Convergence of Overdamped Langevin in KL Divergence and a Hierarchical Entropy Method for Low Dimensional Marginals

Kaiwen Zhang

Courant Institute, New York University

31 January 2026

Primary References

- Santosh S. Vempala and Andre Wibisono. Rapid Convergence of the Unadjusted Langevin Algorithm: Isoperimetry Suffices, pages 381–438. Springer International Publishing, Cham, 2023.
- Daniel Lacker and Fuzhong Zhou. A hierarchical entropy method for the delocalization of bias in high-dimensional langevin monte carlo, 2025.

Week 1: convergence in KL (Vempala-Wibisono)
Week 2: delocalization of bias (Lacker-Zhou)

Introduction

- Consider the overdamped Langevin dynamics:

$$dX_t = -\nabla V(X_t)dt + \sqrt{2}dB_t \tag{1}$$

  and its Euler-Maruyama discretization

$$X_{(k+1)h} = X_{kh} - h\nabla V(X_{kh}) + \sqrt{2}(B_{(k+1)h} - B_{kh}) \tag{2}$$

- Continuous dynamics: the distribution of X_t converges to stationary distribution $\pi \propto e^{-V}$.
- Discrete scheme: the distribution of X_{kh} converges to a biased stationary distribution π_h.
- Convergence and bias bounds in W_2 distance ($O(\sqrt{dh})$) proved by coupling argument.

Definitions: KL divergence

- The KL divergence of ρ w.r.t. a base measure ν is (when they both have densities)

$$H_\nu(\rho) = \int_{\mathbb{R}^n} \rho(x) \log \frac{\rho(x)}{\nu(x)} dx$$

or generally when $\rho << \nu$

$$H_\nu(\rho) = \int_{\mathbb{R}^n} f(x) \log f(x) d\nu(x), f = \frac{d\rho}{d\nu}$$

- The relative fisher information of ρ w.r.t. ν is

$$J_\nu(\rho) = \int_{\mathbb{R}^n} \rho(x) \left\| \nabla \log \frac{\rho(x)}{\nu(x)} \right\|^2 dx$$

or more generally

$$J_\nu(\rho) = \int_{\mathbb{R}^n} \|\nabla \log f(x)\|^2 d\rho(x), f = \frac{d\rho}{d\nu}$$

Definition: Log-Sobolev Inequality

A probability measure ν is said to satisfy the Log-Sobolev inequality with constant α if for any probability measure ρ,

$$H_\nu(\rho) \leq \frac{1}{2\alpha} J_\nu(\rho)$$

Or, equivalently, for any smooth g such that $\mathbb{E}_\nu[g^2] < \infty$,

$$\mathbb{E}_\nu[g^2 \log g^2] - \mathbb{E}_\nu[g^2] \log \mathbb{E}_\nu[g^2] \leq \frac{2}{\alpha} \mathbb{E}_\nu[\|\nabla g\|^2]$$

- Any strongly log-concave measure satisfies the log-Sobolev inequality (Bakry-Emery); preserved under bounded perturbations.
- **Theorem** (Otto-Villani): any measure ν satisfying the log-Sobolev inequality also satisfies the Talagrand inequality, i.e. for all probability measures ρ, $\frac{\alpha}{2} W_2(\rho, \nu)^2 \leq H_\nu(\rho)$
- (Lacker-Zhou) If $\pi \propto e^{-V}$ satisfies log-Sobolev with constant α and ∇V is β-Lipschitz, then $\alpha \leq \beta$.

Convergence of overdamped Langevin: strategy

1. Derive a Fokker-Planck equation satisfied by the distributions of interest
2. Using the PDE, differentiate the KL divergence in time: one needs to resolve the technicality in differentiability -
   - use regularity theory of the Fokker-Planck equation to gain enough derivatives,
   - use the weak formulation to differentiate, or
   - use density of smooth compactly supported functions in distributions to somehow approximate
3. Extract a negative definite term using the log-Sobolev inequality, and examine the rest:
   - For continuous dynamics 1, there is no other term;
   - For discrete dynamics 2, there is a time-discretization error term;
4. Use log-Sobolev and Talagrand inequalities and other statistical bounds to treat the error terms, and deduce a differential inequality;
5. Conclude the decay estimate via integrating the above inequality.

Continuous dynamics: Fokker-Planck I

Assume: the target measure π satisfies log-Sobolev inequality with constant α.

Denote ρ_t as the distribution of X_t. Suppose f is smooth and compactly supported in $\mathbb{R}^n$, apply Ito's formula and use $dX_t = -\nabla V(X_t)dt + \sqrt{2}dB_t$

$$f(X_t) = f(X_0) - \sum_{j=1}^{n} \int_0^t \partial_j f(X_s)\partial_j V(X_s)ds + \sum_{j=1}^{n} \int_0^t \partial_j f(X_s)\sqrt{2}dB_s^j + \sum_{j=1}^{n} \int_0^t \partial_j^2 f(X_s)ds$$

The second term has zero expectation. As a result:

$$\mathbb{E}[f(X_t)] = \mathbb{E}[f(X_0)] - \mathbb{E}[\int_0^t \nabla f(X_s) \cdot \nabla V(X_s)ds] + \mathbb{E}[\int_0^t \Delta f(X_s)ds]$$

and differentiation gives

$$\frac{d}{dt}\mathbb{E}[f(X_t)] = -\mathbb{E}[\nabla f(X_t) \cdot \nabla V(X_t)] + \mathbb{E}[\Delta f(X_t)]$$

Continuous dynamics: Fokker-Planck II

We use the pairing notation $\mathbb{E}[f(X_t)] = \langle \rho_t, f \rangle$

- Left: $\frac{d}{dt}\mathbb{E}[f(X_t)] = \frac{d}{dt}\int_{\mathbb{R}^n} f\rho_t = \int_{\mathbb{R}^n} \frac{d}{dt}(f\rho_t) = \int_{\mathbb{R}^n} f\frac{d}{dt}\rho_t = \langle f, \partial_t \rho_t \rangle$
- Right: divergence term:

$$f\nabla \cdot (\rho_t \nabla V) = \nabla \cdot (f\rho_t \nabla V) - \rho_t \nabla f \cdot \nabla V$$

