Monte Carlo Inference for Intractable Likelihoods

George Deligiannidis[1]
Department of Statistics, Oxford University

Inference in Mathematical Biology - 23/05/2022

[1] joint work w. A.Doucet, M.Pitt, R.Kohn, S.Schmon, P.Vanetti, R.Cornish, L.Middleton, P.Jacob

Latent Variable Models

- Assume $(Z_t)_{t \geqslant 1}$ are i.i.d. latent random variables such that

$$Z_t \overset{\text{i.i.d.}}{\sim} \mu_\theta(\cdot), \quad X_t | (Z_t = z) \sim g_\theta(\cdot | z) \text{ for } t = 1, \ldots, T,$$

  where $X_1, \ldots, X_T$ are the observations.
- The likelihood of θ associated to $X_{1:T} = x_{1:T}$ is

$$p_\theta(x_{1:T}) = \prod_{t=1}^{T} p_\theta(x_t), \text{ where } p_\theta(x_t) = \int \mu_\theta(z_t) g_\theta(x_t | z_t) \, dz_t.$$

- Probabilistic models ubiquitous in ML & Statistics.
- Also state-space models, aka Hidden Markov Models.

Bayesian Inference and MCMC

- Prior density $p(\theta)$.
- Intractable likelihood $p_\theta(x_{1:T}) = \int \cdots \int p_\theta(x_{1:T}, z_{1:T}) dz_{1:T}$.
- Posterior
$$\pi(\theta) = p(\theta | x_{1:T}) \propto p_\theta(x_{1:T}) p(\theta)$$
allows to perform uncertainty quantification.
- Standard MCMC schemes target
$$p(\theta, z_{1:T} | x_{1:T}) \propto p_\theta(z_{1:T}, x_{1:T}) p(\theta)$$
by sampling alternately $Z_{1:T} \sim p_\theta(\cdot | x_{1:T})$ and $\theta \sim p(\cdot | x_{1:T}, Z_{1:T})$.

Standard MCMC Approaches

- **Problem** 1: Difficult to sample from $p_\theta(z_{1:T} | x_{1:T})$.
- **Problem** 2: Even when implementable, can converge very slowly.
- **Problem** 3: For complex generative model, only forward simulation from $\{Z_t\}$ is possible.

Ideal Marginal Metropolis-Hastings

Sampling MH kernel $P_{\mathrm{MH}}(\vartheta, \cdot)$

- Sample $\vartheta' \sim q(\cdot | \vartheta)$.
- With probability

$$1 \wedge \frac{p_{\vartheta'}(x_{1:T})}{p_{\vartheta}(x_{1:T})} \frac{p(\vartheta') q(\vartheta | \vartheta')}{p(\vartheta) q(\vartheta' | \vartheta)},$$

  output ϑ'; otherwise, output ϑ.
- **Problem**: MH cannot be implemented for intractable $p_{\vartheta}(x_{1:T})$.

Pseudo-Marginal Algorithm

- **Running Assumption**: one has access to a *non-negative unbiased estimator* of $p_\theta(x_{1:T})$ obtained by sampling $U \sim m_\theta(\cdot)$ and returning $\widehat{p}_\theta(x_{1:T}; U)$.

Sampling PM kernel $P_{\mathrm{PM}}\{(\vartheta, U), \cdot\}$

- Sample $\vartheta' \sim q(\cdot|\vartheta)$.
- Sample $U' \sim m_{\vartheta'}(\cdot)$ and compute $\widehat{p}_{\vartheta'}(x_{1:T}; U')$.
- With probability

$$1 \wedge \frac{p_{\vartheta'}(x_{1:T}; U')}{p_{\vartheta}(x_{1:T}; U)} \frac{p(\vartheta')\, q(\vartheta|\vartheta')}{p(\vartheta)\, q(\vartheta'|\vartheta)} \underbrace{\frac{\widehat{p}_{\vartheta'}(x_{1:T}; U') / p_{\vartheta'}(x_{1:T})}{\widehat{p}_{\vartheta}(x_{1:T}; U) / p_{\vartheta}(x_{1:T})}}_{\text{noise } LRE(\vartheta, \vartheta')},$$

output (ϑ', U'); otherwise, output (ϑ, U).

