Markov Chain Monte Carlo and Couplings Part II

Pierre E. Jacob

Optimization-Conscious Econometrics Summer School

June 3-6th, 2024

Outline

Outline

Outline

Setting

Target probability distribution π.

MCMC: $X_0 \sim \pi_0$, then $X_t | X_{t-1} \sim P(X_{t-1}, \cdot)$ for $t = 1, 2, \ldots$

Expectation: $\pi(h) = \int h(x)\pi(dx) = \mathbb{E}[h(X)]$.

Marginal after one step:
$\pi_1(dx') = \pi_0 P(dx') = \int \pi_0(dx) P(x, dx')$.

Conditional expectation:
$Ph(x) = \int P(x, dx')h(x') = \mathbb{E}[h(X_1)|X_0 = x]$.

Example: probit regression

Probit regression: $\mathbb{P}(Y_i = 1|X_i) = \Phi(X_i'\beta)$, where Φ is Normal$(0, 1)$ cumulative distribution function (CDF).

Equivalent to:

$$Y_i = \mathbb{1}(Z_i > 0), \quad Z_i = X_i'\beta + \epsilon_i, \quad \epsilon_i \sim \text{Normal}(0, 1).$$

Prior: $\beta \sim \text{Normal}(\mu, \Sigma)$.

Conditional distributions:

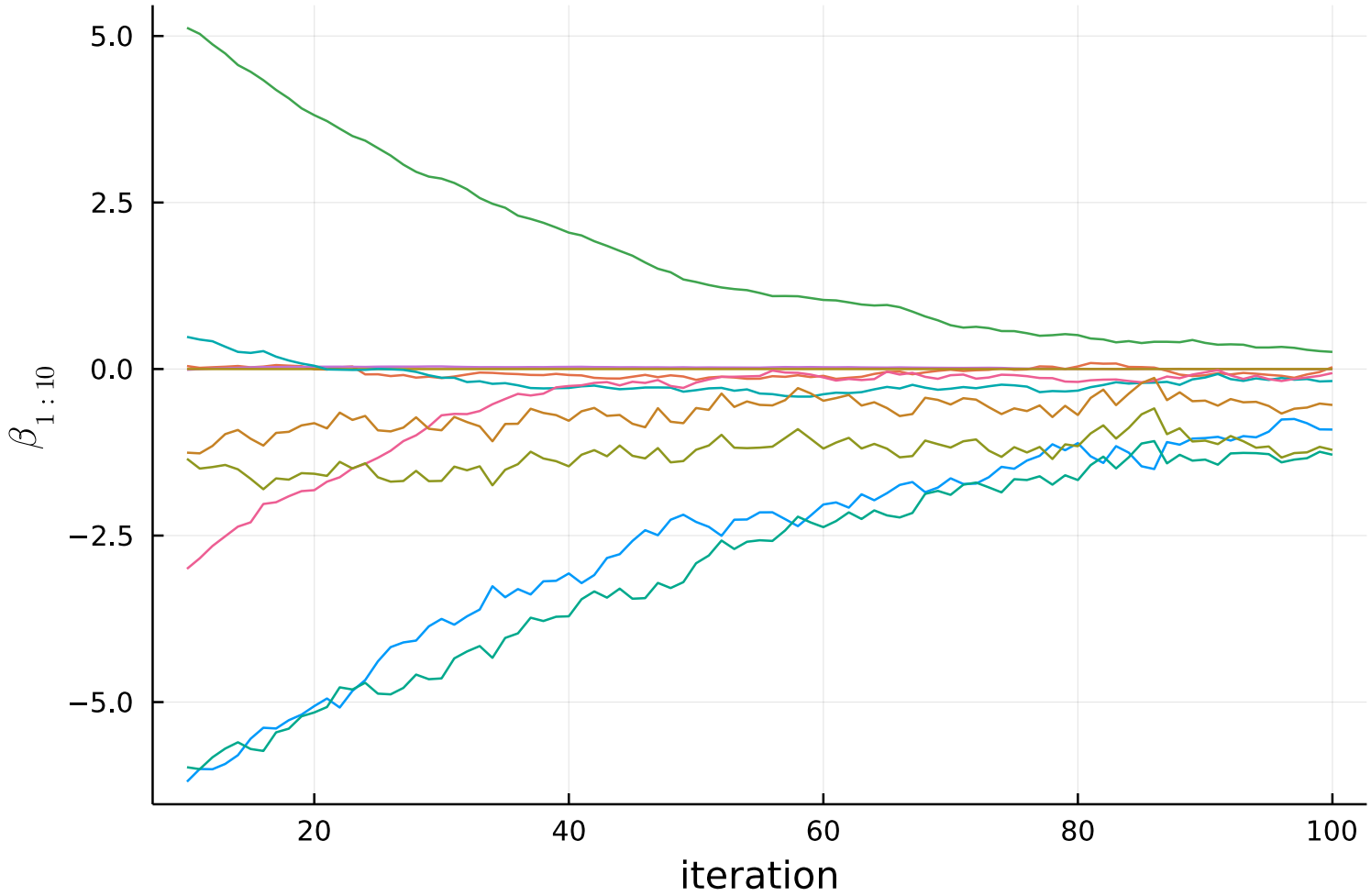
$\forall i$ $Z_i|\beta \sim \text{Normal}(X_i'\beta, 1)$ truncated at 0 on left/right if Y_i is 1/0,