  In spacial integration, the second term vanishes. Thus

$$-\int_{\mathbb{R}^n} \rho_t \nabla f \cdot \nabla V = -\mathbb{E}[\nabla f(X_t) \cdot \nabla V(X_t)] = \langle f, \nabla \cdot (\rho_t \nabla V) \rangle$$

- Right: Laplacian term: using $u\Delta v = \nabla \cdot (u\nabla v) - \nabla u \cdot \nabla v$ twice

$$\mathbb{E}[\Delta f(X_t)] = \int_{\mathbb{R}^n} \rho_t \Delta f = \int_{\mathbb{R}^n} f\Delta \rho_t = \langle f, \Delta \rho_t \rangle$$

Thus ρ_t solves in distributional sense:

$$\partial_t \rho_t = \nabla \cdot (\rho_t \nabla V) + \Delta \rho_t = \nabla \cdot \left(\rho_t \nabla \log \frac{\rho_t}{\pi}\right) \tag{3}$$

Continuous dynamics: differentiating KL

$$\partial_t H_\pi(\rho_t) = \partial_t \int \rho_t \log \frac{\rho_t}{\pi} = \partial_t \int \rho_t \log \rho_t + \rho_t V$$

$$= \int (\partial_t \rho_t) \log \rho_t + \partial_t \rho_t + V \partial_t \rho_t = \int \partial_t \rho_t (\log \rho_t + V) + 0$$

$$= \int (\partial_t \rho_t) \log \frac{\rho_t}{\pi} = \int \nabla \cdot (\rho_t \nabla V + \nabla \rho_t) \log \frac{\rho_t}{\pi}$$

$$= -\int (\rho_t \nabla V + \nabla \rho_t) \cdot \nabla \log \frac{\rho_t}{\pi} = -\int \rho_t (\nabla V + \frac{\nabla \rho_t}{\rho_t}) \cdot \nabla \log \frac{\rho_t}{\pi}$$

$$= -\int \rho_t (-\nabla \log \pi + \nabla \log \rho_t) \cdot \nabla \log \frac{\rho_t}{\pi} = -\int \rho_t \left\| \nabla \log \frac{\rho_t}{\pi} \right\|^2 = -J_\pi(\rho_t)$$

If π satisfies the log-Sobolev inequality, we have $-J_\pi(\rho_t) \leq -2\alpha H_\pi(\rho_t)$. Thus $\partial_t H_\pi(\rho_t) \leq -2\alpha H_\pi(\rho_t)$ and then we deduce

$$\frac{d}{dt}\left(e^{2\alpha t} H_\pi(\rho_t)\right) \leq 0, \text{ so } H_\pi(\rho_t) \leq e^{-2\alpha t} H_\pi(\rho_0)$$

Discrete scheme: assumptions

Recall the scheme

$$X_{(k+1)h} = X_{kh} - h\nabla V(X_{kh}) + \sqrt{2}(B_{(k+1)h} - B_{kh})$$

Target density π: assume

- the target density $\pi \propto \exp(-V)$ satisfies the log-Sobolev inequality with constant α.
- ∇V is Lipschitz with constant β.

Step size h: assume

- $h < \frac{\alpha}{4\beta^2}$
- Consequently: $h < \frac{1}{3\alpha}$ (since $\alpha \leq \beta$)

Discrete scheme: one-step contraction - goal

Consider a one-step interpolation, i.e. for $0 \leq t \leq h$

$$X_t = X_0 - t\nabla V(X_0) + \sqrt{2}B_t$$

Denote ρ_h as the distribution of X_h. Goal:

$$H_\pi(\rho_h) \leq e^{-\alpha h} H_\pi(\rho_0) + \frac{2}{3\alpha}(2hn\beta^2 + 2h^2 n\beta^3) \tag{4}$$

Afterwards, we iterate the estimate for k steps to get:

$$H_\pi(\rho_{kh}) \leq e^{-k\alpha h} H_\pi(\rho_0) + \frac{2}{3\alpha(1 - e^{-\alpha h})}(2hn\beta^2 + 2h^2 n\beta^3) \tag{5}$$

Discrete scheme: one-step contraction - Fokker-Planck

Similar calculations:

$$\mathbb{E}[f(X_t)] = \mathbb{E}[f(X_0)] - \mathbb{E}[\int \nabla f(X_s) \cdot \nabla V(X_0) ds] + \mathbb{E}[\int_0^t \Delta f(X_s) ds]$$

$$\frac{d}{dt}\mathbb{E}[f(X_t)] = -\mathbb{E}[\nabla f(X_t) \cdot \nabla V(X_0)] + \mathbb{E}[\Delta f(X_t)]$$

By def'n of conditional expectation: $\mathbb{E}[XY] = \mathbb{E}[X\mathbb{E}[Y|X]]$. Then

$$\frac{d}{dt}\mathbb{E}[f(X_t)] = -\mathbb{E}[\nabla f(X_t) \cdot \mathbb{E}[\nabla V(X_0)|X_t]] + \mathbb{E}[\Delta f(X_t)]$$

so ρ_t satisfies the following equation in a distributional sense:

$$\partial_t \rho_t = \nabla \cdot (\rho_t b(x,t)) + \Delta \rho_t$$

where $b(x,t) = \mathbb{E}[\nabla V(X_0)|X_t = x]$ is a vector field.

Discrete scheme: one-step contraction - differentiating KL

- Seek: the negative-definite term in continuous dynamics
- Add and subtract:

$$\partial_t \rho_t = \nabla \cdot (\rho_t \nabla V + \rho_t (b(x,t) - \nabla V)) + \Delta \rho_t$$

Similar calculations for $\partial_t H_\pi(\rho_t)$ (all integrals over $\mathbb{R}^n$ unless specified)

$$\begin{aligned}
\partial_t H_\pi(\rho_t) &= \int (\partial_t \rho_t) \log \frac{\rho_t}{\pi} \\
&= \int [\nabla \cdot (\rho_t \nabla V) + \Delta \rho_t] \log \frac{\rho_t}{\pi} + \int \nabla \cdot [\rho_t (b - \nabla V)] \log \frac{\rho_t}{\pi} \\
&= -J_\pi(\rho_t) - \int_{\mathbb{R}^n} \rho_t(x) [\mathbb{E}[\nabla V(X_0) | X_t = x] - \nabla V(x)] \cdot \nabla \log \frac{\rho_t(x)}{\pi(x)} dx \\
&= -J_\pi(\rho_t) + \int_{\mathbb{R}^n} \rho_t(x) [\nabla V(x) - \mathbb{E}[\nabla V(X_0) | X_t = x]] \cdot \nabla \log \frac{\rho_t(x)}{\pi(x)} dx
\end{aligned}$$

