposals performed very similar.

Connection to MALA

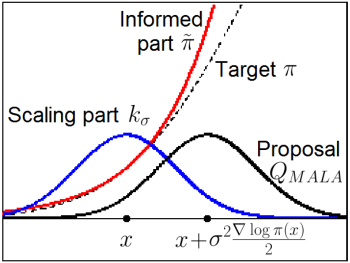
In continuous spaces, to sample from Q_g, one needs to replace $\frac{\pi(y)}{\pi(x)}$ with some approximation $\tilde{\pi}_x(y)$.

E.g.: $\tilde{\pi}_x(y) = \exp\left(\nabla \log \pi(x)(y-x)\right)$ 1st-order Taylor expansion leads to MALA

$$Q_{MALA}(x,y) = N\left(x + \sigma^2 \frac{\nabla \log \pi(x)}{2}, \sigma^2 \mathbb{I}_n\right)$$

$$\propto \sqrt{\exp\left(\nabla \log \pi(x)(y-x)\right)} \exp\left(\frac{|y-x|^2}{2\sigma^2}\right)$$

$$= \sqrt{\tilde{\pi}_x(y)} K_\sigma(x, dy)$$



NB: choice of $g(t)$, $K_\sigma(x, dy)$ and approximation $\tilde{\pi}$ provide large flexibility.

Application to Multiple-try MCMC

Original Multiple-Try kernel (MTM)[1]

1. Sample $y_1, \dots, y_N \overset{iid}{\sim} K_\sigma(x, \cdot)$
2. Choose y from $(y_1, \dots, y_N)$ with probabilities $\propto (\pi(y_1), \dots, \pi(y_N))$
3. Sample $x_1^*, \dots, x_{N-1}^* \overset{iid}{\sim} K_\sigma(y, \cdot)$ and set $x_N^* = x$
4. Accept y with probability $1 \wedge \frac{\pi(x_1^*)+\cdots+\pi(x_N^*)}{\pi(y_1)+\cdots+\pi(y_N)}$

PROBLEM: as $N \to \infty$ MTM converges to MH with $Q_\pi(x, dy) \propto \pi(y) K_\sigma(x, dy)$
$\Rightarrow$ inherently mis-specified for local moves!

[1] Liu&al.(2000)The multiple-try method and local optimization in metropolis sampling. JASA

Application to Multiple-try MCMC

Original Multiple-Try kernel (MTM)[1]

1. Sample $y_1, \dots, y_N \overset{iid}{\sim} K_\sigma(x, \cdot)$
2. Choose y from $(y_1, \dots, y_N)$ with probabilities $\propto (\pi(y_1), \dots, \pi(y_N))$
3. Sample $x_1^*, \dots, x_{N-1}^* \overset{iid}{\sim} K_\sigma(y, \cdot)$ and set $x_N^* = x$
4. Accept y with probability $1 \wedge \frac{\pi(x_1^*)+\dots+\pi(x_N^*)}{\pi(y_1)+\dots+\pi(y_N)}$

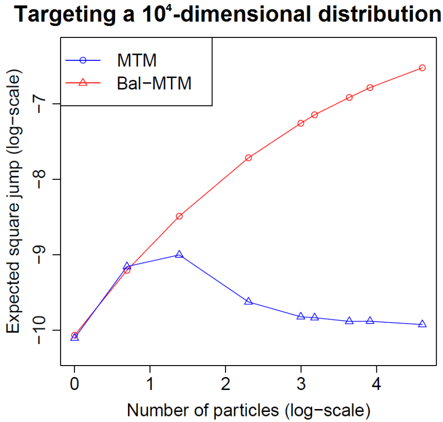
PROBLEM: as $N \to \infty$ MTM converges to MH with $Q_\pi(x, dy) \propto \pi(y) K_\sigma(x, dy)$
$\Rightarrow$ inherently mis-specified for local moves!

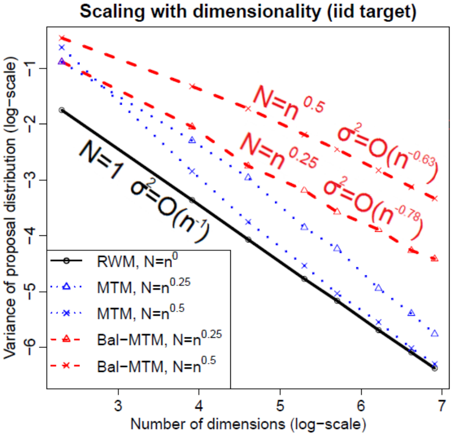
Locally balanced MTM kernel (Bal-MTM)

1. Sample $y_1, \dots, y_N \overset{iid}{\sim} K_\sigma(x, \cdot)$
2. Choose y from $(y_1, \dots, y_N)$ with probabilities $\propto (\sqrt{\pi}(y_1), \dots, \sqrt{\pi}(y_N))$
3. Sample $x_1^*, \dots, x_{N-1}^* \overset{iid}{\sim} K_\sigma(y, \cdot)$ and set $x_N^* = x$
4. Accept y with probability $1 \wedge \frac{\sqrt{\pi}(y)}{\sqrt{\pi}(x)} \frac{\sqrt{\pi}(x_1^*)+\dots+\sqrt{\pi}(x_N^*)}{\sqrt{\pi}(y_1^*)+\dots+\sqrt{\pi}(y_N^*)}$

[1] Liu&al.(2000)The multiple-try method and local optimization in metropolis sampling. JASA

Example: 10^4 dimensional target (iid t-student)





Summary

- MCMC based on uninformed proposals (i.e. RWM) can be slow.
- Biasing proposals towards high-probability regions is a natural thing to do (e.g. gradient-based MCMC), but how this should be done is not obvious.
- Framework of locally-balanced proposal can provide useful guidance to design informed proposals, especially in discrete spaces.

Things which we didn't discuss:

- Approximate versions to achieve a good cost-vs-efficiency trade-off?
- Interpolation between $\sigma \downarrow 0$ and $\sigma \uparrow \infty$?
- Connections to continuous time versions
- ...

References

G. Zanella, **Design of informed local proposals for MCMC in discrete spaces**. *In preparation.*

G. Zanella, **Random Partition Models and Complementary clustering of Anglo-Saxon place-names**. *Annals of Applied Statistics*, 2015.