$\beta|Z_{1:n} \sim \text{Normal}((X'X + \Sigma^{-1})^{-1}(X'Z + \Sigma^{-1}\mu), (X'X + \Sigma^{-1})^{-1})$.

The MCMC algorithm updates $(\beta, Z_{1:n})$ alternately.

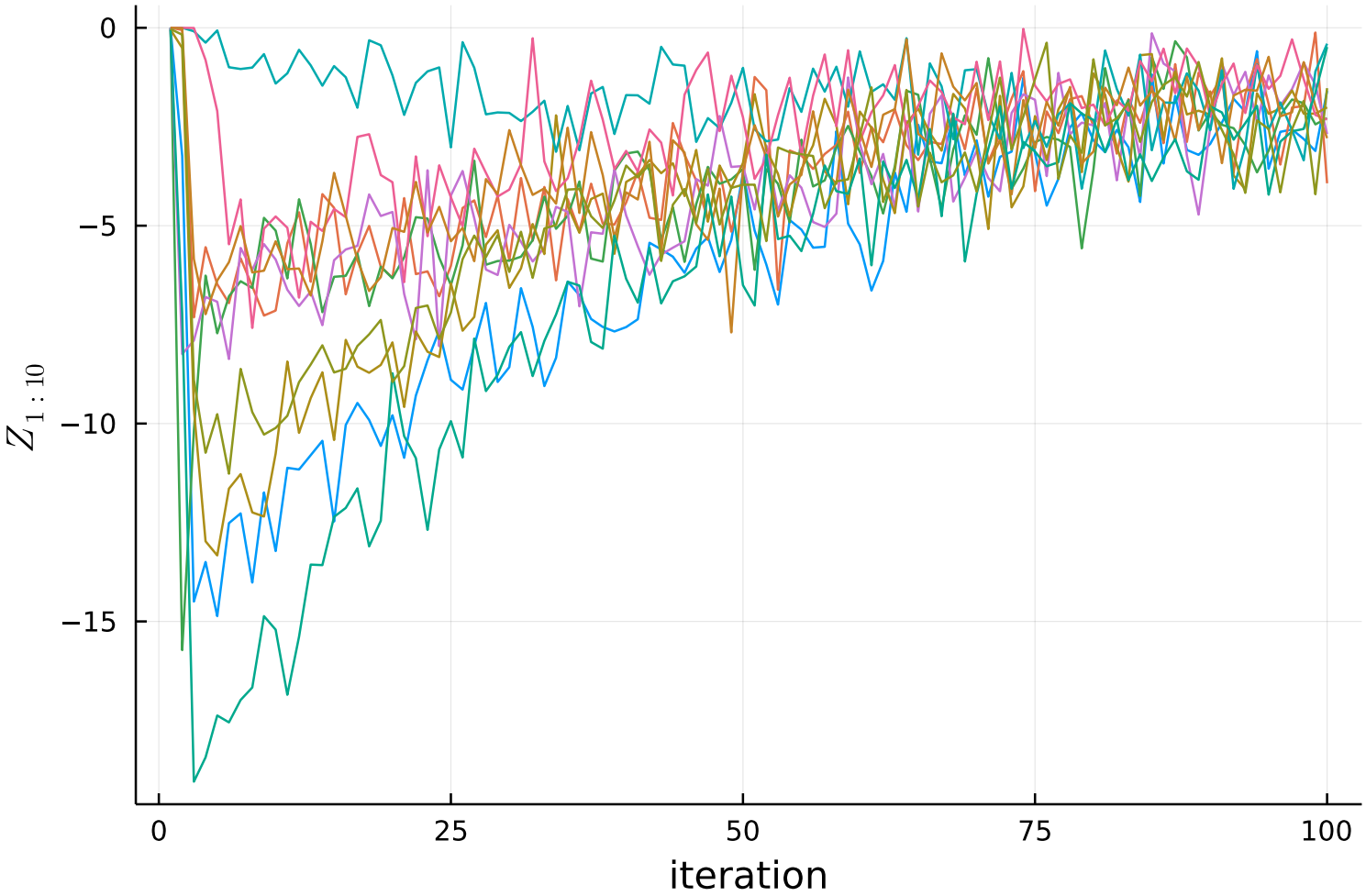
Can start with β from prior, or from MLE, or from $(0, \ldots, 0)$.

