Wasserstein error for Langevin Sampling with Noisy Gradients

Yifei Xiong

Joint work with Prof. Vinayak Rao

Department of Statistics
Purdue University

https://arxiv.org/abs/2610.37858
https://yifei-xiong.github.io/

About this talk

- For this graduate student workshop, I will spend more time on **background and motivation**.
- We will focus on the main idea behind the improved error bound.
- Full proofs and technical details are in the paper.
  Y. Xiong and V. Rao (2026). Wasserstein bounds for Euler and randomized midpoint Langevin sampling with noisy gradients. *arXiv:2610.37858*.

Langevin dynamics

We want samples from a density π on $\mathbb{R}^p$. Write $\pi(x) \propto \exp(-U(x))$, so U is the negative log-density up to a constant.

Langevin dynamics preserves π. With p-dimensional Brownian motion B_t,

$$dX_t = \underbrace{-\nabla U(X_t)\, dt}_{\text{pull toward lower } U} + \underbrace{\sqrt{2}\, dB_t}_{\text{spread the samples}} .$$

Evaluating $\nabla U = -\nabla \log \pi$ requires no normalizing constant of π.

Euler update (approximate simulation)

Over each step of length $\alpha > 0$, Euler holds the gradient fixed at its starting value. With a fresh $Z_{k+1} \sim \mathcal{N}(0, I_p)$,

$$X_{k+1} = X_k - \alpha \nabla U(X_k) + \sqrt{2\alpha}\, Z_{k+1}.$$

Langevin dynamics

We want samples from a density π on $\mathbb{R}^p$. Write $\pi(x) \propto \exp(-U(x))$, so U is the negative log-density up to a constant.

Langevin dynamics preserves π. With p-dimensional Brownian motion B_t,

$$dX_t = \underbrace{-\nabla U(X_t)\, dt}_{\text{pull toward lower } U} + \underbrace{\sqrt{2}\, dB_t}_{\text{spread the samples}} .$$

Evaluating $\nabla U = -\nabla \log \pi$ requires no normalizing constant of π.

Euler update (approximate simulation)

Over each step of length $\alpha > 0$, Euler holds the gradient fixed at its starting value. With a fresh $Z_{k+1} \sim \mathcal{N}(0, I_p)$,

$$X_{k+1} = X_k - \alpha \nabla U(X_k) + \sqrt{2\alpha}\, Z_{k+1}.$$

Wasserstein distance

Write $\mathcal{L}(X_k)$ for the distribution of X_k. After k Euler steps, our sampling error is $W_2(\mathcal{L}(X_k), \pi)$.

Definition

For distributions μ, λ on $\mathbb{R}^p$ with finite second moments,

$$W_2(\mu, \lambda) := \inf_{\mathcal{L}(X)=\mu,\, \mathcal{L}(Y)=\lambda} \left(\mathbb{E}\|X - Y\|_2^2\right)^{1/2}.$$

- The infimum is over all joint distributions of (X, Y) with the given marginals, any such pair gives an upper bound.
- Think of moving mass from μ to λ. W_2 is the smallest root mean squared distance the mass travels.

Approximate gradients

In large-scale problems, evaluating $\nabla U(X_k)$ exactly can be expensive.

- **Minibatches.** In neural-network training, different batches can give different gradients at the same parameters.
- **Low precision.** Large language models are often trained with formats such as BF16. Rounding can bias the computed gradient.

Gradient noise model

$$\widehat{\nabla U}(X_k) - \nabla U(X_k) = \underbrace{b(X_k)}_{\text{bias}} + \underbrace{\xi(X_k)}_{\text{centered noise}}, \qquad \mathbb{E}[\xi(X_k) \mid X_k] = 0.$$

Approximate gradients

In large-scale problems, evaluating $\nabla U(X_k)$ exactly can be expensive.

- **Minibatches.** In neural-network training, different batches can give different gradients at the same parameters.
- **Low precision.** Large language models are often trained with formats such as BF16. Rounding can bias the computed gradient.

Gradient noise model

$$\widehat{\nabla U}(X_k) - \nabla U(X_k) = \underbrace{b(X_k)}_{\text{bias}} + \underbrace{\xi(X_k)}_{\text{centered noise}}, \qquad \mathbb{E}[\xi(X_k) \mid X_k] = 0.$$

Sampling error

Start both chains from the same distribution. Use the Euler update for X_k and the noisy Euler update for X_k^{N}.

$$W_2(\mathcal{L}(X_k^{\mathrm{N}}), \pi) \leq \underbrace{W_2(\mathcal{L}(X_k^{\mathrm{N}}), \mathcal{L}(X_k))}_{\text{noisy Euler to Euler}} + \underbrace{W_2(\mathcal{L}(X_k), \pi)}_{\text{Euler to target}}.$$

For the second term, Dalalyan and Karagulyan (2019), Pedrotti and Whalley (2026) give a finite-step Euler bounds.

