Nonlinear MCMC for Bayesian Machine Learning

James Vuckovic
james@jamesvuckovic.com

November 23, 2022

Outline:

1. Introduction
2. Theory
3. Experiments

Motivation — MCMC and Bayesian ML

- We want to compute the **Bayesian posterior predictive distribution** $P(y|x, \mathcal{D})$.
- Given a parameteric family of models $P(y, x|\theta)$, we can decompose this problem into the **integral**

$$P(y|x, \mathcal{D}) = \int P(y|x, \theta)P(\theta|\mathcal{D})$$

- However, this integral is generally **intractable** so we approximate it with samples $\theta^i \sim P(\theta|\mathcal{D})$ i.e.

$$P(y|x, \mathcal{D}) = \int P(y|x, \theta)P(\theta|\mathcal{D}) \approx \frac{1}{N}\sum_{i=1}^{N} P(y|x, \theta^i)$$

- However, generating exact samples θ^i is **also intractable** so we can use a **Markov kernel** $\mathcal{T}$ with **invariant measure** $P(\theta|\mathcal{D})$ to generate independent Markov chains $\theta^i_{n+1} \sim \mathcal{T}(\theta^i_n, \bullet)$ and approximate

$$P(y|x, \mathcal{D}) = \int P(y|x, \theta)P(\theta|\mathcal{D}) \approx \frac{1}{N}\sum_{i=1}^{N} P(y|x, \theta^i) \approx \frac{1}{N}\sum_{i=1}^{N} P(y|x, \theta^i_{n_{\mathrm{sim}}})$$

- This procedure is **Markov Chain Monte Carlo** (MCMC). The central question of our work: how do we pick a good transition kernel?

Nonlinear MCMC — Mean Field System

- To specify a MCMC algorithm, we need a Markov **transition kernel**
- Our work directly builds on the paper Andrieu et al., 2011 with some important differences & extensions

> **On nonlinear Markov chain Monte Carlo**
>
> CHRISTOPHE ANDRIEU[1], AJAY JASRA[2], ARNAUD DOUCET[3] and PIERRE DEL MORAL[4]

- Consider a family of kernels indexed by probability measures $\eta \in \mathcal{P}(\mathbb{R}^d)$

$$K_\eta(x, \mathrm{d}y) := (1 - \varepsilon)K(x, \mathrm{d}y) + \varepsilon J_\eta(x, \mathrm{d}y)$$

  - K is a *linear* Markov kernel, called the **primary** kernel, and
  - J_η is a family of *nonlinear* "jump **interaction**" Markov kernels
  - $\varepsilon \in]0, 1[$ is a mixture hyperparameter
- Let Q be a *linear* Markov kernel called the **auxiliary** kernel
- We can construct a nonlinear Markov chain $\{(X_n, Y_n)\}_{n=0}^{\infty}$ from K_η as follows

$$\begin{cases} Y_{n+1} \sim Q(Y_n, \bullet) \\ \eta_{n+1} := \mathrm{Distribution}(Y_{n+1}) \qquad Y_0 \sim \eta_0, \;\; X_0 \sim \mu_0 \\ X_{n+1} \sim K_{\eta_{n+1}}(X_n, \bullet) \end{cases} \tag{1}$$

- We pick Q to be $\eta^\star$-invariant, $K, J_{\eta^\star}$ to be π-invariant

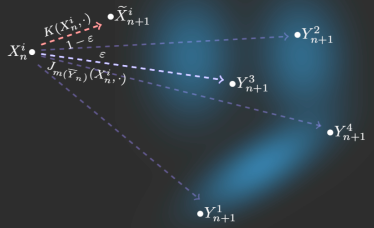
Nonlinear MCMC — Interacting Particle System

- However equation (1) can't be directly simulated due to $\mathrm{Distribution}(Y_n)$
- We can *approximate* $\mathrm{Distribution}(Y_n)$ in the mean field system (1) using a set of particles $\overline{Y}_n := \{Y_n^1, \ldots, Y_n^N\}$ with the empirical measure

$$\mathrm{Distribution}(Y_n) \approx m(\overline{Y}_n) := \frac{1}{N} \sum_{i=1}^{N} \delta_{Y_n^i}$$

- Hence we get the **interacting particle system** which we will *simulate* to obtain MCMC estimates

$$\begin{cases} Y_{n+1}^i \sim Q(Y_n^i, \bullet) \\ \eta_{n+1}^N := m(\overline{Y}_{n+1}) \\ X_{n+1}^i \sim K_{\eta_{n+1}^N}(X_n^i, \bullet) \end{cases} \qquad Y_0^i \overset{iid}{\sim} \eta_0, \; X_0^i \overset{iid}{\sim} \mu_0, \; i = 1, \ldots, N. \tag{2}$$