Example: probit regression



How long should we run this?

Example: probit regression



How long should we run this?

Some questions

Convergence of marginals:

$$|\pi_t - \pi| \to 0,$$

for some choice of distance, at which rate?

Law of large numbers:

$$t^{-1} \sum_{s=0}^{t-1} h(X_s) \xrightarrow[t \to \infty]{a.s.} \pi(h).$$

Central limit theorem:

$$\sqrt{t} \left(t^{-1} \sum_{s=0}^{t-1} h(X_s) - \pi(h) \right) \xrightarrow[t \to \infty]{d} \text{Normal}(0, v(P, h)).$$

Basic convergence properties of MCMC

In MCMC the chain (X_t) is π-invariant by construction.

It is not enough for convergence, e.g. consider $P(x, \cdot) = \delta_x(\cdot)$.

Irreducibility, aperiodicity and π-invariance imply convergence.
Theorem 4 in Roberts & Rosenthal, 2004, *General state space Markov chains and MCMC algorithms*.

LLN: for $\pi(|h|) < \infty$ we have $t^{-1} \sum_{s=0}^{t-1} h(X_s) \to \pi(h)$ a.s.
Theorem 17.0.1 in Meyn & Tweedie, 1993, *Markov chains and stochastic stability*.

Outline

Central limit theorem

Markov kernel P, test function h, might satisfy

$$\sqrt{t}\left(t^{-1}\sum_{s=0}^{t-1} h(X_s) - \pi(h)\right) \to \text{Normal}(0, v(P,h)),$$

where $v(P,h)$ is called the asymptotic variance.

Taking the limit $t \to \infty$ of $\mathbb{V}^\star[t^{-1}\sum_{s=0}^{t-1} h(X_s)]$:

$$v(P,h) = \mathbb{V}^\star(h(X_0)) + 2\sum_{t=1}^{\infty} \mathbb{C}\text{ov}^\star(h(X_0), h(X_t)).$$

It is difficult to approximate $v(P,h)$.

The Poisson equation

Write $Ph(x) = \int P(x, dx')h(x') = \mathbb{E}[h(X_1)|X_0 = x]$.

A function g in $L^1(\pi)$ is said to be a solution of the Poisson equation associated with h and P, if

$$g - Pg = h - \pi(h).$$

For brevity we say that g is *fishy*. Solutions are not unique.