Our focus is on the first term.

One Euler step

First consider zero gradient bias ($b = 0$). From the same fixed state X, the two updates are

$$X_{k+1} = X - \alpha \nabla U(X) + \sqrt{2\alpha}\, \widetilde{Z},$$
$$X_{k+1}^{\mathrm{N}} = X - \alpha \nabla U(X) - \alpha \xi(X) + \sqrt{2\alpha}\, Z.$$

Sharing Z pairs each noisy increment with its own Gaussian part. We must undo the full shift $-\alpha\xi(X)$, giving $W_2^2 \le \alpha^2 \mathbb{E}\|\xi(X)\|^2$.

Could another pairing move the mass a shorter distance?

One Euler step

First consider zero gradient bias ($b = 0$). From the same fixed state X, the two updates are

$$X_{k+1} = X - \alpha \nabla U(X) + \sqrt{2\alpha}\, \widetilde{Z},$$
$$X_{k+1}^{\mathrm{N}} = X - \alpha \nabla U(X) - \alpha \xi(X) + \sqrt{2\alpha}\, Z.$$

Sharing Z pairs each noisy increment with its own Gaussian part. We must undo the full shift $-\alpha\xi(X)$, giving $W_2^2 \le \alpha^2 \mathbb{E}\|\xi(X)\|^2$.

Could another pairing move the mass a shorter distance?

Gaussian gradient noise

In one dimension, let $\xi \sim \mathcal{N}(0, \sigma^2)$ be independent of Z. The increments are

$$\text{Noisy Euler update} \qquad \sqrt{2\alpha}\, Z - \alpha\xi \sim \mathcal{N}(0,\, 2\alpha + \alpha^2\sigma^2),$$

$$\text{Euler update} \qquad \sqrt{2\alpha}\, \widetilde{Z} \sim \mathcal{N}(0,\, 2\alpha).$$

For centered Gaussians, their W_2 distance is the difference between their standard deviations. For fixed $\sigma > 0$, as $\alpha \to 0$,

$$W_2^2 = \left(\sqrt{2\alpha + \alpha^2\sigma^2} - \sqrt{2\alpha}\right)^2 \sim \tfrac{\sigma^4}{8}\alpha^3.$$

(For α, larger exponents are better. Sharing Z has squared cost $\alpha^2\sigma^2$.)

Gaussian gradient noise

In one dimension, let $\xi \sim \mathcal{N}(0, \sigma^2)$ be independent of Z. The increments are

$$\text{Noisy Euler update} \qquad \sqrt{2\alpha}\, Z - \alpha\xi \sim \mathcal{N}(0,\, 2\alpha + \alpha^2\sigma^2),$$

$$\text{Euler update} \qquad \sqrt{2\alpha}\, \widetilde{Z} \sim \mathcal{N}(0,\, 2\alpha).$$

For centered Gaussians, their W_2 distance is the difference between their standard deviations. For fixed $\sigma > 0$, as $\alpha \to 0$,

$$W_2^2 = \left(\sqrt{2\alpha + \alpha^2\sigma^2} - \sqrt{2\alpha}\right)^2 \sim \tfrac{\sigma^4}{8}\alpha^3.$$

(For α, larger exponents are better. Sharing Z has squared cost $\alpha^2\sigma^2$.)

One-step coupling

At state X, the centered noise $\xi(X)$ is *sub-Gaussian with scale* K if, for every $u \in \mathbb{R}^p$,

$$\mathbb{E}\Big[\exp\Big(u^\top \xi(X)\Big) \mid X\Big] \le \exp\big(K^2\|u\|_2^2/2\big)\,.$$

The same K gives a Gaussian tail bound in every direction.

Lemma 1 (Xiong and Rao, 2026)

Fix X. If $\alpha K^2 \le 1$ and $Z \sim \mathcal{N}(0, I_p)$ is independent of $\xi(X)$, then

$$W_2^2\Big(\mathcal{L}(\sqrt{2\alpha}Z - \alpha\xi(X)), \mathcal{N}(0, 2\alpha I_p)\Big) \le \tfrac{2}{3}\alpha^3 K^4 p.$$

Proof idea. Talagrand's inequality bounds W_2^2 by a scaled KL divergence, which the sub-Gaussian assumption controls.

