Comparison of Markov chains via weak Poincaré inequalities

with application to pseudo-marginal MCMC

Sam Power

University of Bristol

14 October, 2022

Links & Acknowledgements

- Main paper today: arXiv 2112.05605;
- Follow-up: arXiv 2208.05239
- All joint work with
  - Christophe Andrieu (Bristol)
  - Anthony Lee (Bristol)
  - Andi Q. Wang (Bristol $\rightsquigarrow$ Warwick)

- Funded by Bayes4Health EPSRC Grant

Setting: Task

- Task: simulation, integration in complex models
  - posterior inference
  - gradient estimation in intractable models
  - ...

Setting: Methodology

- MCMC
- Basic Idea:
  - Specify ‘target measure’ π.
  - Design ergodic Markov kernel P such that $\pi P = \pi$.
  - Initialise X_0.
  - For $n \geqslant 1$, draw $X_n \sim P(X_{n-1}, \cdot)$.
  - Estimate

$$\pi(f) \approx \frac{1}{T} \sum_{(T_0, T_0+T]} f(X_n)$$

Setting: Analysis

- Tasks:
  - Characterise quality of this estimator.
  - Make algorithmic recommendations.

Convergence of Markov Chains

- If P is ergodic, then for any $f \in \mathrm{L}^1(\pi)$, $\mathbb{E}[f(X_t)] \to \pi(f)$.
- How should we quantify this convergence?
  - Should be uniform over (some interesting class of) f
  - TV, KL, V-Norm, Transport, ...?
- Here: convergence in $\mathrm{L}^2(\pi)$.
  - Sufficiently strong for most applications.
  - Relevant to asymptotic variance, confidence intervals for estimates, etc.
- By duality, can translate convergence of expectations into convergence of laws.

$\mathrm{L}^2(\pi)$ Theory of Markov Chains

- Hilbert Spaces
  - $\mathrm{L}^2(\pi) := \{f : \mathcal{X} \to \mathrm{R} : \pi\left(f^2\right) < \infty\}$
  - $\mathrm{L}^2_0(\pi) := \mathrm{L}^2_0(\pi) \cap \{\pi(f) = 0\}$
  - Today: Work with $\mathrm{L}^2_0(\pi)$, (basically) WLOG.
- $\langle f, g \rangle := \pi(f \cdot g)$, $\|f\|_2^2 = \langle f, f \rangle = \pi\left(f^2\right)$.
- $(Pf)(x) := \int P(x, \mathrm{d}y) f(y)$.
  - Today: kernel P is π-reversible, hence operator P is symmetric.
  - Will also assume *positivity*; mostly in the background.

$\mathrm{L}^2(\pi)$ Theory of Markov Chains

- 'Dirichlet form': Let T satisfy $\pi T = \pi$.
  - Then, $\mathcal{E}_T(f, g) := \langle f, (I - T) g \rangle$.
  - Can also write

$$\mathcal{E}_T(f, g) = \frac{1}{2} \int \pi(\mathrm{d}x) \cdot T(x, \mathrm{d}y) \cdot (f(x) - f(y)) \cdot (g(x) - g(y)).$$

    - also 'Dirichlet energy', '(co)variance dissipation'
- Goal: for $f \in \mathrm{L}_0^2(\pi)$, obtain a bound like

$$\|P^n f\|_2^2 \leqslant \text{'rate'}(n) \cdot \text{'size'}(f).$$

Convergence Bounds in $\mathrm{L}^2(\pi)$

- Best case: 'rate' $(n) = (1-\gamma)^n$, 'size' $(f) = \|f\|_2^2$

$$\|P^n f\|_2^2 \leqslant (1-\gamma)^n \cdot \|f\|_2^2.$$

  - Equivalent to 'Poincaré' / 'Spectral Gap' inequality for P^*P:

$$\forall f \in \mathrm{L}_0^2(\pi), \quad \mathcal{E}_{P^*P}(f) \geqslant \gamma \cdot \|f\|_2^2$$

    - For positive P, basically the same as having one for P, up to constants.
  - Holds in many nice cases (well-confined target, good kernels),
    - $\cdots$ but not *all* cases (heavy tails, 'sticky' kernels).
- What can we say about *slower-than-exponential* convergence?