If $\sum_{t \geq 0} \|P^t\{h - \pi(h)\}\|_{L^1(\pi)} < \infty$ then fishy functions exist.

Marie Duflo, 1970, *Opérateurs potentiels des chaînes et des processus de Markov irréductibles.*

Examples of fishy function

The function

$$x \mapsto \sum_{t=0}^{\infty} P^t \left\{ h - \pi(h) \right\}(x),$$

satisfy

$$g - Pg = h - \pi(h),$$

so it is fishy.

Also, if you add any constant c to g, it remains fishy.

We could denote it by 🐟 but we won't.

Central limit theorem

Aiming for a CLT for Markov chain ergodic averages, write

$$\sum_{s=0}^{t-1}\{h(X_s) - \pi(h)\} = \sum_{s=1}^{t}\{g(X_s) - Pg(X_{s-1})\} + g(X_0) - g(X_t).$$

Then apply the central limit theorem for martingale difference sequences, leading to the asymptotic variance

$$v(P, h) = \mathbb{E}^\star[\{g(X_1) - Pg(X_0)\}^2].$$

Chapter 21 in
Douc, Moulines, Priouret & Soulier, 2018, *Markov chains*.

Asymptotic variance

The more familiar form of the asymptotic variance is

$$\lim_{t\to\infty} \mathbb{V}^\star\left(t^{-1/2}\sum_{s=0}^{t-1} h(X_s)\right) = \mathbb{V}^\star(h(X)) + 2\sum_{s=1}^{\infty} \mathbb{C}\mathrm{ov}^\star\left(h(X_0), h(X_s)\right).$$

This expression is equivalent to $\mathbb{E}^\star[\{g(X_1) - Pg(X_0)\}^2]$.

Use $g = \sum_{t=0}^{\infty} P^t\{h - \pi(h)\}$, and $g = h - \pi(h) + Pg$,

$$\mathbb{E}^\star[\{g(X_1) - Pg(X_0)\}^2] = \pi(\{h - \pi(h)\}^2) + 2\pi(\{h - \pi(h)\} \cdot Pg).$$

Chapter 21 in
Douc, Moulines, Priouret & Soulier, 2018, *Markov chains*.

Asymptotic bias

Consider the bias of the average $t^{-1} \sum_{s=0}^{t-1} h(X_s)$.

Its expectation is $t^{-1} \sum_{s=0}^{t-1} P^s h(x_0)$, given $X_0 = x_0 \in \mathbb{X}$.

Therefore

$$\lim_{t \to \infty} t \times \left\{ \mathbb{E}_{x_0}[t^{-1} \sum_{s=0}^{t-1} h(X_s)] - \pi(h) \right\} = g(x_0),$$

where g is the fishy function $\sum_{t=0}^{\infty} P^t \{h - \pi(h)\}$.

Kontoyiannis & Dellaportas, 2009, *Notes on using control variates for estimation with reversible MCMC samplers.*

Outline

Convergence of marginals

Denote by π_t the distribution of X_t at time t.

Rate of marginal convergence: smallest function $\varphi : \mathbb{N} \to \mathbb{R}_+$ such that:

$$|\pi_t - \pi| \leq \varphi(t).$$

Typically we obtain upper bounds on the rate of convergence.

Geometric ergodicity: upper bound $C\rho^t$ for some $\rho \in (0, 1)$.

Polynomial ergodicity: upper bound $Ct^{-\kappa}$ for some $\kappa > 0$.

Mixing time: for some $\varepsilon > 0$,

$$t_\varepsilon = \inf\{t \in \mathbb{N} : \ |\pi_t - \pi| \leq \varepsilon\}.$$

Coupling of probability distributions

Let p and q be two probability distributions on $\mathbb{X}$.