Specific Nonlinear Interactions

- We need to make choices for nonlinear "jump" interactions J_η. We use two proposals in Andrieu et al., 2011
- Define the potential function $G(x) := \frac{\pi(x)}{\eta^\star(x)}$
- **Boltzmann-Gibbs transformation**

$$J_\eta^{BG}(x, \mathrm{d}y) = \Psi_G(\eta)(\mathrm{d}y); \quad \Psi_G(\eta)(\mathrm{d}y) = \frac{G(y)}{\eta(G)}\eta(\mathrm{d}y)$$

  - This uses G to re-weight the distribution η
  - If $\eta = m(\{Y^1, \ldots, Y^N\})$, then this amounts to using a *softmax* over the log-potentials of each particle:

$$\Psi_G(m(\overline{Y})) = \sum_{i=1}^{N} \frac{G(Y^i)}{\sum_{j=1}^{N} G(Y^j)} \delta_{Y^i}$$

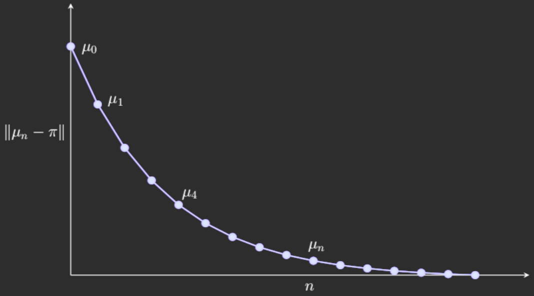
- **Accept-Reject Interaction**

$$J_\eta^{AR}(x, \mathrm{d}y) = \underbrace{\alpha(x, y)\eta(\mathrm{d}y)}_{\text{accept}} + \underbrace{\left(1 - \int \alpha(x, y)\eta(\mathrm{d}y)\right)\delta_x(\mathrm{d}y)}_{\text{reject}}; \quad \alpha(x, y) := 1 \wedge \frac{G(y)}{G(x)}$$

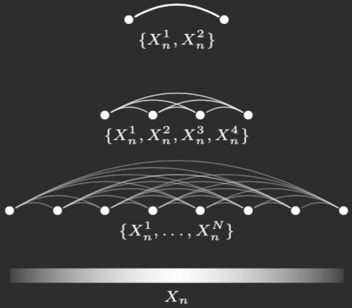
  - This is an "adaptive Metropolis-Hastings" where the proposal distribution is η
- π is $J_{\eta^\star}$-invariant for each of these choices

Convergence Notions For Nonlinear MCMC

- We denote:
  - $\mu_n := \mathrm{Distribution}(X_n)$, i.e. the distribution of the mean-field system (1)
  - $\mu_n^N := \mathrm{Distribution}(X_n^1)$, i.e. the distribution of one particle from the interacting particle system (2) (it doesn't matter which particle)
- Two different phenomena to characterize convergence of nonlinear MCMC
  1. **Long-time convergence**: Does $\|\mu_n - \pi\| \to 0$ and, if so, at what rate?
  2. **Propagation of Chaos**: Do groups of *dependent* particles become *independent* as $N \to \infty$?



Long-time Convergence



Propagation of Chaos

Main Theorem

Theorem 1 (Convergence of Nonlinear MCMC)

Under suitable conditions on K_η and Q, there exist fixed constants $C_1, C_2, C_3 > 0$, a function $\mathcal{R} : [0, \infty[\to [1, \infty[$, and $\rho > 0$ s.t.

$$\|\mu_n^N - \pi\|_{tv} \leq C_1 \frac{1}{N} \mathcal{R}(1/N) + C_2 \rho^n + C_3 n \rho^n.$$

- Theorem 1 says it's not *sufficient* to only let $n \to \infty$ to ensure $\mu_n \to \pi$
- It's easy to come up with cases where for any $N > 0$, π is not $K_{m(\overline{Y}_n)}$-invariant for *any* n. Hence having $N \to \infty$ is also *necessary* (at least in general)

Corollary 2 (Adapted from Sznitman, 1991, Theorem 2.2)

Suppose that Theorem 1 applies to K_η. Let $\overline{X}_n := \{X_n^1, \ldots, X_n^N\}$ be the interacting particle system from (2). Then for every $n \in \mathbb{N}$ and $f \in \mathcal{B}_b(\mathbb{R}^d)$ we have

$$\lim_{N \to \infty} \mathbb{E}\left[\left|\frac{1}{N} \sum_{i=1}^{N} f(X_n^i) - \mu_n(f)\right|\right] = 0.$$