Pseudo-Marginal Algorithm

- **Fact**: PM algorithm is a valid MCMC algorithm to sample $\pi(\theta)$.
- PM is a standard MH algorithm using proposal $q(\theta'|\theta)\, m_{\theta'}(u')$ targetting

$$\int \overline{\pi}(\theta, u)\mathrm{d}u = \pi(\theta) \underbrace{\frac{\int \widehat{p}_\theta(x_{1:T}; u)\, m_\theta(u)\, \mathrm{d}u}{p_\theta(x_{1:T})}}_{=1 \text{ by unbiasedness}} = \pi(\theta).$$

- To use MCMC, having access to an non-negative unbiased estimator of likelihood is sufficient.

Likelihood Estimators

- For latent variable models, one has

$$p_\theta(x_{1:T}) = \prod_{t=1}^{T} p_\theta(x_t), \text{ where } p_\theta(x_t) = \int \mu_\theta(z_t)\, g_\theta(x_t|\, z_t)\, \mathrm{d}z_t.$$

- A non-negative unbiased estimator is given by

$$\widehat{p}_\theta(x_{1:T}; U) = \prod_{t=1}^{T} \left\{ \frac{1}{N} \sum_{k=1}^{N} g_\theta\left(x_t|\, Z_t^k \right) \right\}, \quad Z_t^k \overset{\text{i.i.d.}}{\sim} \mu_\theta(\cdot),$$

- For state-space models, an alternative is to use particle filters
- The estimator is unbiased, relative variance is bounded uniformly over T if $N \propto T$.

Asymptotic Variance of MCMC Estimators

- Consider estimate $\frac{1}{t}\sum_{k=1}^{t} h(\vartheta_k)$ of $\int h(\theta)\,\pi(\theta)\,d\theta$ where $\vartheta_k \sim K(\vartheta_{k-1},\cdot)$ for K kernel $\pi-$invariant.
- This estimate satisfies a $\sqrt{t}-$CLT with asymptotic variance

$$\mathrm{Var}_{\pi}(h) \times \mathrm{IACT}(K,h),$$

where

$$\mathrm{IACT}(K,h) = 1 + 2\sum_{i=1}^{\infty} \mathrm{corr}_{\vartheta_0\sim\pi,\vartheta_i\sim Q^i}\left\{h(\vartheta_0), h(\vartheta_i)\right\}.$$

- IACT measures the loss in precision relative to using i.i.d. samples from the target.
- Intuitively, $\mathrm{IACT}(P_{\mathrm{PM}},h) \nearrow$ as $\sigma^2(\theta) := \mathrm{Var}\{\log \widehat{p}_\theta(x_{1:T};U)\} \nearrow$.
- Empirical results confirm intuition.

Log(IACT) of PM for State-Space Model

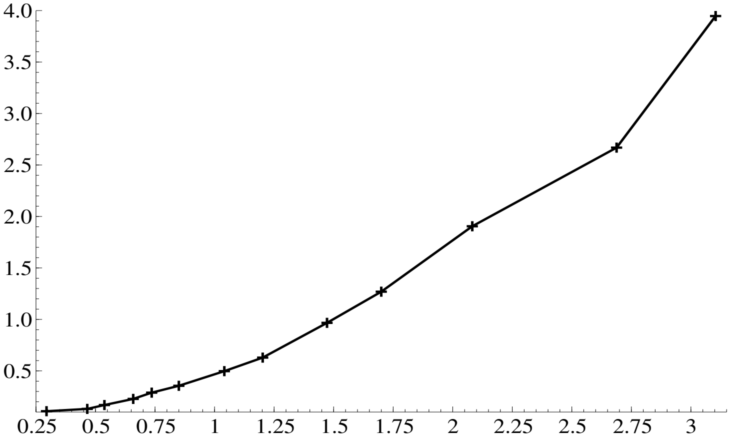

Figure: Average over the 9 parameter components of the log-integrated autocorrelation time of PM chain as a function of $\sigma^2 := \sigma^2\left(\overline{\theta}\right)$ for $\overline{\theta}$ central parameter value.