Slower-than-Exponential Convergence in $\mathrm{L}^2(\pi)$

- Let ψ denote a sequence with $\psi(n) \in \exp(-o(n)) \cap o(1)$.
- Might try 'rate' $= \psi$, 'size' $= \|\cdot\|_2^2$, i.e. $\|P^n f\|_2^2 \leqslant \psi(n) \cdot \|f\|_2^2$.
  - Actually, this implies exponential convergence! So, no use for bad chains.
- Might try 'rate' $= \psi$, 'size' $= \Phi$ with $\Phi \gg \|\cdot\|_2^2$, i.e. $\|P^n f\|_2^2 \leqslant \psi(n) \cdot \Phi(f)$.
  - Only ask for convergence for a 'nice' set of f.
  - This *can* hold!
  - Moreover, this is equivalent to a different functional inequality:

$$\text{for } s > 0, \quad \|f\|_2^2 \leqslant s \cdot \mathcal{E}_{P*P}(f) + \beta(s) \cdot \Phi(f)$$

    which is known as a *weak* Poincaré inequality (RW2001).
  - Typically $\Phi = \mathrm{osc}(\cdot)^2$.

Weak Poincaré Inequalities

- Weak Poincaré inequality (WPI):

$$\text{for } s > 0, \quad \|f\|_2^2 \leqslant s \cdot \mathcal{E}_{P^*P}(f) + \beta(s) \cdot \Phi(f)$$

- Optimising over s yields

$$\frac{\mathcal{E}_{P^*P}(f)}{\Phi(f)} \geqslant K^*\left(\frac{\|f\|_2^2}{\Phi(f)}\right).$$

  with K^* positive, increasing, convex, 'flat' at 0.
- By Grönwall, can show that $\|P^n f\|_2^2 \leqslant \gamma(n) \cdot \Phi(f)$, where γ solves

$$\frac{\mathrm{d}\gamma}{\mathrm{d}t} = -K^*(\gamma)$$

$$\gamma(0) = \|f\|_2^2 / \Phi(f).$$

Pause

- Recap
  - Markov chains are useful for computation.
  - Markov chain convergence bounds are useful for complexity theory.
  - $\mathrm{L}^2(\pi)$ convergence theory is clean and relevant.
  - Functional inequalities can characterise $\mathrm{L}^2(\pi)$ convergence rates.
- Next:
  - When would WPIs be necessary?
  - How can we prove WPIs?

Causes for Slower-than-Exponential Convergence

- Broadly, common scenarios are
  - π has heavy tails, relative to the size of the moves which P makes.
    - $\approx$ optimising a function which is very flat
  - P is an approximation of some 'nice' kernel $\hat{P}$, with approximation error which is unbounded in a suitable sense.
    - $\approx$ optimising a function using inexact gradients
- In this work, we consider the **second** scenario.
  - Motivated by 'exact approximation' methodologies for inference in intractable likelihood models.

Case Study: Inference in State Space Models

- Consider a state space model, specified by

$$x_t | x_{t-1} \sim M_\theta \left(x_{t-1}, \mathrm{d}x_t \right)$$
$$y_t | x_t \sim G_\theta \left(\mathrm{d}y_t | x_t \right).$$

- Observe $\{y_t\}_{0 \leqslant t \leqslant T}$, infer θ.
  - $\{x_t\}_{0 \leqslant t \leqslant T}$ is **not** observed.
- For many estimators, computation requires access to $p_\theta \left(y \right)$.
  - This is not directly available, and is given by a high-dimensional integral.
  - One can emulate the ‘ideal’ strategy by approximating this quantity.