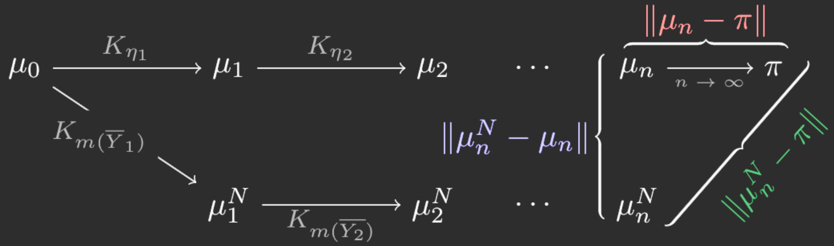
Proof Sketch — Triangle Inequality

Main ingredients of proof:

1. Ergodicity of K, Q
2. "Regularity" of $\eta \mapsto J_\eta$
3. Use the triangle inequality as follows:

$$\|\mu_n^N - \pi\| \leq \underbrace{\|\mu_n^N - \mu_n\|}_{\text{propagation of chaos}} + \underbrace{\|\mu_n - \pi\|}_{\text{long-time convergence}}$$

$$\lesssim \frac{1}{N}\mathcal{R}\left(\frac{1}{N}\right) + \rho^n + n\rho^n$$



Proof Sketch — Lipschitz Regularity of J

- We use Lipschitz regularity of J to ensure that as $\eta_n \to \eta^\star$, $J_{\eta_n} \to J_{\eta^\star}$

$$\|J_\eta - J_{\eta'}\| \lesssim \|\eta - \eta'\| \quad \eta, \eta' \in \mathcal{P}(\mathbb{R}^d)$$

- This ensures that $K_{\eta_n} \to K_{\eta^\star}$
- Note: the LHS norm and the RHS norms are *different* – one is over kernels, the other is over probability measures

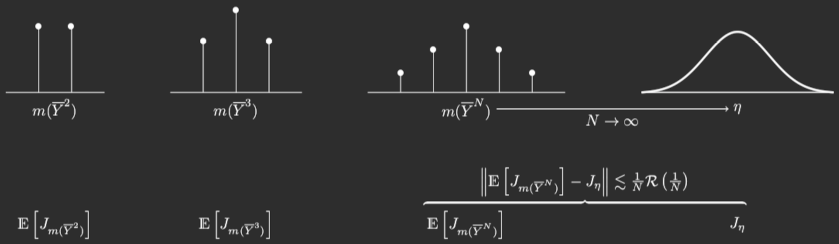
$$\eta_0 \xrightarrow{Q} \eta_1 \xrightarrow{Q} \eta_2 \quad \cdots \quad \underbrace{\eta_n \quad \cdots \quad \eta^\star}_{\|\eta_n - \eta^\star\| \lesssim \delta^n \xrightarrow{n\to\infty} 0}$$

$$J_{\eta_1} \qquad J_{\eta_2} \quad \cdots \quad \overbrace{J_{\eta_n} \quad \cdots \quad J_{\eta^\star}}^{\|J_{\eta_n} - J_{\eta^\star}\| \lesssim \|\eta_n - \eta^\star\| \xrightarrow[n\to\infty]{} 0}$$

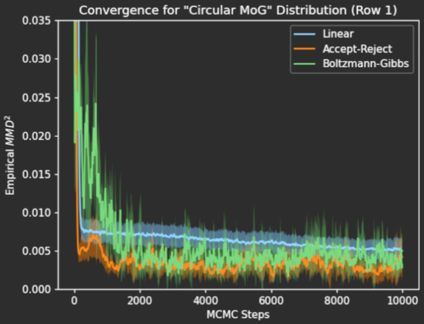
Proof Sketch — LLN Regularity of J

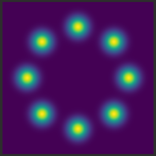
- We use particles $\overline{Y} := (Y^1, \ldots, Y^N)$ to empirically approximate a measure $\eta \in \mathcal{P}(\mathbb{R}^d)$ by $\eta \approx m(\overline{Y}) = \frac{1}{N}\sum_{i=1}^N \delta_{Y^i}$
- As $N \to \infty$, various LLN (or CLT) results tell us $m(\overline{Y}) \to \eta$
- We want this same "LLN" regularity of the nonlinear interaction J_η, i.e. as $N \to \infty$, we want $J_{m(\overline{Y})} \to J_\eta$
- Additional considerations:
  - The convergence should not depend on which $\eta \in \mathcal{P}(\mathbb{R}^d)$ we're approximating
  - Since $\overline{Y}$ is random $J_m(\overline{Y})$ is a *random kernel* which is hard to work with. To simplify, we mean convergence in a suitable *expectation* sense:

$$\mathbb{E}[J_{m(\overline{Y})} f(x)] \to J_\eta f(x) \quad \forall x \in \mathbb{R}^d,\ f \text{ in a suitable set of functions}$$