Discrete scheme: one-step contraction - bounding error I

- Error comes from time-discretization.
- Using $u \cdot v \le \|u\| \cdot \|v\| = 2\|u\| \cdot \frac{1}{2}\|v\| \le \|u\|^2 + \frac{1}{4}\|v\|^2$,

$$\begin{aligned}\partial_t H_\pi(\rho_t) &\le -J_\pi(\rho_t) + \frac{1}{4}\int \rho_t \left\| \nabla \log \frac{\rho_t}{\pi} \right\|^2 \\ &\qquad + \mathbb{E}\left[\left\| \nabla V(X_t) - \mathbb{E}[\nabla V(X_0)|X_t] \right\|^2 \right] \\ &= -\frac{3}{4} J_\pi(\rho_t) + \mathbb{E}\left[\left\| \nabla V(X_t) - \mathbb{E}[\nabla V(X_0)|X_t] \right\|^2 \right] \qquad (6)\end{aligned}$$

- $\nabla V(X_t)$ is X_t-measurable; apply conditional Jensen:

$$\begin{aligned}&\mathbb{E}\left[\|\nabla V(X_t) - \mathbb{E}[\nabla V(X_0)|X_t]\|^2\right] = \mathbb{E}\left[\|\mathbb{E}[\nabla V(X_t) - \nabla V(X_0)|X_t]\|^2\right] \\ &\le \mathbb{E}[\mathbb{E}[\|\nabla V(X_t) - \nabla V(X_0)\|^2|X_t]] = \mathbb{E}[\|\nabla V(X_t) - \nabla V(X_0)\|^2]\end{aligned}$$

Discrete scheme: one-step contraction - bounding error II

- Apply Lipschitz condition on ∇V, def'n of X_t, independence between X_0 and B_t:

$$\mathbb{E}[\|\nabla V(X_t) - \nabla V(X_0)\|^2] \leq \beta^2 \mathbb{E}[\|X_t - X_0\|^2]$$
$$= \beta^2 \mathbb{E}\left[\|-t\nabla V(X_0) + \sqrt{2}B_t\|^2\right] = \beta^2(t^2\mathbb{E}[\|\nabla V(X_0)\|^2] + 2\mathbb{E}[\|B_t\|^2])$$
$$= \beta^2 t^2 \mathbb{E}[\|\nabla V(X_0)\|^2] + 2tn\beta^2$$

- No info on distribution to X_0, so bridge to π: $\forall$ r.v. $Y \sim \pi$:

$$\begin{aligned}\mathbb{E}[\|\nabla V(X_0)\|^2] &= \mathbb{E}[\|\nabla V(Y) + \nabla V(X_0) - \nabla V(Y)\|^2] \\ &\leq 2\mathbb{E}[\|\nabla V(Y)\|^2] + 2\mathbb{E}[\|\nabla V(X_0) - \nabla V(Y)\|^2] \\ &\leq 2n\beta + \beta\mathbb{E}[\|X_0 - Y\|^2]\end{aligned}$$

- Choose Y with optimal coupling s.t. $\mathbb{E}[\|X_0 - Y\|^2] = W_2(\rho_0, \pi)^2$

Discrete scheme: one-step contraction - bounding error IV

- Otto-Villani: log-Sobolev imples Talagrand, $W_2(\rho_0, \pi)^2 \leq \frac{2}{\alpha} H_\pi(\rho_0)$;
- Error term estimate:

$$\mathbb{E}[\|\nabla V(X_t) - \nabla V(X_0)\|^2] \leq 2tn\beta^2 + 2t^2 n\beta^3 + 4\frac{\beta^4 t^2}{\alpha} H_\pi(\rho_0)$$

- In inequality 6, use log-Sobolev to relate $J_\pi(\rho_t)$ and $H_\pi(\rho_t)$:

$$\begin{aligned}\partial_t H_\pi(\rho_t) &\leq -\frac{3}{4} J_\pi(\rho_t) + 2tn\beta^2 + 2t^2 n\beta^3 + \frac{4t^2\beta^4}{\alpha} H_\pi(\rho_0) \\ &\leq -\frac{3\alpha}{2} H_\pi(\rho_t) + \frac{4t^2\beta^4}{\alpha} H_\pi(\rho_0) + 2tn\beta^2 + 2t^2 n\beta^3\end{aligned}$$

- Upper bound error by replacing t with h, integrate from $t = 0$ to h:

$$H_\pi(\rho_h) \leq \left[e^{-\frac{3\alpha h}{2}} + \frac{8h^2\beta^4}{3\alpha^2}(1 - e^{-\frac{3\alpha h}{2}})\right] H_\pi(\rho_0) + \frac{2}{3\alpha}(2hn\beta^2 + 2h^2 n\beta^3)$$

Discrete scheme: clearing constants and conclusion

- Goal: coefficient in front of $H_\pi(\rho_0)$ is less than 1
- Use: $h \leq \frac{\alpha}{4\beta^2} \implies \frac{8h^2\beta^4}{3\alpha^2} \leq \frac{8\beta^4}{3\alpha^2}\frac{\alpha^2}{16\beta^4} = \frac{1}{6}$,
  $1 - e^{-x} \leq x \implies 1 - e^{-\frac{3\alpha h}{2}} \leq \frac{3\alpha h}{2}$, and $h \leq \frac{1}{3\alpha} \implies \frac{3\alpha h}{2} \leq \frac{1}{2}$
- Overall estimate:

$$
\begin{aligned}
&e^{-\frac{3\alpha h}{2}} + \frac{8h^2\beta^4}{3\alpha^2}(1 - e^{-\frac{3\alpha h}{2}}) \\
&\quad\leq e^{-\frac{3\alpha h}{2}} + \frac{1}{6}\frac{3\alpha h}{2} = e^{-\frac{3\alpha h}{2}} + \frac{1}{4}\alpha h \\
&\quad= e^{-\frac{3\alpha h}{2}}(1 + \frac{1}{4}\alpha h e^{\frac{3\alpha h}{2}}) \leq e^{-\frac{3\alpha h}{2}}(1 + \frac{1}{4}\alpha h e^{\frac{1}{2}}) \\
&\quad\leq e^{-\frac{3\alpha h}{2}}(1 + \frac{1}{2}\alpha h) \leq e^{-\frac{3\alpha h}{2}} e^{\frac{\alpha h}{2}} = e^{-\alpha h}
\end{aligned}
$$