Bayesian Inference in State Space Models

- An 'ideal' approach in the Bayesian setting could be to posit a prior p_0 on θ, and then sample from the posterior measure using the Random Walk Metropolis algorithm:
    1. At θ,
        1.1 Propose $\theta' \sim \mathcal{N}(\theta, \sigma^2 \cdot \mathrm{Id})$.
        1.2 Evaluate $r(\theta, \theta') = \frac{p_0(\theta') \cdot p_{\theta'}(y)}{p_0(\theta) \cdot p_\theta(y)}$.
        1.3 With probability $\min\{1, r(\theta, \theta')\}$, move to θ'; otherwise, remain at θ.
- This cannot be implemented as-is.
- Can we approximate this procedure?

Likelihood Estimation in State Space Models

- It is known (DM04) that by use of Sequential Monte Carlo methods / 'particle filters', one can obtain estimators of the (marginal) likelihood $p_\theta(y)$ which are positive and unbiased.
- For an observation sequence of length T, under suitable assumptions, one can control the relative variance of these estimators by using $N \propto T$ particles.
  - Thus, high-quality estimators of the marginal likelihood can be obtained at a computational cost of $\mathcal{O}(T^2)$.
- Particle Marginal Metropolis-Hastings (PMMH; ADH10) proposes to use this estimator to approximate the MH ratio.

Pseudo-Marginal MCMC

- PMMH is an instance of the *Pseudo-Marginal* approach to MCMC (AR09), in which intractable likelihood terms are replaced by positive, unbiased estimators.
- It can be shown that when implemented correctly, these methods indeed admit the **correct** invariant measure.
- Such methods are often termed ‘exact approximations’.

Some notation

- Need to formalise how the estimator $\hat{p}_\theta(y)$ is generated.
- In this context, it is useful to view unbiased estimators as mean-1 multiplicative perturbations of the truth, e.g.
    1. Draw $w \sim Q_\theta(\mathrm{d}w)$, and
    2. Observe $\hat{p}_\theta(y) = p_\theta(y) \cdot w$,

  where $\int Q_\theta(\mathrm{d}w) \cdot w = 1$.
- We **do not** observe w, but its behaviour is key to the analysis.
- w will be referred to as 'weights' in what follows.

Particle Marginal Metropolis-Hastings

1. At (θ, w),
   1.1 Propose $\theta' \sim \mathcal{N}\left(\theta, \sigma^2 \cdot \mathrm{Id}\right)$.
   1.2 Draw $w' \sim Q_{\theta'}\left(\mathrm{d}w'\right)$ (implicit via particle filter).
   1.3 Evaluate

$$r\left((\theta, w), \left(\theta', w'\right)\right) = \frac{p_0\left(\theta'\right) \cdot \hat{p}_{\theta'}(y)}{p_0(\theta) \cdot \hat{p}_\theta(y)} = \frac{p_0\left(\theta'\right) \cdot p_{\theta'}(y) \cdot w'}{p_0(\theta) \cdot p_\theta(y) \cdot w}.$$

   1.4 With probability $\min\left\{1, r\left((\theta, w), \left(\theta', w'\right)\right)\right\}$, move to $\left(\theta', w'\right)$; otherwise, remain at (θ, w).

Convergence of PMMH

- Some known results:
  - PMMH admits the invariant measure

    $$\Pi(\mathrm{d}\theta, \mathrm{d}w) \propto p_0(\mathrm{d}\theta) \cdot p_\theta(y) \cdot Q_\theta(\mathrm{d}w) \cdot w,$$

    which admits the posterior measure as a marginal.
  - (Spectral Gap for P) + ($\Pi - \operatorname{ess\,sup} w < \infty$) $\implies$ (Spectral Gap for $\tilde{P}$)
  - ($\Pi(\{\theta : Q_\theta - \operatorname{ess\,sup} w = \infty\}) > 0$) $\implies$ (No Spectral Gap for $\tilde{P}$)
- Intuitively, if the weights are unbounded, but *light-tailed*, then the convergence behaviour of PMMH should still be favourable, even if slower-than-exponential.
- We will see that WPIs can capture this intuition.