A joint distribution Γ on $\mathbb{X} \times \mathbb{X}$ is a coupling of p and q if its first marginal is p and its second marginal is q.

$$\int \Gamma(x, dy) = p(x), \quad \int \Gamma(dx, y) = q(y).$$

Total variation distance

For two variables $X \sim p$ and $Y \sim q$,

$$\begin{aligned}
\|p - q\|_{\text{TV}} &= \sup_{A \in \mathcal{X}} \mathbb{P}(X \in A) - \mathbb{P}(Y \in A) \\
&= \frac{1}{2} \int |p(x) - q(x)| \, dx \\
&= 1 - \int \min(p(x), q(x)) dx \\
&= \inf_{(X,Y) \in \text{coupling}(p,q)} \mathbb{P}(X \neq Y)
\end{aligned}$$

Optimal transport distance

Let d be a distance on the state space.

Monge–Kantorovich or Wasserstein distance between p and q:

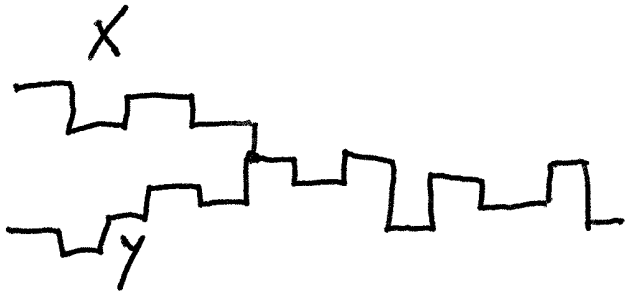
$$
\begin{aligned}
W_k(p, q) &= \left(\inf_{(X,Y) \in \text{coupling}(p,q)} \mathbb{E}[d(X, Y)^k] \right)^{1/k} \\
&= \left(\int_0^1 |F_p^{-1}(t) - F_q^{-1}(t)|^k dt \right)^{1/k}, \\
W_1(p, q) &= \sup_{h \in 1-\text{Lipschitz}} |p(h) - q(h)| \\
&= \int_{\mathbb{R}} |F_p(t) - F_q(t)| dt.
\end{aligned}
$$

Outline

Couplings of Markov chains

Construct a joint process (X_t, Y_t) such that $Y_t \sim \pi$ for all $t \geq 0$, and marginally both chains evolve according to same kernel P.

Suppose that there exists τ a random variable such that $X_t = Y_t$ for all $t \geq \tau$.



Then

$$\begin{aligned} \|\pi_t - \pi\|_{\mathrm{TV}} &= \|\mathcal{L}(X_t) - \mathcal{L}(Y_t)\|_{\mathrm{TV}} \\ &\leq \mathbb{P}(X_t \neq Y_t) = \mathbb{P}(\tau > t), \end{aligned}$$

where $\|\cdot\|_{\mathrm{TV}}$ is the total variation distance.

Bru & Yor, 2002, *Comments on the life and mathematical legacy of Wolfgang Doeblin.*

Couplings of Markov chains

Coupling techniques have proved very successful, in some cases giving precise rates of convergence.

See for example

Jerrum, 1998, *Mathematical foundations of the MCMC method.*

Using W_1 instead of TV,

$$
\begin{aligned}
& W_1(\pi_t, \pi) \\
&= \inf_{X_t, Y_t \in \text{coupling}(\pi_t, \pi)} \mathbb{E}[d(X_t, Y_t)] \\
&\leq \mathbb{E}[d(X_t, Y_t)].
\end{aligned}
$$

Recall that we cannot sample $Y_0 \sim \pi$.