Computational Complexity vs Statistical Efficiency

- To reduce variance of log-likelihood ratio we need more particles, i.e. more compute per iteration.
- **Aim**: Minimize the "computational time" w.r.t. $\sigma^2 := \sigma^2\left(\overline{\theta}\right)$

$$\mathrm{CT}\left(P_{\mathrm{PM}}, h\right) = \frac{\mathrm{IACT}\left(P_{\mathrm{PM}}, h\right)}{\sigma^2}$$

  as $\sigma^2 \propto 1/N$ and computational efforts proportional to N.
- Dependence of $\sigma^2(\theta)$ on θ unclear.
- Direct analysis of CT is very complex as intractable.

Computational time for state-space model

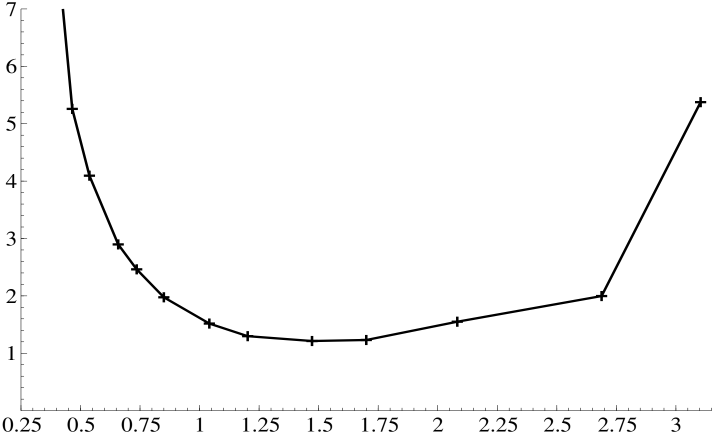

Figure: Computational time as a function of σ

Asymptotic Properties of Noise for PM

- Let $\theta \in \mathbb{R}^d$: asymptotic study of MCMC always relies on $d \to \infty$ and T fixed, here fixed d and $T \to \infty$.
- **Proposition**. Let $N = \beta T$ for $\alpha > 0$ then the error in log-likelihood $E_T = \log\{\widehat{p}_\theta(X_{1:T}; U)/p_\theta(X_{1:T})\}$ satisfies CLT [2]

$$E_T | \mathcal{X}^T \Rightarrow \mathcal{N}\left\{-\sigma^2(\theta)/2, \sigma^2(\theta)\right\} \text{ (proposal } U \sim m_\theta(\cdot))$$

$$E_T | \mathcal{X}^T \Rightarrow \mathcal{N}\left\{\sigma^2(\theta)/2, \sigma^2(\theta)\right\} \text{ (equilibrium } U \sim \overline{\pi}(\cdot|\theta))$$

  and, at equilibrium,

$$\log LRE_T\left(\vartheta, \vartheta + \frac{\xi}{\sqrt{T}}\right)\Bigg| \mathcal{X}^T \Rightarrow \mathcal{N}\left\{-2\sigma^2(\theta), 2\sigma^2(\theta)\right\}.$$

- PM estimator needs $N \propto T$ to control variance of $\log LRE_T$.

[2] Bérard, Del Moral and Doucet, A., 2014. Elect. J. Proba., 19, pp.1-28.