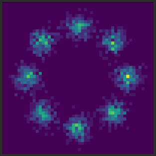


Experiments — 2-dimensional (Circular Mog[1])

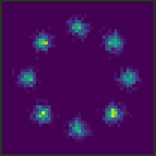




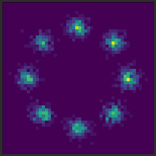
Target Density



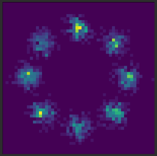
Ground Truth



Linear (MALA)



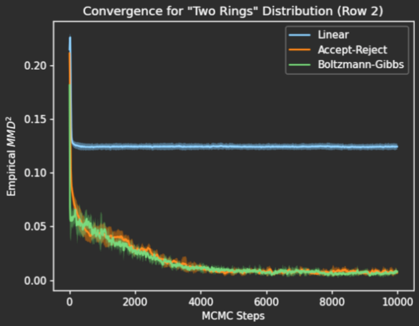
Nonlinear (AR)

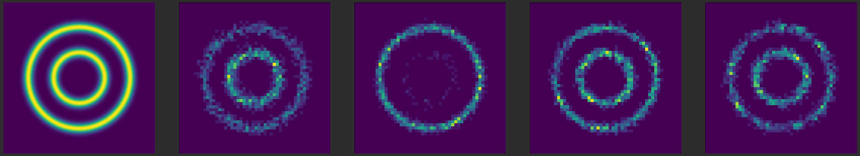


Nonlinear (BG)

[1] Density from Stimper, Schölkopf, and Hernández-Lobato, 2022

Experiments — 2-dimensional (Two Rings[2])

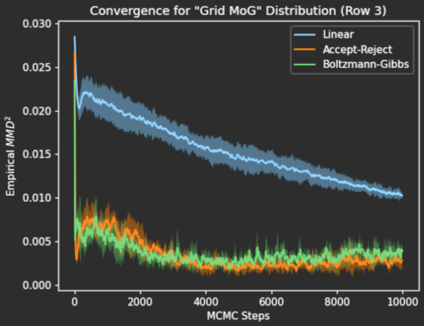


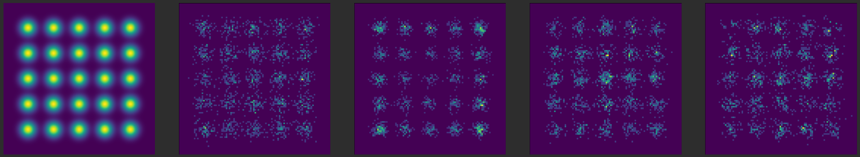


Target Density | Ground Truth | Linear (MALA) | Nonlinear (AR) | Nonlinear (BG)

[2] Density from Stimper, Schölkopf, and Hernández-Lobato, 2022

Experiments — 2-dimensional (Grid Mog[3])

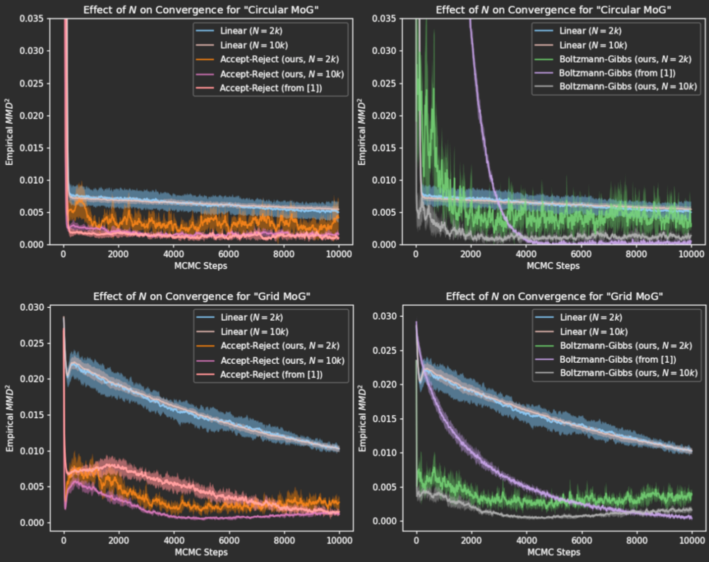




Target Density	Ground Truth	Linear (MALA)	Nonlinear (AR)	Nonlinear (BG)

