Quantitative convergence analysis of hypocoercive sampling dynamics

Jianfeng Lu (鲁剑锋)

Duke University
jianfeng@math.duke.edu

MATH+IMS Joint Applied Mathematics Colloquium
Chinese University of Hong Kong, September 2020

Joint with
Yu Cao (NYU)
Lihan Wang (Duke)

Duke UNIVERSITY

$$\mathbb{E}_{X\sim\mu} f(X) \approx \frac{1}{N}\sum_{i=1}^{N} f(X_i)$$

Sampling high dimensional probability distributions is a ubiquitous challenge in many fields:

- computational statistical mechanics;
- machine learning;
- Bayesian statistics;
- high-dimensional PDEs;
- quantum many-body problems;
- ...

A popular approach is Markov chain Monte Carlo: $\{X_i\}$ sampled from a Markov chain $X_{i+1} \sim p(\cdot|X_i)$ with invariant measure μ.

$$\mathbb{E}_{X\sim\mu} f(X) \approx \frac{1}{N}\sum_{i=1}^{N} f(X_i)$$

Markov chain Monte Carlo: $\{X_i\}$ sampled from a Markov chain $X_{i+1} \sim p(\cdot|X_i)$ with invariant measure μ.

Central limit theorem holds for "nice" Markov chains:

$$\sqrt{N}\Big(\frac{1}{N}\sum_{i=1}^{N} f(X_i) - \mathbb{E}_{\mu} f(X)\Big) \xrightarrow{d} \mathcal{N}(0,\sigma^2)$$

with asymptotic variance

$$\sigma^2 = \mathrm{var}[f(X_i)] + 2\sum_{k=1}^{\infty} \mathrm{cov}[f(X_i), f(X_{i+k})].$$

Efficiency of the MCMC sampler desires

- Short burn-in period;
- Small asymptotic variance.

Efficiency of the MCMC sampler desires

- Short burn-in period;
- Small asymptotic variance.

This talk: Continuous state space $x \in \mathbb{R}^d$, in particular $d \gg 1$

Common design principle:
Construct a continuous time Markov process and then discretize.

Example: Overdamped Langevin dynamics for $\mathrm{d}\mu \propto e^{-U(x)}\,\mathrm{d}x$

$$\mathrm{d}x_t = -\nabla U(x_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t.$$

[Rossky, Doll, Friedman 1978]; [Besag 1994]; [Roberts, Tweedie 1996]

Hope for fast convergence to equilibrium of the sampling dynamics.

Overdamped Langevin dynamics for $\mathrm{d}\mu \propto e^{-U(x)}\,\mathrm{d}x$

$$\mathrm{d}x_t = -\nabla U(x_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t.$$

The Fokker-Planck equation (backward Kolmogorov equation)

$$\partial_t h = -\nabla_x U \cdot \nabla_x h + \Delta_x h, \qquad h(0,x) = h_0(x).$$

The convergence of Fokker-Planck equation is well understood, as the generator is self-adjoint and coercive with respect to L^2_μ.

Assumption (Poincaré inequality for μ)

$$\int \Big(h - \int h\,\mathrm{d}\mu\Big)^2 \mathrm{d}\mu \le \frac{1}{m}\int |\nabla_x h|^2\,\mathrm{d}\mu$$

This implies that the overdamped dynamics has convergence rate m.

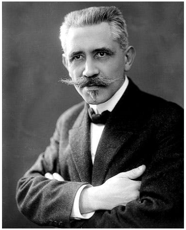
$$\|h(t,\cdot) - \int h(t,\cdot)\,\mathrm{d}\mu\|_{L^2(\mu)} \le e^{-mt}\|h(0,\cdot) - \int h(0,\cdot)\,\mathrm{d}\mu\|_{L^2(\mu)}$$

Our motivation is to establish quantitative convergence rate estimate for hypocoercive sampling dynamics.

Our first example is the (underdamped) Langevin dynamics

$$
\begin{aligned}
\mathrm{d}x_t &= v_t \, \mathrm{d}t \\
\mathrm{d}v_t &= -\nabla U(x_t) \, \mathrm{d}t - \gamma v_t \, \mathrm{d}t + \sqrt{2\gamma} \, \mathrm{d}W_t
\end{aligned}
$$

Here γ is a friction parameter.



Paul Langevin (1872–1946)

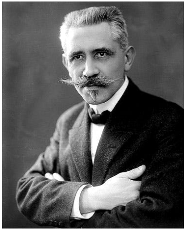
Our motivation is to establish quantitative convergence rate estimate for hypocoercive sampling dynamics.