Empirical vs Assumed Distributions for State-Space Model

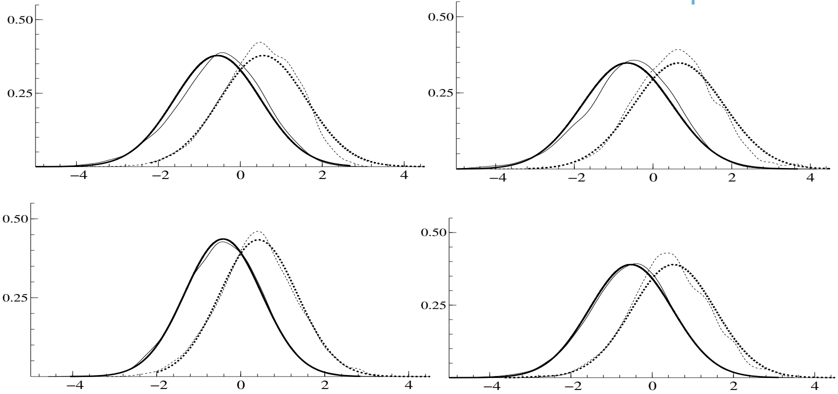

Figure: Empirical distributions (dashed) vs assumed Gaussians (solid) of Z at $\overline{\theta}$ (left) and marginalized over samples from $\pi(\theta)$ (center) and $\int \pi(d\vartheta)\, q(\theta|\,\vartheta)$ (right) for $T = 40$, $T = 300$ and $T = 2700$.

Large Sample Analysis of PM

- **Assumption**: Posterior concentrates at rate $1/\sqrt{T}$ around $\widehat{\theta}_T \xrightarrow{P} \overline{\theta}$ and proposal is $\theta'|\theta \sim \mathcal{N}(\theta, 1/T)$ say.
- Center chain at $\widehat{\theta}_T$ and rescale by $\sqrt{T}$:

$$(\Theta_T, \mathbf{E}_T) := \{\widetilde{\vartheta}_i := \sqrt{T}(\vartheta_i - \widehat{\theta}_T), \qquad E_i := \log\left\{\widehat{p}_{\vartheta_i}(X_{1:T}; U_i) / p_{\vartheta_i}(X_{1:T})\right\}\}_{i \geqslant 0}.$$

- **Proposition:** $(\Theta_T, \mathbf{E}_T)_{T \geqslant 1}$ converges weakly as $T \to \infty$ to a stationary MC of kernel P^{σ}_{PM} given for $(\widetilde{\theta}, e) \neq (\widetilde{\theta}', e')$ by

$$u(\widetilde{\theta}' - \widetilde{\theta})\varphi(e'; -\sigma^2/2, \sigma^2) \min\left\{1, \frac{\varphi(\widetilde{\theta}'; 0, \Sigma)}{\varphi(\widetilde{\theta}; 0, \Sigma)} \exp\left(e' - e\right)\right\} d\theta' de'$$

with invariant density $\varphi(\widetilde{\theta}; 0, \Sigma)\varphi(e; \sigma^2/2, \sigma^2)$ for $\sigma^2 := \sigma^2\left(\overline{\theta}\right)$.

Example on Random Effect Models

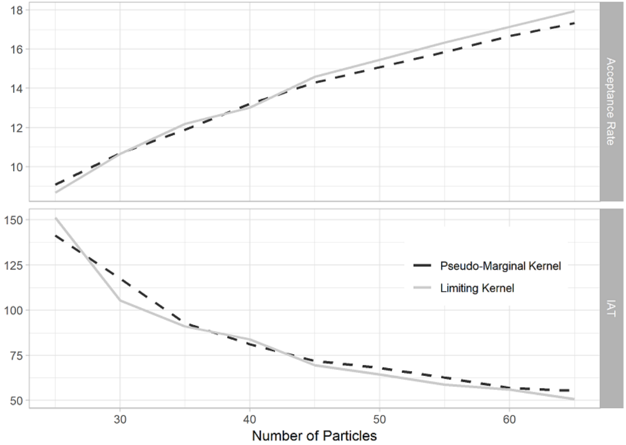

Figure: Acceptance rate (top) and IACT (bottom) for P_{PM} (solid) and P^{σ}_{PM} (dashed) as a function of σ.