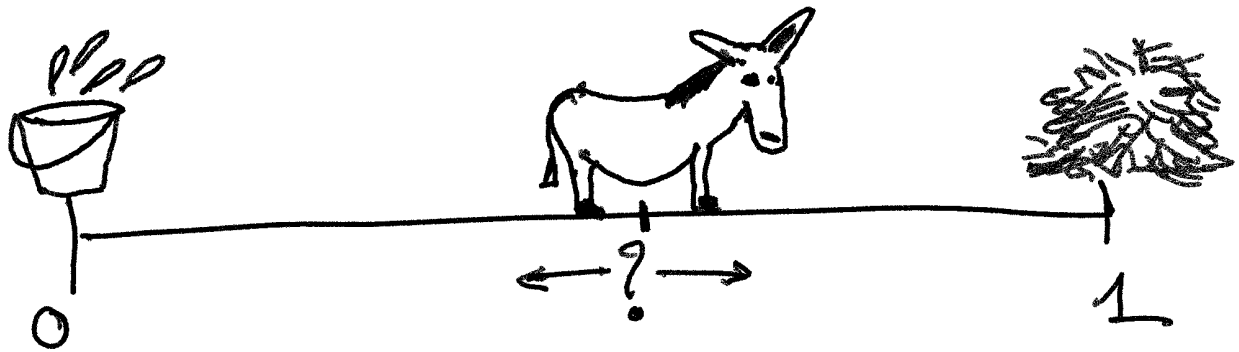
Donkey walk

Let $U_{t,\curvearrowleft}, U_{t,\curvearrowright} \sim \text{Uniform}(0,1)$ be independent.
Consider the chain (Z_t) on $[0,1]$ evolving as

$$\text{(step to the left)} \quad Z_{t-1/2} = U_{t,\curvearrowleft} Z_{t-1},$$
$$\text{(step to the right)} \quad Z_t = Z_{t-1/2} + U_{t,\curvearrowright}(1 - Z_{t-1/2}),$$

or, in one step:

$$Z_t = U_{t,\curvearrowleft}(1 - U_{t,\curvearrowright}) Z_{t-1} + U_{t,\curvearrowright}.$$



Letac, 2002, *Donkey walk and Dirichlet distributions.*

Donkey walk

A "common random numbers" coupling

$$Z_t = U_{t,\curvearrowleft}(1 - U_{t,\curvearrowright})Z_{t-1} + U_{t,\curvearrowright},$$
$$\tilde{Z}_t = U_{t,\curvearrowleft}(1 - U_{t,\curvearrowright})\tilde{Z}_{t-1} + U_{t,\curvearrowright},$$

leads to

$$|\pi_t - \pi|_{W_1} \leq \left(\frac{1}{4}\right)^t \mathbb{E}\left[\left|Z_0 - \tilde{Z}_0\right|\right].$$

Very explicit rate. Turns out to be sharp in this case.

Langevin diffusion with strongly convex potential

Langevin diffusion:

$$dX_t = -\nabla U(X_t)dt + \sqrt{2}dB_t,$$

where U is the potential function.

Introduce (Y_t) that evolves with the same Brownian motion (B_t). Then $\alpha_t = X_t - Y_t$ satisfy

$$\frac{d\alpha_t}{dt} = -(\nabla U(X_t) - \nabla U(Y_t)).$$

In particular,

$$\frac{d|\alpha_t|^2}{dt} = -2\langle \nabla U(X_t) - \nabla U(Y_t), X_t - Y_t \rangle \leq -2K|\alpha_t|^2,$$

if we assume that $\nabla^2 U \geq K\mathrm{Id}$ (U is K-strongly convex).

Langevin diffusion with strongly convex potential

Gronwall's lemma: $f'(t) \leq Cf(t) \Rightarrow f(t) \leq \exp(Ct)f(0)$.

Therefore:

$$|\alpha_t|^2 \leq \exp(-2Kt)|\alpha_0|^2.$$

Then

$$\begin{aligned} W_2^2(\pi_t, \pi) &\leq \mathbb{E}[|X_t - Y_t|^2| \\ &\leq 2(\mathbb{E}[|X_0|^2] + \mathbb{E}[|Y_0|^2]) \exp(-2Kt). \end{aligned}$$

Let's assume finiteness of $\mathbb{E}[|X_0|^2]$ and $\mathbb{E}[|Y_0|^2]$.

$\Rightarrow$ Geometric ergodicity of the Langevin diffusion.