Comparison of Markov Chains

- Analysing a Markov chain from scratch is challenging.
- It is often easier to show that a similar, simpler Markov chain has a certain behaviour, and then analyse your original chain perturbatively.
- We will analyse the PMMH chain by comparing it to a suitable embedding of the ideal RWMH chain.
- This comparison takes place at the level of *Dirichlet forms*.

‘Embedded’ Random-Walk Metropolis-Hastings

1. At (θ, w),
   1.1 Propose $\theta' \sim \mathcal{N}\left(\theta, \sigma^2 \cdot \mathrm{Id}\right)$.
   1.2 Draw $w' \sim w' \cdot Q_{\theta'}\left(\mathrm{d}w'\right)$.
   1.3 Evaluate

$$r\left((\theta, w), \left(\theta', w'\right)\right) = \frac{p_0\left(\theta'\right) \cdot p_{\theta'}(y)}{p_0(\theta) \cdot p_\theta(y)} = r\left(\theta, \theta'\right).$$

   1.4 With probability $\min\left\{1, r\left(\theta, \theta'\right)\right\}$, move to $\left(\theta', w'\right)$; otherwise, remain at (θ, w).

Dirichlet Forms and PMMH

- Recall the definition of the Dirichlet form:

$$\mathcal{E}_T(f,g) = \frac{1}{2}\int \pi(\mathrm{d}x)\cdot T(x,\mathrm{d}y)\cdot(f(x)-f(y))\cdot(g(x)-g(y)).$$

- Note that if $T(x,\mathrm{d}y) = T(x,\{x\})\cdot\delta_x + T_+(x,\mathrm{d}y)$, then it holds that

$$\mathcal{E}_T(f,g) = \frac{1}{2}\int \pi(\mathrm{d}x)\cdot T_+(x,\mathrm{d}y)\cdot(f(x)-f(y))\cdot(g(x)-g(y)),$$

i.e. we can ignore the ‘lazy’ part of the chain.

Dirichlet Forms and PMMH

- The Dirichlet form of the embedded chain writes as

$$\mathcal{E}_P(f) = \frac{1}{2} \cdot \int \Pi(\mathrm{d}\theta, \mathrm{d}w) \cdot \mathcal{N}\left(\theta'; \theta, \sigma^2 \cdot \mathrm{Id}\right) \cdot w' \cdot Q_{\theta'}\left(\mathrm{d}w'\right)$$
$$\cdot \min\left\{1, r\left(\theta, \theta'\right)\right\} \cdot \left(f(\theta, w) - f\left(\theta', w'\right)\right)^2 .$$

- The Dirichlet form of the PMMH chain writes as

$$\mathcal{E}_{\tilde{P}}(f) = \frac{1}{2} \cdot \int \Pi(\mathrm{d}\theta, \mathrm{d}w) \cdot \mathcal{N}\left(\theta'; \theta, \sigma^2 \cdot \mathrm{Id}\right) \cdot Q_{\theta'}\left(\mathrm{d}w'\right)$$
$$\cdot \min\left\{1, r\left((\theta, w), \left(\theta', w'\right)\right)\right\} \cdot \left(f(\theta, w) - f\left(\theta', w'\right)\right)^2 .$$

Comparing the Acceptance Probabilities

- Our goal is to write $\mathcal{E}_P(f) \lesssim \mathcal{E}_{\tilde{P}}(f) + \text{`slack'}$
- Compare the integrands:

$$\frac{\text{Integrand}_P}{\text{Integrand}_{\tilde{P}}} = \frac{w' \cdot \min\left\{1, r\left(\theta, \theta'\right)\right\}}{\min\{1, r((\theta, w), (\theta', w'))\}}$$

- Using $\min\{1, a \cdot b\} \geqslant \min\{1, a\} \cdot \min\{1, b\}$, we obtain that

$$\frac{\text{Integrand}_P}{\text{Integrand}_{\tilde{P}}} \leqslant \max\left\{w, w'\right\}$$