Our first example is the (underdamped) Langevin dynamics

$$\mathrm{d}x_t = v_t\,\mathrm{d}t$$

$$\mathrm{d}v_t = -\nabla U(x_t)\,\mathrm{d}t - \gamma v_t\,\mathrm{d}t + \sqrt{2\gamma}\,\mathrm{d}W_t$$

Here γ is a friction parameter.



Paul Langevin (1872–1946)

As $\gamma \to \infty$, and after a time rescaling, we will recover the overdamped Langevin dynamics $\mathrm{d}x_t = -\nabla U(x_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t$.

The invariant measure of the Langevin dynamics is given by

$$\rho_\infty(\mathrm{d}x, \mathrm{d}v) = \frac{1}{Z} e^{-U(x) - \frac{1}{2}|v|^2}\,\mathrm{d}x\,\mathrm{d}v,$$

where Z is the normalizing constant. The marginal distribution is μ.

Langevin dynamics

$$\mathrm{d}x_t = v_t\,\mathrm{d}t$$
$$\mathrm{d}v_t = -\nabla U(x_t)\,\mathrm{d}t - \gamma v_t\,\mathrm{d}t + \sqrt{2\gamma}\,\mathrm{d}W_t$$

The corresponding backward Kolmogorov equation, known as the kinetic Fokker-Planck equation, is given by

$$\partial_t f = \mathcal{L} f$$
$$f(0,x,v) = f_0(x,v)$$

with the generator given by $\mathcal{L} = \mathcal{L}_{\mathsf{ham}} + \gamma\mathcal{L}_{\mathsf{FD}}$ with

$$\mathcal{L}_{\mathsf{ham}} = v\cdot\nabla_x - \nabla_x U\cdot\nabla_v \qquad \text{and} \qquad \mathcal{L}_{\mathsf{FD}} = \Delta_v - v\cdot\nabla_v$$

We can verify that $\mathcal{L}^*\rho_\infty = 0$, and thus ρ_∞ the invariant measure.

Recall that the overdamped Langevin dynamics converges with rate m, where m is the Poincaré constant of μ.

Question: Any improvement by the underdamped Langevin dynamics?

Recall that the overdamped Langevin dynamics converges with rate m, where m is the Poincaré constant of μ.

Question: Any improvement by the underdamped Langevin dynamics?

Theorem (Cao-L.-Wang 2019)

For convex U satisfying $|\mathrm{Hess}\, U| \lesssim (1 + |\nabla U|)$ and superlinear as $|x| \to \infty$,

$$\|f(t,\cdot) - \int f(t,\cdot)\, \mathrm{d}\rho_\infty\|_{L^2(\rho_\infty)} \leq C_0 \exp(-\lambda t) \|f(0,\cdot) - \int f(0,\cdot)\, \mathrm{d}\rho_\infty\|_{L^2(\rho_\infty)}$$

with explicit estimate of λ as

$$\lambda = \sqrt{m} \log\Big(1 + \frac{\gamma\sqrt{m}}{c_0(\sqrt{m} + \gamma)^2}\Big)$$

Recall that the overdamped Langevin dynamics converges with rate m, where m is the Poincaré constant of μ.

Question: Any improvement by the underdamped Langevin dynamics?

Theorem (Cao-L.-Wang 2019)

For convex U satisfying $|\mathrm{Hess}\, U| \lesssim (1 + |\nabla U|)$ and superlinear as $|x| \to \infty$,

$$\|f(t,\cdot) - \int f(t,\cdot)\, \mathrm{d}\rho_\infty\|_{L^2(\rho_\infty)} \leq C_0 \exp(-\lambda t) \|f(0,\cdot) - \int f(0,\cdot)\, \mathrm{d}\rho_\infty\|_{L^2(\rho_\infty)}$$

with explicit estimate of λ as

$$\lambda = \sqrt{m} \log\Big(1 + \frac{\gamma\sqrt{m}}{c_0(\sqrt{m} + \gamma)^2}\Big)$$

$$= \mathcal{O}(\sqrt{m}) \qquad \text{if we take } \gamma = \mathcal{O}(\sqrt{m}).$$

Results available for more general case; we will not discuss those here.