- Hence we have proved the one-step estimate

$$H_\pi(\rho_h) \leq e^{-\alpha h} H_\pi(\rho_0) + \frac{2}{3\alpha}(2hn\beta^2 + 2h^2 n\beta^3)$$

Convergence of marginal distributions: setup and notations

- $X_t \in \mathbb{R}^n$. Denote the set of indices as $[n] = \{1, 2, ..., n\}$. We use $u \subset [n]$ to refer to a subset of dimensions.
- X_t^u means the dimensions in u of variable X_t, ρ_t^u refers to the marginal distribution
- Given $\pi \propto e^{-V}$, form a graph whose vertices are indices $[n]$. Define $i \sim j$ if $\partial_{ij} V$ is not identically zero.
- $\nabla_u f$ means the gradient of f in the directions in u
- $N(u) = \cup_{i \in u} N(i)$ is the neighborhood of u. Note that $\nabla_u V$ only depends on the coordinates in $N(u)$

Recall the scheme:

$$X_{(k+1)h} = X_{kh} - h\nabla V(X_{kh}) + \sqrt{2}(B_{(k+1)h} - B_{kh})$$

Notation: for $u \subset [n]$, denote

$$H_t(u) := H_{\pi^u}(\rho_t^u)$$

Assumptions

On target distribution π:

- $\pi \sim \exp(-V)$ satisfies the log-Sobolev inequality with constant α.
- the potential V is such that ∇V is Lipschitz with constant β.
- there exists $\gamma > 0$ such that for any $u \subset \{1, 2, ..., n\}, x^u \in \mathbb{R}^u$, and any probability measures μ on $\mathbb{R}^u$, we have

$$\|\mathbb{E}_\mu[\nabla_u V(\cdot, x^u)] - \mathbb{E}_{Y\sim\pi}[\nabla_u V(Y)|Y^u = x^u]\|^2 \leq \frac{2\gamma\beta^2}{\alpha} H_{\pi_{x^u}^{N(u)\setminus u|u}}(\mu)$$

Here, the notation $\pi_{x^u}^{N(u)\setminus u|u}$ means the conditional distribution of $N(u)$ coordinates under an overall joint distribution π, given that the u coordinates are x^u.

On initial condition:

- there exists $C_0 \geq 0$ such that $H_0(u) \leq C_0|u|$ for all $u \subset [n]$.

Conditions on step size h depends on growth conditions on the size of neighborhoods $N_k(u)$. Will specify later.

Deriving Fokker-Planck: I

Suppose $dX_t = -\nabla V(X_0) + \sqrt{2}dB_t$ for $0 \leq t \leq h$. Expand via Ito:

$$\phi(X_t^u) = \phi(X_0^u) + \sum_{i \in u} \int_0^t \partial_i \phi(X_s^u) dX_s^{(i)} + \frac{1}{2} \sum_{i,j \in u} \int_0^t \partial_{ij} \phi(X_s^u) d\left\langle X^{(i)}, X^{(j)} \right\rangle_s$$

Using the SDE for dX_s, one term vanishes in expectation, so after time differentiation:

$$\frac{d}{dt}\mathbb{E}[\phi(X_t^u)] = -\mathbb{E}[\nabla_u \phi(X_t^u) \cdot \nabla_u V(X_0)] + \mathbb{E}[\Delta_u \phi(X_t^u)]$$

Make everything dependent on ρ_t^u using conditional expectation:

$$\frac{d}{dt}\mathbb{E}[\phi(X_t^u)] = -\mathbb{E}[\nabla_u \phi(X_t^u) \cdot \mathbb{E}[\nabla_u V(X_0) | X_t^u]] + \mathbb{E}[\Delta_u \phi(X_t^u)]$$

This is the distributional form of the Fokker-Planck equation

$$\partial_t \rho_t^u = \nabla_u \cdot (\rho_t^u b^u(x^u, t)) + \Delta_u \rho_t^u, \text{ where } b^u(x^u, t) = \mathbb{E}[\nabla_u V(X_0) | X_t^u = x^u]$$

Deriving Fokker-Planck: II

Closer view:

$$\partial_t \rho_t^u = \nabla_u \cdot (\rho_t^u b^u(x^u, t)) + \Delta_u \rho_t^u \tag{7}$$

$$b^u(x^u, t) = \mathbb{E}[\nabla_u V(X_0) | X_t^u = x^u]$$

Recall:

- coefficient in full-dimensions continuum dynamics - ∇V.
- $\nabla_u(x)$ depends on coordinates in $N(u)$.

Two sources of error:

- time discretization - inferring info on X_0 based on X_t;
- spacial localization - inferring info on $N(u)$ based on X_t^u.

Error terms: I

Differentiate KL divergence (formally):

$$
\begin{aligned}
\frac{d}{dt}\int \rho_t^u \log\frac{\rho_t^u}{\pi^u} &= \frac{d}{dt}\int \rho_t^u \log \rho_t^u - \frac{d}{dt}\int \rho_t^u \log \pi^u \\
&= \int (\partial_t \rho_t^u)\log\frac{\rho_t^u}{\pi^u} = \int \nabla_u \cdot (\rho_t^u b^u(x^u,t) + \nabla_u \rho_t^u)\log\frac{\rho_t^u}{\pi^u} \\
&= -\int (\rho_t^u b^u(x^u,t) + \nabla_u \rho_t^u)\cdot \nabla_u \log\frac{\rho_t^u}{\pi^u}
\end{aligned}
$$

- Q: what coefficient multiplied to ρ_t^u can give us the term $\rho_t^u \|\nabla_u \log\frac{\rho_t^u}{\pi^u}\|^2$?
- A: $-\nabla_u \log \pi^u$. (Note: it is *not* $\nabla_u V$)