Practical Guidelines

- Use P_{PM}^{σ} as a proxy for P_{PM} and optimize

$$\mathrm{CT}(P_{\mathrm{PM}}^{\sigma}, h) = \frac{\mathrm{IACT}(P_{\mathrm{PM}}^{\sigma}, h)}{\sigma^2},$$

  as we expect

$$\mathrm{CT}(P_{\mathrm{PM}}, h) \overset{T\to\infty}{\to} \mathrm{CT}(P_{\mathrm{PM}}^{\sigma}, h).$$

- For "good" proposals, select $\sigma \approx 1.0$ whereas for "poor" proposals, select $\sigma \approx 1.7$.
- When you have no clue about the proposal efficiency,
- If $\sigma_{opt} = 1.0$ and you pick $\sigma = 1.7$, computing time increases by $\approx 150\%$.
- If $\sigma_{\mathrm{opt}} = 1.7$ and you pick $\sigma = 1.0$, computing time increases by $\approx 50\%$.
- If $\sigma_{\mathrm{opt}} = 1.0$ or $\sigma_{\mathrm{opt}} = 1.7$ and you pick $\sigma = 1.2 - 1.3$, computing time increases by $\approx 15\%$.

Computational cost of PM

- PM scales in $\mathcal{O}(T^2)$ at each iteration as $N \propto T$ required to control σ.
- For i.i.d. data, simulated likelihood works well with $N \propto T^{1/2+\varepsilon}$.
- Is it possible to improve PM?
- **Problem**: $p_{\theta'}(x_{1:T})/p_{\theta}(x_{1:T})$ is estimated by dividing independent estimators of $p_{\theta}(x_{1:T})$ and $p_{\theta'}(x_{1:T})$.

Correlated Pseudo-Marginal Algorithm

- Likelihood estimator $\widehat{p}_\theta(x_{1:T}; U)$ uses $U \sim m_\theta(\cdot)$.
- Use reparameterization trick so that $U \sim \mathcal{N}(0, I)$ (Glasserman, 1990; Kingma & Welling, 2014).
- Correlate estimators $\widehat{p}_\theta(x_{1:T}; U)$ of $p_\theta(x_{1:T})$ and $\widehat{p}_{\theta'}(x_{1:T}; U')$ of $p_{\theta'}(x_{1:T})$ using for $\rho > 0$

$$U' = \rho U + \sqrt{1-\rho^2}\varepsilon, \; \varepsilon \sim \mathcal{N}(0, I).$$

Correlated Pseudo-Marginal Algorithm

Sampling CPM kernel $P_{\mathrm{CPM}}\left\{(\vartheta, U), \cdot\right\}$

- Sample $\vartheta' \sim q\left(\cdot \mid \vartheta\right)$.
- Sample $\varepsilon \sim \mathcal{N}\left(0, I\right)$, set $U' = \rho U + \sqrt{1-\rho^2}\varepsilon$, compute $\widehat{p}_{\vartheta'}\left(x_{1:T}; U'\right)$.
- With probability

$$1 \wedge \frac{p_{\vartheta'}\left(x_{1:T}; U'\right)}{p_{\vartheta}\left(x_{1:T}; U\right)} \frac{p\left(\vartheta'\right) q\left(\vartheta \mid \vartheta'\right)}{p\left(\vartheta\right) q\left(\vartheta' \mid \vartheta\right)} \underbrace{\frac{\widehat{p}_{\vartheta'}\left(x_{1:T}; U'\right) / p_{\vartheta'}\left(x_{1:T}\right)}{\widehat{p}_{\vartheta}\left(x_{1:T}; U\right) / p_{\vartheta}\left(x_{1:T}\right)}}_{\text{noise } LRE(\vartheta, \vartheta')},$$

output (ϑ', U'); otherwise, output (ϑ, U).

Asymptotic Properties of Noise for CPM

- **Proposition**. Let $N \to \infty$ as $T \to \infty$ with $N = o(T)$. For U at equilibrium and $U' = \rho U + \sqrt{1-\rho^2}\varepsilon$ with $\rho = \exp\left(-\psi \frac{N}{T}\right)$ then

$$\log LRE_T\left(\vartheta, \vartheta + \frac{\xi}{\sqrt{T}}\right)\Big| \mathcal{X}^T, \mathcal{U}^T \Rightarrow \mathcal{N}\left\{-\frac{\kappa^2(\theta)}{2}, \kappa^2(\theta)\right\}.$$

- PM estimator needs $N \propto T$ to control variance of $\log LRE_T$, CPM can use $N \propto \log T$.
- CLT is conditional on the observations and the auxiliary variables.
- Asymptotically the log-ratio decouples from the current location of the chain, more robust.