Euler after k steps

Let U be m-strongly convex with M-Lipschitz gradient. Assume $\alpha \le \min\{1/(m+M), 1/K^2\}$ and $\mathbb{E}\|b(X_k^{\mathrm{N}})\|^2 \le \delta^2 p$ at each step.

Noisy Euler and Euler

$$W_2(\mathcal{L}(X_k^{\mathrm{N}}), \mathcal{L}(X_k)) \le \frac{\delta\sqrt{p}}{m} + \alpha K^2 \sqrt{\frac{(m+M)p}{3mM}}.$$

Adding the Euler bound for $U \in C^2$ (Pedrotti and Whalley, 2026) gives

$$W_2(\mathcal{L}(X_k^{\mathrm{N}}), \pi) \le \underbrace{(1-m\alpha)^k W_2(\mathcal{L}(X_0^{\mathrm{N}}), \pi)}_{\text{initialization}}$$

$$+ \underbrace{12\alpha M\sqrt{p/m}}_{\text{discretization}} + \underbrace{\frac{\delta\sqrt{p}}{m}}_{\text{bias}} + \underbrace{\alpha K^2 \sqrt{\frac{(m+M)p}{3mM}}}_{\text{centered noise}}.$$

Euler after k steps

Let U be m-strongly convex with M-Lipschitz gradient. Assume $\alpha \le \min\{1/(m+M), 1/K^2\}$ and $\mathbb{E}\|b(X_k^{\mathrm{N}})\|^2 \le \delta^2 p$ at each step.

Noisy Euler and Euler

$$W_2\big(\mathcal{L}(X_k^{\mathrm{N}}), \mathcal{L}(X_k)\big) \le \frac{\delta\sqrt{p}}{m} + \alpha K^2 \sqrt{\frac{(m+M)p}{3mM}}.$$

Adding the Euler bound for $U \in C^2$ (Pedrotti and Whalley, 2026) gives

$$W_2\big(\mathcal{L}(X_k^{\mathrm{N}}), \pi\big) \le \underbrace{(1-m\alpha)^k W_2\big(\mathcal{L}(X_0^{\mathrm{N}}), \pi\big)}_{\text{initialization}}$$

$$+ \underbrace{12\alpha M\sqrt{p/m}}_{\text{discretization}} + \underbrace{\frac{\delta\sqrt{p}}{m}}_{\text{bias}} + \underbrace{\alpha K^2 \sqrt{\frac{(m+M)p}{3mM}}}_{\text{centered noise}}.$$

Euler bounds compared

Each column sums to a bound on $W_2\big(\mathcal{L}(X_k^{\mathrm{N}}), \pi\big)$.

	Dalalyan & Karagulyan $\mathbb{E}\|\xi\|^2 \le \sigma^2 p$	This work sub-Gaussian noise
Initialization	$(1-m\alpha)^k W_2(\mathcal{L}(X_0^{\mathrm{N}}), \pi)$	$(1-m\alpha)^k W_2(\mathcal{L}(X_0^{\mathrm{N}}), \pi)$
Discretization	$1.65\frac{M}{m}\sqrt{\alpha p}$	$12\alpha M\sqrt{p/m}$
Bias	$\delta\sqrt{p}/m$	$\delta\sqrt{p}/m$
Centered noise	$\dfrac{\sigma^2\sqrt{\alpha p}}{1.65M+\sigma\sqrt{m}}$	$\alpha K^2\sqrt{\dfrac{(m+M)p}{3mM}}$

Randomized midpoint (RM): another Langevin approximation

Over a time interval of length α, the diffusion follows

$$X_{t+\alpha} = X_t - \int_0^\alpha \nabla U(X_{t+s})\, ds + \sqrt{2}\,(B_{t+\alpha} - B_t).$$

Choose $\tau \sim \mathrm{Unif}(0,1)$. Along a fixed path, $\alpha \nabla U(X_{t+\alpha\tau})$ averages to the drift integral.

Two-stage update

First approximate $X_{t+\alpha\tau}$ by Y_k, then use its gradient for the full step.

$$Y_k = X_k - \alpha\tau \nabla U(X_k) + \sqrt{2\alpha\tau}\, Z',$$
$$X_{k+1} = X_k - \alpha \nabla U(Y_k) + \sqrt{2\alpha}\, Z.$$

The increments share Z' through $Z = \sqrt{\tau}\, Z' + \sqrt{1-\tau}\, Z''$. Z', Z'' are independent standard Gaussians, independent of τ.