Error terms: II

Add and subtract to reveal the negative definite term:

$$\begin{aligned}
&\frac{d}{dt}\int \rho_t^u \log \frac{\rho_t^u}{\pi^u} \\
&= -\int (-\rho_t^u \nabla_u \log \pi^u + \nabla_u \rho_t^u) \cdot \nabla_u \log \frac{\rho_t^u}{\pi^u} \\
&\quad + \rho_t^u (b^u(x^u, t) + \nabla_u \log \pi^u) \cdot \nabla_u \log \frac{\rho_t^u}{\pi^u} \\
&= -\int \rho_t^u \left\| \nabla_u \log \frac{\rho_t^u}{\pi^u} \right\|^2 + \rho_t^u (b^u(x^u, t) + \nabla_u \log \pi^u) \cdot \nabla_u \log \frac{\rho_t^u}{\pi^u}
\end{aligned}$$

- The first term we keep, the second term we bound.
- Bounding the 2nd term should eat away some of the constant in the first term and leave us with $\mathbb{E}\left[\|b^u(X_t^u, t) + \nabla_u \log \pi^u(X_t^u)\|^2\right]$.
- Could do this by either $u \cdot v \leq \|u\|^2 + \frac{1}{4}\|v\|^2$ (Vempala-Wibosono) or by completing the square (Lacker-Zhou).

Error terms: III

Arrive at the bound: (recall ρ_t^u is the distribution of X_t^u)

$$\partial_t H_{\pi^u}(\rho_t^u) \leq -\frac{1}{2} J_{\pi^u}(\rho_t^u) + \frac{1}{2}\mathbb{E}\left[\|b^u(X_t^u, t) + \nabla_u \log \pi^u(X_t^u)\|^2\right] \qquad (8)$$

Study the error term:

- first, $b^u(X_t^u, t) = \mathbb{E}[\nabla_u V(X_0)|X_t^u]$;
- whereas:

$$\begin{aligned}\nabla_u \log \pi^u(x^u) &= \frac{1}{\pi^u(x^u)}\nabla_u \pi^u(x^u) = \frac{1}{\pi^u(x^u)}\nabla_u \int \pi(\overline{x}, x^u) d\overline{x} \\ &= \frac{1}{\pi^u(x^u)}\nabla_u \int e^{-V(\overline{x}, x^u)} d\overline{x} \\ &= -\int \nabla_u V(\overline{x}, x^u)\frac{e^{-V(\overline{x}, x^u)}}{\pi^u(x^u)} d\overline{x} \\ &= -\mathbb{E}_{Y\sim\pi}[\nabla_u V(Y)|Y^u = x^u]\end{aligned}$$

Error terms: IV

Idea: separate error into two terms, one for time discretization, one for what's left.
For any $\epsilon \in (0,1)$, $(x+y)^2 \leq \frac{1}{\epsilon}x^2 + \frac{1}{1-\epsilon}y^2$, so:

$$
\begin{aligned}
&\mathbb{E}[\|b^u(X_t^u, t) + \nabla_u \log \pi^u(X_t^u)\|^2] \\
=&\mathbb{E}[\|b^u(X_t^u, t) - \mathbb{E}[\nabla_u V(X_t)|X_t^u] + \mathbb{E}[\nabla_u V(X_t)|X_t^u] + \nabla_u \log \pi^u(X_t^u)\|^2] \\
\leq&\frac{1}{1-\epsilon}\mathbb{E}[\|b^u(X_t^u, t) - \mathbb{E}[\nabla_u V(X_t)|X_t^u]\|^2] \qquad \ldots \text{"continuity term"} \\
&+ \frac{1}{\epsilon}\mathbb{E}[\|\mathbb{E}[\nabla_u V(X_t)|X_t^u] + \nabla_u \log \pi^u(X_t^u)\|^2] \quad \ldots \text{"hierarchy term"} \\
\leq&\frac{1}{1-\epsilon}\mathbb{E}[\|\nabla_u V(X_0) - \nabla_u V(X_t)\|^2] \\
&+ \frac{1}{\epsilon}\mathbb{E}[\|\mathbb{E}[\nabla_u V(X_t)|X_t^u] - \mathbb{E}_{Y\sim\pi}[\nabla_u V(Y)|Y^u = X_t^u]\|^2]
\end{aligned}
$$

Error terms: conclusion

- Supposing π satisfies the log-Sobolev inequality, its marginal π^u also obeys the log-Sobolev inequality with the same constant - switches J back to H.
- Thus for any $\epsilon \in (0, 1)$ we have

$$\partial_t H_{\pi^u}(\rho_t^u) \le -\alpha H_{\pi^u}(\rho_t^u) + A_t^1(u) + A_t^2(u) \tag{9}$$

where

$$\begin{aligned} A_t^1(u) =& \frac{1}{2(1-\epsilon)} \mathbb{E}[\|\nabla_u V(X_0) - \nabla_u V(X_t)\|^2] \\ A_t^2(u) =& \frac{1}{2\epsilon} \int \|\mathbb{E}[\nabla_u V(X_t) | X_t^u = x^u] \\ & \qquad - \mathbb{E}_{Y\sim\pi}[\nabla_u V(Y) | Y^u = x^u]\|^2 \rho_t^u(dx^u) \end{aligned}$$

Continuity term: I

- $\nabla_u V$ is a β-Lipschitz function of coordinates in $N(u)$.
- Denote $w = N(u)$. Then: ($|w|$ means no. of coordinates in w)

$$\begin{aligned}\mathbb{E}[\|\nabla_u V(X_0) - \nabla_u V(X_t)\|^2] &\leq \beta^2 \mathbb{E}[\|X_t^w - X_0^w\|^2] \\ &= \beta^2 \mathbb{E}[\|-t\nabla_w V(X_0) + \sqrt{2} B_t^w\|^2] \\ &= \beta^2 t^2 \mathbb{E}[\|\nabla_w V(X_0)\|^2] + 2\beta^2 t|w|\end{aligned}$$

Treat $\mathbb{E}[\|\nabla_w V(X_0)\|^2]$: try the Vempala-Wibisono argument via coupling

$$\begin{aligned}\mathbb{E}[\|\nabla_w V(X_0)\|^2] &\leq 2\mathbb{E}[\|\nabla_w V(X_0) - \nabla_w V(Y)\|^2] + 2\mathbb{E}[\|\nabla_w V(Y)\|^2] \\ &\leq 2\beta^2 \mathbb{E}[\|X_0^{N(w)} - Y^{N(w)}\|^2] + 2\mathbb{E}[\|\nabla_w V(Y)\|^2]\end{aligned}$$

- To bound 1st term: vary Y over all random variables with $N(w)$ marginal $\pi^{N(w)}$, select optimal coupling
- To bound 2nd term: sample the remaining coordinates of Y by conditioning π on the $N(w)$ coordinates.