Large sample analysis of CPM

- Let $\Theta_T := \{\widetilde{\vartheta}_i = \sqrt{T}(\vartheta_i - \widehat{\theta}_T)\}_{i \geqslant 0}$ be the stationary *non-Markovian* sequence of CPM targetting $p(\theta | X_{1:T})$.
- **Proposition**: $\{\Theta_T\}_{T \geqslant 1}$ converges weakly as $T \to \infty$ to a stationary MC of kernel $P^{\kappa}_{\mathrm{CPM}}$ given for $\widetilde{\theta} \neq \widetilde{\vartheta}$ by

$$\upsilon(\widetilde{\theta}' - \widetilde{\theta})\mathbb{E}_{R \sim \mathcal{N}(-\kappa^2/2, \kappa^2)} \left[1 \wedge \frac{\varphi(\widetilde{\theta}'; 0, \Sigma)}{\varphi(\widetilde{\theta}; 0, \Sigma)} \exp\left(R\right) \right] \mathrm{d}\widetilde{\theta}'$$

  where $\kappa := \kappa\left(\overline{\theta}\right)$ and invariant density $\varphi\left(\widetilde{\theta}; 0, \Sigma\right)$.
- CPM is more subtle than PM: $\mathrm{CT}(P_{\mathrm{CPM}}, h) \to \mathrm{CT}(P^{\kappa}_{\mathrm{CPM}}, h)$ only if $\sqrt{T}/N = O\left(1\right)$.
- Further analysis provides guidelines on selection of parameters ψ, N.

Example: Gaussian Latent Variable Model

MH ($T = 8,000$)		IACT(θ)	
		15.6	
PM			
N		RIACT(θ)	RCT(θ)
5000		2.2	11210
CPM ($\rho = 0.9963$)			
N	κ	RIACT(θ)	RCT(θ)
10	3.1	14.0	126.2
20	2.2	4.7	93.3
25	2.0	2.8	69.3
35	1.7	1.7	61.1
56	1.3	1.6	87.0

Here $\text{RIACT} = \text{IACT}/\text{IACT}_{MH}$ and $\text{RCT} = N \times \text{RIACT}$. Improvement by 180 fold.

Discussion

- MCMC can use unbiased likelihood estimator.
- The smaller the variance, the better the performance.
- Precise guidelines for optimizing computational complexity/statistical efficiency are available.
- Extensions to HMC and Slice sampling are feasible but poorly understood.
- Scalability with T remains unimpressive.

MCMC for Large Datasets

- Consider
$$p(\theta | x_{1:T}) \propto p(\theta) \prod_{t=1}^{T} p_\theta(x_t),$$
where $p_\theta(x_t)$ can now be evaluated but $T \gg 1$.
- Standard MCMC like MH too expensive $O(T)$ at each iteration.
- Subsampling MCMC methods
  - SGLD (Pages & Lamberton, 2002; Welling & Teh, 2011; Chatterji et al., 2018),
  - Subsampling MH (Bardenet et al., 2014; Korattikara, 2014).
  - Firefly (McLaurin & Adams, 2014).
  - PDMP (Bouchard-Cote et al. 2018, Bierkens et al. 2018).

Factorized MH for Large Datasets

- Use a factorized acceptance probability.

Sampling FMH kernel $P_{\mathrm{FMH}}(\vartheta, \cdot)$

- Sample $\vartheta' \sim q(\cdot | \vartheta)$.
- With probability

$$\alpha_{\mathrm{FMH}}(\vartheta, \vartheta') = \underbrace{1 \wedge \frac{p(\vartheta')\, q(\vartheta | \vartheta')}{p(\vartheta)\, q(\vartheta' | \vartheta)}}_{:= \alpha_0(\vartheta, \vartheta')} \prod_{t=1}^{T} \underbrace{1 \wedge \frac{p_{\vartheta'}(x_t)}{p_{\vartheta}(x_t)}}_{:= \alpha_t(\vartheta, \vartheta')},$$

output ϑ'; otherwise, output ϑ.