Comparing the Acceptance Probabilities

- It thus holds that

$$\mathcal{E}_P(f) \leqslant \frac{1}{2} \cdot \int \Pi(\mathrm{d}\theta, \mathrm{d}w) \cdot \mathcal{N}\left(\theta'; \theta, \sigma^2 \cdot \mathrm{Id}\right) \cdot Q_{\theta'}\left(\mathrm{d}w'\right) \cdot \max\left\{w, w'\right\}$$
$$\cdot \min\left\{1, r\left((\theta, w), \left(\theta', w'\right)\right)\right\} \cdot \left(f(\theta, w) - f\left(\theta', w'\right)\right)^2.$$

- Compare again

$$\mathcal{E}_{\tilde{P}}(f) = \frac{1}{2} \cdot \int \Pi(\mathrm{d}\theta, \mathrm{d}w) \cdot \mathcal{N}\left(\theta'; \theta, \sigma^2 \cdot \mathrm{Id}\right) \cdot Q_{\theta'}\left(\mathrm{d}w'\right)$$
$$\cdot \min\left\{1, r\left((\theta, w), \left(\theta', w'\right)\right)\right\} \cdot \left(f(\theta, w) - f\left(\theta', w'\right)\right)^2.$$

Truncation and Comparison

- Only remaining difference: $\max$ term
- We partition the space into a ‘good set’ on which this factor is well-controlled, and its complement

$$\text{for } s > 0, \quad A(s) := \left\{ \left(\theta, w, \theta', w' \right) : \max \left\{ w, w' \right\} \leqslant s \right\}.$$

- On $A(s)$, we use $\max \leqslant s$.
- On $A(s)^{\complement}$, we use $\left(f(\theta, w) - f\left(\theta', w' \right) \right)^2 \leqslant \operatorname{osc}^2(f)$.
- One subsequently sees that

$$\mathcal{E}_P(f) \leqslant s \cdot \mathcal{E}_{\tilde{P}}(f) + \beta(s) \cdot \Phi(f)$$

with $\beta(s) = \Pi(w > s)$, $\Phi = \operatorname{osc}^2$.

Deducing and Applying the WPI

- Under the assumption that P admits a spectral gap $\gamma_P > 0$, we combine this with our 'relative' WPI to see that

$$\|f\|_2^2 \leqslant s \cdot \mathcal{E}_{\tilde{P}}(f) + \beta(\gamma_P \cdot s) \cdot \Phi(f),$$

  i.e. a legitimate WPI for $\tilde{P}$.
- One can immediately read off several conclusions, e.g.
  - If $\Pi(w^p) < \infty$, then $\beta \in \mathcal{O}(s^{-p})$, $\|P^n f\|_2^2 \in \mathcal{O}(n^{-p})$.
  - If $\log w \sim \mathcal{N}$, then for all $p > 0$, it holds that $\beta \in o(s^{-p})$, $\|P^n f\|_2^2 \in o(n^{-p})$
  - Reducing only the variance of w will improve constants, but not rates.
  - Improving the integrability of w will improve rates.

Further comments on the techniques

- Paper contains additional applications, e.g. $\#$ particles in PMMH, MCMC-ABC (Approximate Bayesian Computation), asymptotic variance, ...
- Proofs have a common, simple structure.
  - Compare integrands of Dirichlet forms (multiplicatively).
  - Identify family of 'good' sets.
  - Partition space appropriately.
- Results are robust to inner workings of algorithms
  - e.g. for PMMH, estimator just needs to be positive and unbiased.
  - Didn't need to come from a particle filter!
  - Any valid estimator with good tails gives a good bound.

Recap

- $L^2(\pi)$ convergence analysis of MCMC via Functional Inequalities
- Comparison theory is well-suited to many popular algorithms, particularly for intractable likelihood models.
- WPIs are a usable theoretical tool.
  - When applicable, proofs are usually clean and interpretable.
  - Not suitable for **all** perturbation-type chains, e.g. SG-MCMC.
- Subsequent and ongoing work:
  - General foundations of WPIs for Markov chains
  - Connections to e.g. conductance, isoperimetry.
  - Establishing (strong) PIs for ideal chains.
  - Applications to other ‘exact approximate’ methods.