Then the 1st term becomes Wasserstein distance, and the 2nd term is bounded using the generator. There is another approach though. . .

Continuity term: II

Lacker-Zhou: appeals to Kantorovich duality

- For any continuous functions f, g such that $f(x^{N(w)}) - g(y^{N(w)}) \leq \|x^{N(w)} - y^{N(w)}\|^2$ for $\rho_0^{N(w)}$-almost every $x^{N(w)}$ and $\pi^{N(w)}$-almost every $y^{N(w)}$, one has

$$\int f d\rho_0^{N(w)} - \int g d\pi^{N(w)} \leq W_2^2(\rho_0^{N(w)}, \pi^{N(w)})$$

- Choose $f(x^{N(w)}) = \frac{1}{2\beta^2}\|\nabla_w V(x^{N(w)})\|^2$, and $g(y^{N(w)}) = \frac{1}{\beta^2}\|\nabla_w V(y^{N(w)})\|^2$.
- Pointwise, $\|\nabla_w V(x^{N(w)})\|^2 \leq 2\beta^2\|x^{N(w)} - y^{N(w)}\|^2 + 2\|\nabla_w V(y^{N(w)})\|^2$
- Therefore $f(x^{N(w)}) - g(y^{N(w)}) \leq \|x^{N(w)} - y^{N(w)}\|^2$

Continuity term: III

Note that since $\nabla_w V$ only depends on $N(w)$ variables,

$$\int_{\mathbb{R}^{N(w)}} \nabla_w V(x^{N(w)}) d\rho_0^{N(w)} = \mathbb{E}[\|\nabla_w V(X_0)\|^2]$$

Apply Kantorovich duality:

$$\frac{1}{2\beta^2}\mathbb{E}[\|\nabla_w V(X_0)\|^2] - \frac{1}{\beta^2}\mathbb{E}_\pi[\|\nabla_w V(Y)\|^2] \leq W_2^2(\rho_0^{N(w)}, \pi^{N(w)})$$

Rearrange

$$\mathbb{E}[\|\nabla_w V(X_0)\|^2] \leq 2\beta^2 W_2^2(\rho_0^{N(w)}, \pi^{N(w)}) + 2\mathbb{E}_\pi[\|\nabla_w V(Y)\|^2]$$

The marginals satisfy the Talagrand inequality. Thus

$$\mathbb{E}[\|\nabla_w V(X_0)\|^2] \leq \frac{4\beta^2}{\alpha} H_{\pi^{N(w)}}(\rho_0^{N(w)}) + 2\mathbb{E}_\pi[\|\nabla_w V(Y)\|^2]$$

Continuity term: IV

Now treat $\mathbb{E}_\pi[\|\nabla_w V(Y)\|^2]$

$$\mathbb{E}_\pi[\|\nabla_w V(Y)\|^2] = \int \nabla_w V(y) \cdot \nabla_w V(y) e^{-V(y)} dy$$

$$= -\int \nabla_w V(y) \cdot \nabla_w \pi(y) dy = \int \Delta_w V(y)\pi(y) dy = \mathbb{E}_\pi[\Delta_w V(Y)] \leq \beta |w|$$

Putting everything together,

$$A_t^1(u) \leq \frac{1}{2(1-\epsilon)}\left[\beta^2 t^2 \left(\frac{4\beta^2}{\alpha} H_{\pi^{N(w)}}(\rho_0^{N(w)}) + 2\beta|w|\right) + 2\beta^2 t|w|\right]$$
$$= \frac{2\beta^4 t^2}{\alpha(1-\epsilon)} H_{\pi^{N(w)}}(\rho_0^{N(w)}) + \frac{\beta^2 t(\beta t + 1)}{1-\epsilon}|w|$$

We might introduce the notation $H_t(u) = H_{\pi^u}(\rho_t^u)$ and bound $t \leq h$ to rewrite the estimate as

$$A_t^1(u) \leq \frac{2\beta^4 h^2}{\alpha(1-\epsilon)} H_0(N_2(u)) + \frac{\beta^2 h(\beta h + 1)}{1-\epsilon}|N(u)|$$

Hierarchy term

Now we treat A_t^2. Recall its definition:

$$A_t^2(u) = \frac{1}{2\epsilon}\int \|\mathbb{E}[\nabla_u V(X_t)|X_t^u = x^u] - \mathbb{E}_{Y\sim\pi}[\nabla_u V(Y)|Y^u = x^u]\|^2 \rho_t^u(dx^u)$$

We specifically had an assumption for this: assume there exists $\gamma > 0$ such that for any $u \subset \{1, 2, ..., n\}, x^u \in \mathbb{R}^u$, and any probability measures μ on $\mathbb{R}^u$, we have

$$\|\mathbb{E}_\mu[\nabla_u V(\cdot, x^u)] - \mathbb{E}_{Y\sim\pi}[\nabla_u V(Y)|Y^u = x^u]\|^2 \le \frac{2\gamma\beta^2}{\alpha} H_{\pi_{x^u}^{N(u)\setminus u|u}}(\mu)$$

Therefore

$$A_t^2(u) \le \frac{\gamma\beta^2}{\alpha\epsilon}\int_{\mathbb{R}^u} H_{\pi_{x^u}^{N(u)\setminus u|u}}(\rho_{t,x^u}^{N(u)\setminus u|u}) d\rho_t^u(x^u)$$

Bounding individual terms: conclusion

Apply the chain rule for KL divergence: Suppose we have coordinates x, y, and ν, μ joint densities on (x, y). Then

$$H_\nu(\mu) = H_{\nu_x}(\mu_x) + \int H_{\nu_{y|x}}(\mu_{y|x}) d\mu_x(x)$$

where the subscripted measures are marginals and conditional measures respectively.