Properties of Factorized MH

- P_{FMH} is $p\left(\theta \middle| x_{1:T}\right)$-reversible thus $p\left(\theta \middle| x_{1:T}\right)$-invariant
- Lower acceptance probability

$$\alpha_{\mathrm{FMH}}\left(\vartheta, \vartheta'\right) \leqslant \alpha_{\mathrm{MH}}\left(\vartheta, \vartheta'\right).$$

- Peskun's theorem implies

$$\mathrm{IACT}\left(P_{\mathrm{FMH}}, h\right) \geqslant \mathrm{IACT}\left(P_{\mathrm{MH}}, h\right).$$

Re-interpretation of acceptance probability

- Define Bernoulli $B_t \overset{\text{ind}}{\sim} \text{Ber}(1-\alpha_t(\vartheta,\vartheta'))$ for $t \in \{0,1,\ldots,T\}$.
- We have

$$\begin{aligned}
\mathbb{P}(\exists t \in \{0,\ldots,T\} : B_t = 1) &= 1 - \mathbb{P}(\forall t \in \{0,\ldots,T\} : B_t = 0) \\
&= 1 - \prod_{t=0}^{T} \alpha_t(\vartheta,\vartheta') \\
&= 1 - \alpha_{\text{FMH}}(\vartheta,\vartheta')
\end{aligned}$$

- Suggest sampling sequentially $B_0,\ldots,B_t$ and stop first time $B_t = 1$ (reject). If $B_0 = B_1 = \cdots = B_T = 0$, then accept: requires going through whole dataset!
- Assume we have $\alpha_t(\vartheta,\vartheta') \geqslant \overline{\alpha}$; e.g. Lipschitz assumption on $\theta \mapsto \log p_\theta(x_t)$ uniform in t.
- Accepted proposal requires $O((1-\overline{\alpha})T)$ likelihood evaluations.

Properties

- Under concentration, to ensure $(1-\overline{\alpha})T = O(1)$ and geometric ergodicity, one needs proposal of s.d. $1/T$ for FMH instead of $1/\sqrt{T}$ for MH.
- Alternative decomposition

$$\pi(\theta) \propto \underbrace{p(\theta)\prod_{t=1}^{T}\widehat{p}_{\theta}(x_t)}_{\widehat{\pi}(\theta)\ \text{approximation}} \prod_{t=1}^{T} \underbrace{\frac{p_{\theta}(x_t)}{\widehat{p}_{\theta}(x_t)}}_{\text{reduce variability}}$$

so using a $\widehat{\pi}(\theta)$-reversible proposal

$$\alpha_{\mathrm{FMH}}(\vartheta,\vartheta') = \prod_{t=1}^{T} 1 \wedge \frac{p_{\vartheta'}(x_t)/\widehat{p}_{\vartheta'}(x_t)}{p_{\vartheta}(x_t)/\widehat{p}_{\vartheta'}(x_t)}$$

- In this scenario, one can ensure $(1-\overline{\alpha})T = O(1)$, geometric ergodicity and proposal of s.t.d. $1/\sqrt{T}$.

Some References

1. G. D., A.Doucet, & M.K. Pitt, "The correlated pseudo-marginal method", *J. Royal Stat. Soc.* B, 2018.
2. A.Doucet., M.K. Pitt, G. D., & R. Kohn, "Efficient implementation of MCMC when using an unbiased likelihood estimator", *Biometrika*, 2015.
3. S. Schmon, G. D., A.Doucet & M.K. Pitt, "Large sample asymptotics of the PM algorithm", arXiv:1806.10060.
4. P. Vanetti, A. Bouchard-Cote, G. D. & A.Doucet, "Piecewise-deterministic MCMC", arXiv:1707.05296.
5. Middleton, L., Deligiannidis, G., Doucet, A., Jacob, P. E. (2020). Unbiased Markov chain Monte Carlo for intractable target distributions. Electronic Journal of Statistics, 14(2), 2842-2891.