Pages 22-23 in Villani, 2009, *Optimal transport: old and new.*

Unadjusted Langevin algorithm

Same reasoning works for ULA, with common noise:

$$X^\epsilon_{t+1} = X^\epsilon_t + \epsilon \nabla \log \pi(X^\epsilon_t) + \sqrt{2\epsilon} Z_{t+1},$$
$$Y^\epsilon_{t+1} = Y^\epsilon_t + \epsilon \nabla \log \pi(Y^\epsilon_t) + \sqrt{2\epsilon} Z_{t+1}, \quad Z_{t+1} \sim \text{Normal}(0, 1).$$

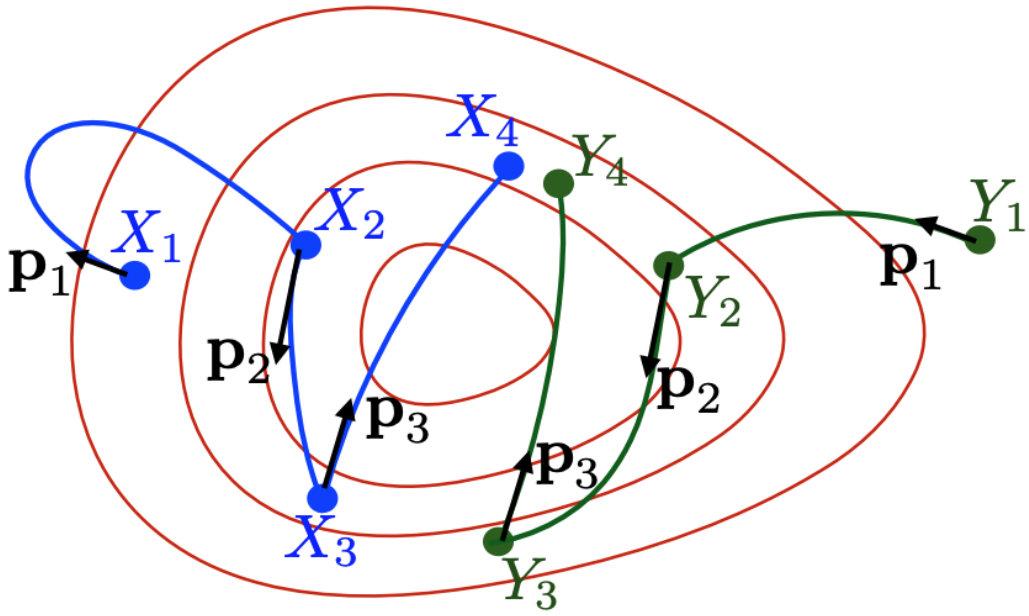
Then

$$|X^\epsilon_{t+1} - Y^\epsilon_{t+1}|^2 \leq \left(1 - \frac{2\epsilon K L}{K + L}\right) |X^\epsilon_t - Y^\epsilon_t|^2,$$

where $L\text{Id} \geq \nabla^2 U \geq K\text{Id}$.

Appendix A of Wisibono, 2018, *Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem.*

Hamiltonian Monte Carlo



Coupling two copies $X_1, X_2, \ldots$ (blue) and $Y_1, Y_2, \ldots$ (green) of HMC by choosing same momentum at every step.

Figure 2 of Mangoubi & Smith, 2017, *Rapid mixing of HMC strongly log-concave distributions.*

Example of recent convergence results

HMC (best $K > 1$)	Initialization	Extra assumption	#Gradient Evals
[BPR+13]	warm	product distribution	$d^{\frac{1}{4}}\dagger$
[CDWY20]	$\mathcal{N}(x^*, L^{-1}\mathbb{I}_d)$	Hessian Lipschitz	$d^{\frac{11}{12}}\kappa$
this work	warm	strongly Hessian Lipschitz	$d^{\frac{1}{4}}\kappa$
MALA ($K = 1$)	-	-	-
[RR98]	warm	product distribution	$d^{\frac{1}{3}}\dagger$
[DCWY19] [CDWY20]	warm / $\mathcal{N}(x^*, L^{-1}\mathbb{I}_d)$	N/A	$\max\{d^{\frac{1}{2}}\kappa^{\frac{3}{2}}, d\kappa\}$
[LST20]	$\mathcal{N}(x^*, L^{-1}\mathbb{I}_d)$	N/A	$d\kappa$
[CLA+21]	warm	N/A	$d^{\frac{1}{2}}\kappa^{\frac{3}{2}}$
[WSC22]	warm	N/A	$d^{\frac{1}{2}}\kappa$
[AC23]	$\mathcal{N}(x^*, L^{-1}\mathbb{I}_d)$	N/A	$d^{\frac{1}{2}}\kappa$
this work	warm	strongly Hessian Lipschitz	$d^{\frac{3}{7}}\kappa$