Thus

$$A_t^2(u) \leq \frac{\gamma\beta^2}{\alpha\epsilon}(H_t(N(u)) - H_t(u))$$

and we have

$$\partial_t H_t(u) \leq -\alpha H_t(u) + \frac{2\beta^4 h^2}{\alpha(1-\epsilon)} H_0(N_2(u)) + \frac{\beta^2 h(\beta h + 1)}{1-\epsilon}|N(u)| + \frac{\gamma\beta^2}{\alpha\epsilon}(H_t(N(u)) - H_t(u)) \quad (10)$$

One-step estimate: setup

- Goal: integrate inequality from 0 to h.
- Separate time-dependent terms from time-independent ones.
- Define $C_0(u) := \frac{2\beta^4 h^2}{\alpha(1-\epsilon)} H_0(N_2(u)) + \frac{\beta^2 h(\beta h+1)}{1-\epsilon} |N(u)|$ (time-independent)
- Suppose F is a function that takes a subset $u \subset [n]$ and returns a number. Define operator A by

$$A[F](u) = \frac{\gamma\beta^2}{\alpha\epsilon}[F(N(u)) - F(u)]$$

- Summarize inequality 10 as

$$\partial_t H_t(u) \leq (A - \alpha) H_t(u) + C_0(u) \tag{11}$$

One-step estimate: operator A as generator

Define $\Lambda(t)$ as a Poisson process with rate $\lambda := \frac{\gamma\beta^2}{\alpha\epsilon}$ starting at 0. Then, consider the subset-valued jump process $\mathcal{X}_t = N_{\Lambda(t)}(\mathcal{X}_0), \mathcal{X}_t \subset [n]$. Let us calculate the generator of this process.

$$
\begin{aligned}
\mathcal{L}f(u) &= \lim_{t\downarrow 0} \frac{1}{t}[\mathbb{E}[f(N_{\Lambda(t)}(u))] - f(u)] \\
&= \lim_{t\to 0} \frac{1}{t}\left[-f(u) + \sum_{k=0}^{\infty} e^{-\lambda t}\frac{(\lambda t)^k}{k!} f(N_k(u))\right] \\
&= \lim_{t\to 0} \frac{1}{t}\left[-f(u) + e^{-\lambda t} f(u) + e^{-\lambda t}(\lambda t) f(N(u)) + o(t)\right] \\
&= \lim_{t\to 0} \frac{1}{t}\left[(e^{-\lambda t} - 1) f(u) + e^{-\lambda t}(\lambda t) f(N(u)) + o(t)\right] = \lambda[f(N(u)) - f(u)]
\end{aligned}
$$

Therefore, A is the generator for the process $\mathcal{X}_t$, which means:

- $e^{tA}f(u) = \mathbb{E}[f(\mathcal{X}_t)|\mathcal{X}_0 = u]$;
- as a consequence, if $f \leq g$, $e^{tA}f \leq e^{tA}g$;
- if $f(u) \leq f(N(u))$, then the map $t \mapsto e^{tA}f$ is non-decreasing in t.

One-step estimate: integrate in time

- Via Duhamel's principle, integrate from 0 to h

$$H_h(u) \leq e^{(A-\alpha)h} H_0(u) + \int_0^h e^{(A-\alpha)(h-s)} C_0(u) ds$$

- **Claim**: $C_0(u) \leq C_0(N(u))$. Indeed, $|N(u)| \leq |N(N(u))|$, and by chain rule of KL divergence, the difference $H_0(N(u)) - H_0(u)$ is an integral of a nonnegative integrand against a conditional measure.
- Bound $e^{(h-s)A} C_0(u) \leq e^{hA}(u)$ for all $s \in (0, h)$. Then

$$H_h(u) \leq e^{-\alpha h} e^{hA} H_0(u) + e^{-\alpha h} \int_0^h e^{\alpha s} e^{hA} C_0(u) ds$$

and simplifying

$$H_h(u) \leq e^{-\alpha h} e^{hA} H_0(u) + \frac{1}{\alpha}(1 - e^{-\alpha h}) e^{hA} C_0(u) \qquad (12)$$

One-step estimate: so far

We have shown:

$$H_h(u) \leq e^{-\alpha h} e^{hA} H_0(u) + \frac{1}{\alpha}(1 - e^{-\alpha h}) e^{hA} C_0(u)$$

Substituting $C_0(u) := \frac{2\beta^4 h^2}{\alpha(1-\epsilon)} H_0(N_2(u)) + \frac{\beta^2 h(\beta h+1)}{1-\epsilon} |N(u)|$ gives

$$\begin{aligned} H_h(u) \leq\ & e^{-\alpha h} e^{hA} H_0(u) + \\ & \frac{1}{\alpha}(1 - e^{-\alpha h}) e^{hA} \left[\frac{2\beta^4 h^2}{\alpha(1-\epsilon)} H_0(N_2(u)) + \frac{\beta^2 h(\beta h + 1)}{1-\epsilon} |N(u)| \right] \end{aligned}$$

Next goal: find the operation acted on $H_0(u)$ to seek contraction.

One-step estimate: regroup terms

Goal: find the operation acted on $H_0(u)$ to seek contraction.

- Define a "neighboring" operator $\mathcal{N}$: $\mathcal{N}[f](u) = f(N(u))$
- Things that are not H_0: define $G(u) = \frac{\beta^2 h(\beta h+1)}{1-\epsilon}|u|$, then

$$H_h(u) \leq \left[e^{-\alpha h}e^{hA} + \frac{1}{\alpha}(1-e^{-\alpha h})e^{hA}\frac{2\beta^4 h^2}{\alpha(1-\epsilon)}\mathcal{N}^2\right]H_0(u) + \frac{1}{\alpha}(1-e^{-\alpha h})e^{hA}\mathcal{N}G(u)$$

Helpful observations: $\mathcal{N}$ and A are linear, and they commute:

$$A\mathcal{N}[f](u) = Af(N(u)) = \lambda[f(N_2(u)) - f(N(u))]$$
$$\mathcal{N}A[f](u) = \mathcal{N}\lambda[f(N(u)) - f(u)] = \lambda[f(N_2(u)) - f(N(u))]$$

Therefore, after simplification

$$H_h(u) \leq \left[e^{-\alpha h} + \frac{2\beta^4 h^2(1-e^{-\alpha h})}{\alpha^2(1-\epsilon)}\mathcal{N}^2\right]e^{hA}H_0(u) + \frac{1-e^{-\alpha h}}{\alpha}e^{hA}\mathcal{N}G(u) \tag{13}$$

Case of polynomial sparsity: setup and strategy

Setup:

- For some $c \geq 1, p \geq 1$, $\max_{i\in[n]} |N_{k+1}(i)| \leq c(1+k^p), \forall k \geq 0$.
- Assumed: there exists $C_0 \geq 0$ s.t. $H_0(u) \leq C_0|u|$ for all $u \subset [n]$.
- The monotonicity of e^{tA} and $\mathcal{N}$ allows us to replace $H_0(u)$ with $|u|$ in the estimate.