Exponential convergence with rate $\sqrt{m}$ (setting $\int f \, \mathrm{d}\rho_\infty = 0$)

$$\|f(t,\cdot)\|_{L^2(\rho_\infty)} \leq C_0 \exp(-c\sqrt{m}t)\|f(0,\cdot)\|_{L^2(\rho_\infty)}$$

- The $\mathcal{O}(\sqrt{m})$ convergence rate is optimal, as can be seen when U is a Gaussian (so explicit calculation can be done);
- First result in literature for sharp $\sqrt{m}$ convergence rate (acceleration compared with overdamped dynamics with rate m);
- Convergence in L^2 implies convergence of density in χ^2-divergence, and thus in relatively entropy and total variation distance with $\mathcal{O}(\sqrt{m})$ rate.

Our analysis also applies to piecewise deterministic Markov process: Deterministic trajectory between Poisson clocks for random bounces and velocity refreshment

Randomized Hamiltonian Monte Carlo

[Duane, Kennedy, Pendleton, Roweth 1987]; [Bou-Rabee, Sanz-Serna 2017]

$$\mathcal{L} = \underbrace{v \cdot \nabla_x - \nabla_x U \cdot \nabla_v}_{\text{Hamiltonian flow}} + \gamma(\Pi_v - \mathcal{I})$$

Π_v: projection on Gaussian (velocity refreshment)

$$(\Pi_v f)(t,x) := \int f(t,x,v)\kappa(\,\mathrm{d}v)$$

Deterministic Hamiltonian flow in between random (Poisson clock with rate γ) velocity refreshment drawn from Gaussian.

Zigzag sampler (ZZ) [Bierkens, Fearnhead, Roberts 2019]

$$\mathcal{L} = v \cdot \nabla_x + \sum_{k=1}^{d} \underbrace{(v_k \partial_{x_k} U)_+}_{\text{bouncing rate}} (\mathcal{B}_k - \mathcal{I}) + \gamma(\Pi_v - \mathcal{I})$$

with bouncing operators (flipping the k-th velocity component)

$$\mathcal{B}_k f = f(x, v - 2v_k e_k), \qquad k = 1, \cdots, d$$

Bouncy particle sampler (BPS)

[Peters, de With 2012] [Bouchard-Côté, Vollmer, Doucet 2018]

$$\mathcal{L} = v \cdot \nabla_x + (v \cdot \nabla U)_+ (\mathcal{B} - \mathcal{I}) + \gamma(\Pi_v - \mathcal{I})$$

with bouncing operator (flipping wrt hyperplane perpendicular to ∇U)

$$\mathcal{B} f = f\left(x, v - 2(v \cdot \nabla U) \frac{\nabla U}{|\nabla U|^2}\right)$$

Promising approaches in the context of stochastic gradient.

Theorem (L.-Wang 2020)

For convex U satisfying $|\mathrm{Hess}\, U| \lesssim (1 + |\nabla U|)$ and superlinear as $|x| \to \infty$, all three PDMPs converge exponentially to equilibrium with rates (after an optimal choice of velocity freshment rate γ)

$$\nu = \begin{cases} \mathcal{O}(\sqrt{m}), & \text{for RHMC;} \\ \mathcal{O}\left(\frac{\sqrt{m}}{\sqrt{L/m}}\right), & \text{for ZZ;} \\ \mathcal{O}\left(\frac{\sqrt{m}}{\sqrt{d}}\right), & \text{for BPS,} \end{cases}$$

where for the zigzag sampler, we assume in addition that $\nabla^2 U \leq L$.

Results available for more general case; we will not discuss those here.

The rate is optimal for RHMC; for ZZ and BPS, our rate estimate is more quantitative than previous results in [Deligiannidis, Paulin, Bouchard-Côté, Doucet 2018]; [Andrieu, Durmus, Nüsken, Roussel 2018] (which only considered rate dependence in d).

Convergence analysis of kinetic Fokker-Planck equation

$$\partial_t f = \mathcal{L} f = \big(\mathcal{L}_{\mathsf{ham}} + \gamma \mathcal{L}_{\mathsf{FD}}\big) f; \qquad f(0, x, v) = f_0(x, v),$$

where

$$\mathcal{L}_{\mathsf{ham}} = v \cdot \nabla_x - \nabla_x U \cdot \nabla_v \qquad \text{and} \qquad \mathcal{L}_{\mathsf{FD}} = \Delta_v - v \cdot \nabla_v$$

The operator is not elliptic and only hypo-elliptic [Hörmander 1967] (as the diffusion is degenerate in the x direction).