Table 1. Summary of ϵ-mixing time in TV distance of HMC and MALA for sampling a L-log-smooth and m-strongly log-concave target under a warm start or a Gaussian initialization $\mathcal{N}(x^*, L^{-1}\mathbb{I}_d)$, where x^* denotes the unique mode of the target density. In the results with extra assumptions, the Hessian Lipschitz or strongly Hessian Lipschitz coefficients are considered constant. The MALA results can be considered as HMC results with the number of leapfrog steps K chosen to be 1. These statements hide constants and logarithmic factors in d, ϵ^{-1} or $\kappa = L/m$. $\dagger$ The κ dependency is unknown and higher-order derivatives are assumed in [RR98] and [BPR+13].

Chen & Gatmiry, 2023, *When does Metropolized Hamiltonian Monte Carlo provably outperform Metropolis-adjusted Langevin algorithm?*

Outline

Langevin diffusion as gradient descent

Langevin diffusion:

$$dX_t = -\nabla U(X_t)dt + \sqrt{2}dB_t.$$

Its density $x \mapsto \rho(x,t)$ satisfies the Fokker–Planck equation:

$$\frac{d\rho}{dt} = \mathrm{div}(\rho \nabla U) + \Delta \rho,$$

which is the gradient flow of the relative entropy $\mathrm{KL}(\rho|\pi) = \int \log(\rho/\pi)\rho$, in the space of measure with the 2-Wasserstein metric:

$$\frac{d\rho}{dt} = -\nabla_{W_2} \mathrm{KL}(\rho|\pi).$$

Jordan, Kinderlehrer & Otto, 1998, *The variational formulation of the Fokker—Planck equation.*

Unadjusted Langevin = forward flow discretization

We can write unadjusted Langevin as

$$X^\epsilon_{t+1/2} = X^\epsilon_t + \epsilon \nabla \log \pi(X^\epsilon_t),$$
$$X^\epsilon_{t+1} = X^\epsilon_{t+1/2} + \sqrt{2\epsilon} Z_{t+1}.$$

In the space of measures:

$$\rho^\epsilon_{t+1/2} = (I - \epsilon \nabla \log \pi)_\sharp \rho^\epsilon_t,$$
$$\rho^\epsilon_{t+1} = \text{Normal}(0, 2\epsilon I) * \rho^\epsilon_{t+1/2}.$$

First step is gradient descent for $\mathbb{E}_\rho[\log \pi]$.
Second step is gradient flow for the negative entropy $-\mathbb{E}_\rho[\log \rho]$.
Sum of these two terms: relative entropy $\text{KL}(\rho|\pi)$.
Bernton, 2018, *Langevin Monte Carlo and JKO splitting*,
Wisibono, 2018, *Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem.*

Coming soon. . .

- The theory of MCMC helps to understand the pros and cons of classes of MCMC algorithms, how mixing times may scale with the number of data points, the number of covariates, or the "regularity" of the target distribution.
- However, it (typically) won't tell you how many iterations to perform. Some exceptions exist.

  Rosenthal, 1995, *Minorization conditions and convergence rates for Markov chain Monte Carlo*.
- Practitioners need to choose a number of iterations, or come up with a stopping criterion.
- In the next two lectures, we'll see methods that employ couplings to address these practical questions.

Outline

Outline