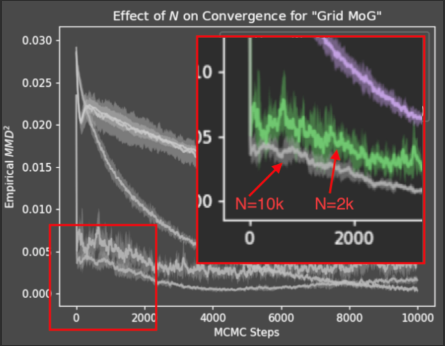
[3] Density from Zhang et al., 2019

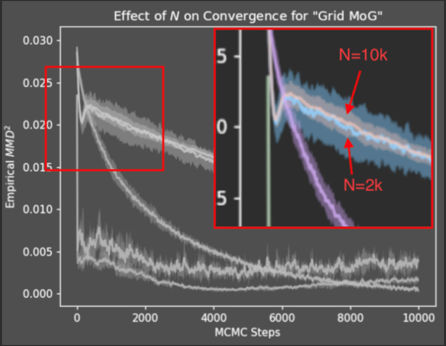
Experiments — 2-dimensional



Accept-Reject — Boltzmann-Gibbs

Experiments — 2-dimensional

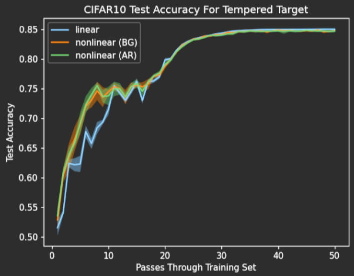




- As predicted by Theorem 1, the nonlinear MCMC convergence rate **depends substantially** on the choice of N (left)
- This is **not true** for the linear MCMC convergence rate, which only reduces variance by increasing N (right)

Experiments — ResNet18 on CIFAR10

- Setup:
  - Likelihood is $P(y|x,\theta)$ is ResNet-18 (He et al., 2016)
  - Prior is $P(\theta)$ is $\mathcal{N}(0, 10^{-4}I)$
  - Auxiliary target density is $\eta^\star(\theta) \propto P(\theta|\mathcal{D})^{1/\tau_1}$
  - Target density is $\pi(\theta) \propto P(\theta|\mathcal{D})^{1/\tau_2}$
  - Auxiliary kernel Q is RMS-ULA,
  - Primary kernel K is ULA,
  - Minibatch size is 256



Algorithm	Test Accuracy (↑) Tempered	Expected Calibration Error (↓) Tempered
Linear	$85.01_{\pm 0.19}$	$0.26_{\pm 0.014}$
Nonlinear (BG)	$84.74_{\pm 0.08}$	$0.16_{\pm 0.03}$
Nonlinear (AR)	$84.67_{\pm 0.23}$	$0.15_{\pm 0.05}$

Conclusion

- What did we do?
  1. We analyzed the convergence of a variation on the family of nonlinear Markov chain Monte Carlo methods proposed in Andrieu et al., 2011
  2. Our proof decomposes into two separate results on long-time and large-particle convergence
  3. We also applied our theory to two specific choices of samplers also introduced in Andrieu et al., 2011
  4. We did some experiments 2-dimensional experiments that demonstrate *superior performance* provided one can choose $\eta^\star$ properly
  5. We did some large-scale experiments on CIFAR10 that show our methods are *feasible* and *comparable but not better* than the linear methods on CIFAR10
- What's next?
  1. Investigate how to choose $\eta^\star$ better in high dimension (e.g. for neural networks)
  2. Expand to more high-dimensional settings and develop better recipes for MCMC in Bayesian ML

Thank you!

References

- Andrieu, Christophe et al. (2011). "On nonlinear Markov chain Monte Carlo". In: *Bernoulli* 17.3, pp. 987 –1014. DOI: 10.3150/10-BEJ307. URL: https://doi.org/10.3150/10-BEJ307.
- He, Kaiming et al. (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.
- Stimper, Vincent, Bernhard Schölkopf, and José Miguel Hernández-Lobato (2022). "Resampling Base Distributions of Normalizing Flows". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 4915–4936.
- Sznitman, Alain-Sol (1991). "Topics in propagation of chaos". In: *Ecole d'été de probabilités de Saint-Flour XIX—1989*. Springer, pp. 165–251.
- Zhang, Ruqi et al. (2019). "Cyclical Stochastic Gradient MCMC for Bayesian Deep Learning". In: *International Conference on Learning Representations*.