In particular, we cannot hope for the exponential convergence to follow from a Poincaré (coercivity) estimate for $\mathcal{L}$. As a result, the constant $C_0 > 1$ is unavoidable in the decay estimate

$$\|f(t, \cdot)\|_{L^2(\rho_\infty)} \leq C_0 \exp(-c\sqrt{m}t) \|f(0, \cdot)\|_{L^2(\rho_\infty)}.$$

Previous results on quantitative convergence of Langevin dynamics:

- Convergence in $H^1_{\rho_\infty}$ norm [Villani 2009];
- Convergence in a modified $L^2_{\rho_\infty}$ norm [Dolbeault, Mouhot, Schmeiser 2009; 2015] (also earlier idea from [Herau 2006]).
  This was applied to kinetic Fokker-Planck equation by [Roussel, Stoltz 2018], which gives explicit rate estimates, though not sharp
- Very recent result based on resolvent analysis using Schur complement [Bernard, Fathi, Levitt, Stoltz 2020]
- Convergence in Wasserstein distance: using Bakry-Émery framework [Boudoin 2016]; by coupling approaches [Eberle, Guillin, Zimmer 2019; Dalalyan, Riou-Durand 2018]
- Convergence based on Lyapunov function [Mattingly, Stuart, Higham 2002]

Our analysis method was inspired by a recent variational framework [Armstrong, Mourrat 2019], which implicitly used the bracket condition dating back to [Hörmander 1967].

As $\mathcal{L}$ is not coercive, the idea is to resort to augmenting the state space by a time interval $I = (0, T)$ equipped with Lebesgue measure λ. Since in time, the diffusion in v direction will propagate to the x direction.

Our analysis method was inspired by a recent variational framework [Armstrong, Mourrat 2019], which implicitly used the bracket condition dating back to [Hörmander 1967].

As $\mathcal{L}$ is not coercive, the idea is to resort to augmenting the state space by a time interval $I = (0, T)$ equipped with Lebesgue measure λ. Since in time, the diffusion in v direction will propagate to the x direction.

Let κ be the Gaussian measure in velocity ($\rho_\infty(\mathrm{d}x\,\mathrm{d}v) = \mu(\mathrm{d}x)\kappa(\mathrm{d}v)$). The exp. conv. follows from an energy estimate combined with

Theorem (Poincaré inequality in time augmented state space)

$$\|f - (f)_{\lambda\times\mu}\|_{L^2(\lambda\times\mu;L^2_\kappa)} \lesssim \big(1 + \frac{1}{T\sqrt{m}}\big)\|f - \Pi_v f\|_{L^2(\lambda\times\mu;L^2_\kappa)} + \big(\frac{1}{\sqrt{m}} + T\big)\|\partial_t f - \mathcal{L}_{ham} f\|_{L^2(\lambda\times\mu;H^{-1}_\kappa)},$$

where $(f)_{\lambda\times\mu} := \frac{1}{T}\int f(t,x,v)\,\mathrm{d}t\,\mathrm{d}\rho_\infty$.

Proof sketch: Without loss of generality, we assume $(f)_{\lambda\times\mu} = 0$.
By triangular inequality

$$\|f\|_{L^2(\lambda\times\mu;L^2_\kappa)} \leq \|f - \Pi_v f\|_{L^2(\lambda\times\mu;L^2_\kappa)} + \|\Pi_v f\|_{L^2(\lambda\times\mu)}.$$

Proof sketch: Without loss of generality, we assume $(f)_{\lambda\times\mu} = 0$.

By triangular inequality

$$\|f\|_{L^2(\lambda\times\mu;L^2_\kappa)} \leq \|f - \Pi_v f\|_{L^2(\lambda\times\mu;L^2_\kappa)} + \|\Pi_v f\|_{L^2(\lambda\times\mu)}.$$

For the underdamped Langevin, using Gaussian Poincaré inequality, we have

$$\|f - \Pi_v f\|_{L^2(\lambda\times\mu;L^2_\kappa)} \leq \|\nabla_v f\|_{L^2(\lambda\times\mu;L^2_\kappa)}.$$

The estimate for $\|\Pi_v f\|_{L^2(\lambda\times\mu)}$ is more tricky.

We desire to control it with the help of $\|\partial_t f - \mathcal{L}_{\mathrm{ham}} f\|_{L^2(\lambda\times\mu;H^{-1}_\kappa)}$, thus, we need to "introduce derivatives".