Thus we study how e^{tA} and $\mathcal{N}$ act on size function $S(u) := |u|$. We will take $\epsilon = \frac{1}{2}$.
Note that because A and $\mathcal{N}$ commute, if we iterate the one-step estimate for k steps, we encounter e^{khA} instead of e^{hA}, and the other operation

$$\left[e^{-\alpha h} + \frac{2\beta^4 h^2(1-e^{-\alpha h})}{\alpha^2(1-\epsilon)}\mathcal{N}^2\right]$$

is just repeated.

Case of polynomial sparsity: action of N and A

First, $|N_{k+1}(u)| \leq \sum_{i \in u} |N_{k+1}(i)| \leq c(1+k^p)|u|$, so $\mathcal{N} S \leq cS$.
Next, for e^{hA}: (using $c \geq 1$)

$$e^{hA}S(u) = \mathbb{E}[S(\mathcal{X}_h)|\mathcal{X}_0 = u] = \mathbb{E}[|N_{\Lambda(h)}(u)|] = \sum_{l=0}^{\infty} |N_l(u)|\mathbb{P}(\Lambda(h) = l)$$

$$\leq |u|\mathbb{P}(\Lambda(h) = 0) + \sum_{l=1}^{\infty} c(1 + (l-1)^p)|u|\mathbb{P}(\Lambda(h) = l)$$

$$\leq c(1 + 0^p)|u|\mathbb{P}(\Lambda(h) = 0) + \sum_{l=1}^{\infty} c(1 + l^p)|u|\mathbb{P}(\Lambda(h) = l) = c|u|\mathbb{E}[1 + \Lambda(h)^p]$$

Poisson moment bound: $\mathbb{E}[\Lambda(h)^p] \leq (\lambda h + p)^p$.
Thus we have

$$e^{hA}S(u) \leq c(1 + (\lambda h + p)^p)|u| = c(1 + (\lambda h + p)^p)S(u)$$

Case of polynomial sparsity: towards multi-step estimate

Observation:

- $\mathcal{N}$ - multiplication by a constant per application
- e^{tA} - size grows in polynomial speed as propagated forward
- Therefore, directly looking at the multi-step estimate, we only need the base in front of $e^{khA}H_0$ to be less than 1 to beat the polynomial.

Using monotonicity and commutativity of e^{tA} and $\mathcal{N}$, (taking $\epsilon = \frac{1}{2}$)

$$
\begin{aligned}
H_{kh}(u) \le & \left[e^{-\alpha h} + \frac{4\beta^4 h^2(1-e^{-\alpha h})}{\alpha^2}\mathcal{N}^2\right]^k e^{khA}H_0(u) \\
& + \frac{1-e^{-\alpha h}}{\alpha}\sum_{j=0}^{k-1}\left[e^{-\alpha h} + \frac{4\beta^4 h^2(1-e^{-\alpha h})}{\alpha^2}\mathcal{N}^2\right]^j e^{(j+1)hA}\mathcal{N}G(u)
\end{aligned} \tag{14}
$$

We need $e^{-\alpha h} + \frac{4\beta^4 h^2}{\alpha^2}(1-e^{-\alpha h})c^2 < 1$

Case of polynomial sparsity: conclusion

The constant is essentially $e^{-\alpha h} + O(h^2)(1 - e^{-\alpha h})$.
Assume $h \le h^* := \frac{\alpha}{4c\beta^2}$. If $h \le h^*$, we have

$$e^{-\alpha h} + \frac{4\beta^4 h^2}{\alpha^2}(1 - e^{-\alpha h})c^2 \le e^{-\alpha h} + \frac{1}{4}(1 - e^{-\alpha h}) = 1 - \frac{3}{4}(1 - e^{-\alpha h})$$

Clearly the constant is less than 1, but let's clean up:

- since $\alpha \le \beta, c \ge 1$, we have $\alpha h \le \frac{\alpha^2}{4c\beta^2} \le \frac{1}{4}$
- $1 - e^{-x} \ge e^{-\frac{1}{4}}x$ for all $x \le \frac{1}{4}$, because at $x = 0$ they coincide, and $\frac{d}{dx}(1 - e^{-x}) = e^{-x} \ge e^{-\frac{1}{4}}$ for all $x \le \frac{1}{4}$

Thus

$$1 - \frac{3}{4}(1 - e^{-\alpha h}) \le 1 - \frac{3}{4}e^{-\frac{1}{4}}\alpha h \le 1 - \frac{1}{2}\alpha h \le e^{-\frac{\alpha h}{2}}$$

where we used $\frac{3}{4}e^{-\frac{1}{4}} > 0.5$ and $1 - x \le e^{-x}$.
This completes the proof of convergence. We also note that the bias term G depends on the dimension of u, not the whole system.

Case of exponential sparsity: overview

Setup:

- **Assume**: there exists $1 \leq r < 1 + \frac{\alpha^2}{\gamma\beta^2}$, $c \geq 1$ such that for each $k \geq 0$, any $i \in [n]$, $\max_i |N_{k+1}(i)| \leq cr^k$.
- As before $\mathcal{N}S \leq cS$
- $e^{tA}S(u) \leq c|u|\mathbb{E}[r^{\Lambda(t)}] \leq ce^{\lambda(r-1)t}S(u)$ - no more polynomial growth.

Overall bound:

$$\left[e^{-\alpha h} + \frac{2\beta^4 h^2(1-e^{-\alpha h})}{\alpha^2(1-\epsilon)}\mathcal{N}^2\right]^k e^{khA}H_0(u) \leq C_0\left[e^{-\alpha h} + \frac{2\beta^4 c^2 h^2(1-e^{-\alpha h})}{\alpha^2(1-\epsilon)}\right]^k ce^{\lambda kh(r-1)}|u|$$

Two degrees of freedom: ϵ and h. Make ϵ close to 1 and h small to obtain contraction and conclude.

Conclusion

Lacker-Zhou has another section on the case of "weak" interactions, which replaces the graphical model of potential V with another model where the form of V is more specific. There the analysis is similar in spirit but different in technical details, because in that case the operators $\mathcal{N}$ and A do not commute.
Thank you!