The idea is to construct "test functions" $(\phi_0, \phi_1, \dots, \phi_d) \in H_0^1(\lambda \times \mu)$ by solving a divergence equation

$$-\partial_t \phi_0 + \nabla_x^* \cdot \phi = -\partial_t \phi_0 + \sum_{i=1}^{d} \left(-\partial_{x_i} \phi_i + \partial_{x_i} U \phi_i\right) = \Pi_v f$$

with Dirichlet boundary conditions (note that $\int_{I \times \mathbb{R}^d} \Pi_v f \, \mathrm{d}t \mu(\mathrm{d}x) = 0$).

Lemma

$$\left(\sum_{i=0}^{d} \|\phi_i\|^2_{L^2(\lambda \times \mu)}\right)^{1/2} \lesssim \max\{m^{-1/2}, T\} \, \|\Pi_v f\|_{L^2(\lambda \times \mu)};$$

$$\left(\sum_{i,j=0}^{d} \|\partial_j \phi_i\|^2_{L^2(\lambda \times \mu)}\right)^{1/2} \lesssim \left(1 + m^{-1/2} T^{-1}\right) \|\Pi_v f\|_{L^2(\lambda \times \mu)}.$$

$$
\begin{aligned}
\|\Pi_v f\|^2_{L^2(\lambda\times\mu)} &= \int_{I\times\mathbb{R}^d} \Pi_v f\big(-\partial_t\phi_0 + \nabla_x^* \cdot \phi\big)\,\mathrm{d}t\mu(\,\mathrm{d}x) \\
&= \int_{I\times\mathbb{R}^d\times\mathbb{R}^d} \Pi_v f\big(-\partial_t\phi_0 + v\cdot\nabla_x\phi_0 \\
&\qquad + v\cdot\partial_t\phi - \sum_i v_i v\cdot\partial_{x_i}\phi + \phi\cdot\nabla_x U\big)\,\mathrm{d}t\rho_\infty(\,\mathrm{d}x\,\mathrm{d}v) \\
&\qquad\qquad \text{(reintroducing } v \text{ using Gaussianity)}
\end{aligned}
$$

After splitting $\Pi_v f$ into f and $\Pi_v f - f$, using integration by parts, we get

$$
\begin{aligned}
\|\Pi_v f\|^2_{L^2(\lambda\times\mu)} \le{}& \|\partial_t f - \mathcal{L}_{\mathrm{ham}} f\|_{L^2(\lambda\times\mu;H_\kappa^{-1})}\|\phi_0 - v\cdot\phi\|_{L^2(\lambda\times\mu;H_\kappa^1)} \\
&+ \big\|-\partial_t\phi_0 + v\cdot\nabla_x\phi_0 + v\cdot\partial_t\phi - \sum_i v_i v\cdot\partial_{x_i}\phi \\
&\qquad + \phi\cdot\nabla_x U\big\|_{L^2(\lambda\times\mu;L^2_\kappa)}\|f - \Pi_v f\|_{L^2(\lambda\times\mu;L^2_\kappa)}.
\end{aligned}
$$

The Poincaré inequality follows from estimate of ϕ and assumption of U.

Quantitative convergence for hypocoercive sampling dynamics based on time-augmented Poincaré inequalities.

- Underdamped Langevin dynamics;
- Randomized Hamiltonian Monte Carlo;
- Zigzag sampler;
- Bouncy particle sampler.

Quantitative convergence for hypocoercive sampling dynamics based on time-augmented Poincaré inequalities.

- Underdamped Langevin dynamics;
- Randomized Hamiltonian Monte Carlo;
- Zigzag sampler;
- Bouncy particle sampler.

We did not discuss in this talk the convergence of the sampling algorithm based on discretization; non-asymptotic analysis of those has been an active research area in machine learning and statistics literature. See e.g.,

- Discretized underdamped Langevin dynamics: [Cheng, Chatterji, Bartlett, Jordan 2018] [Dalalyan, Riou-Durand 2018] [Mou, Ma, Wainwright, Bartlett, Jordan 2019]; [Shen, Lee 2019];
- Discretized Hamiltonian Monte Carlo: [Mangoubi, Vishnoi 2018]; [Lee, Song, Vempala 2018]; [Chen, Vempala 2019]; [Bou-Rabee, Eberle, Zimmer 2020];

Thank you! Any questions?

Email: jianfeng@math.duke.edu

URL: http://www.math.duke.edu/~jianfeng/

References:

- with Yu Cao and Lihan Wang, *On explicit L^2-convergence rate estimate for underdamped Langevin dynamics*, arXiv:1908.04746
- with Lihan Wang, *On explicit L^2-convergence rate estimate for piecewise deterministic Markov process*, arXiv:2007.14927