Randomized midpoint (RM): another Langevin approximation

Over a time interval of length α, the diffusion follows

$$X_{t+\alpha} = X_t - \int_0^\alpha \nabla U(X_{t+s})\, ds + \sqrt{2}\,(B_{t+\alpha} - B_t).$$

Choose $\tau \sim \mathrm{Unif}(0,1)$. Along a fixed path, $\alpha \nabla U(X_{t+\alpha\tau})$ averages to the drift integral.

Two-stage update

First approximate $X_{t+\alpha\tau}$ by Y_k, then use its gradient for the full step.

$$Y_k = X_k - \alpha\tau \nabla U(X_k) + \sqrt{2\alpha\tau}\, Z',$$
$$X_{k+1} = X_k - \alpha \nabla U(Y_k) + \sqrt{2\alpha}\, Z.$$

The increments share Z' through $Z = \sqrt{\tau}\, Z' + \sqrt{1-\tau}\, Z''$. Z', Z'' are independent standard Gaussians, independent of τ.

Noisy randomized midpoint

The two stages use fresh gradient estimates at X_k^{N} and Y_k^{N}.

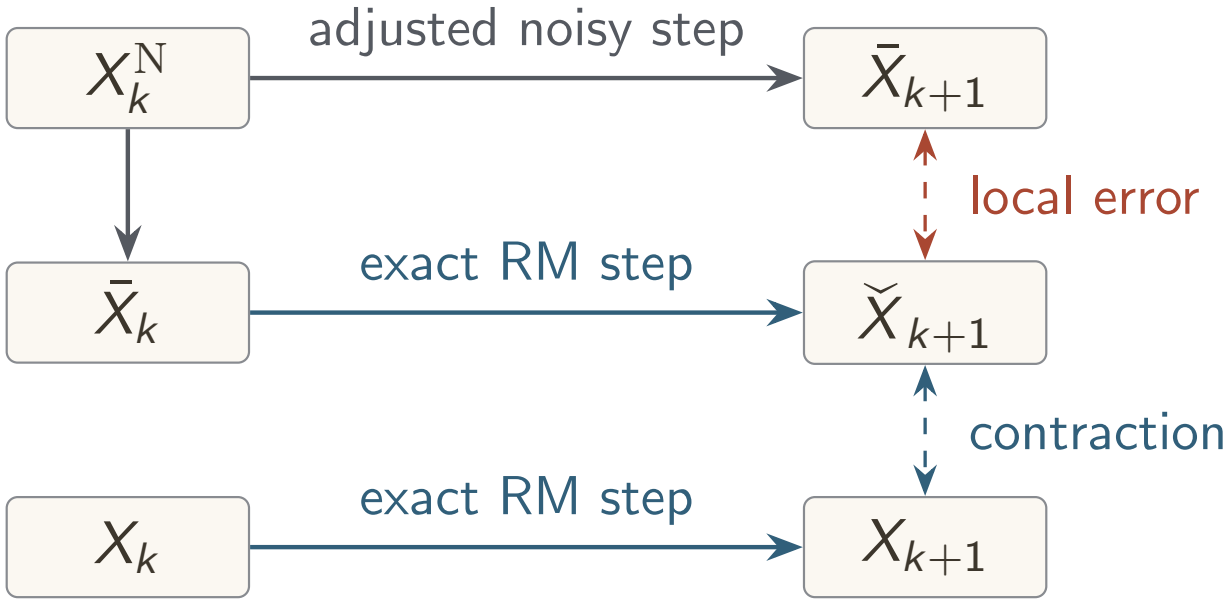
$$Y_k^{\mathrm{N}} = X_k^{\mathrm{N}} - \alpha\tau\widehat{\nabla U}(X_k^{\mathrm{N}}) + \sqrt{2\alpha\tau}\, Z',$$

$$X_{k+1}^{\mathrm{N}} = X_k^{\mathrm{N}} - \alpha\widehat{\nabla U}(Y_k^{\mathrm{N}}) + \sqrt{2\alpha}\, Z.$$

- The first gradient error changes where we evaluate the second gradient.
- Through that midpoint Y_k^{N}, the second error can depend on the Gaussian draw Z.

Intermediate RM states

$\bar{X}_k = X_k^{\mathrm{N}} + \alpha \xi_k^{(2)}$ (remove the previous centered stage-2 error)



Both exact steps share the random time and Gaussian draws.

RM sampling error

Start both chains from the same distribution. Take $\kappa = M/m$ and $\alpha \leq \min\{1/(8M\sqrt{\kappa}), 1/K^2\}$.

Noisy versus exact RM

$$W_2(\mathcal{L}(X_k^{\mathrm{N}}), \mathcal{L}(X_k)) \leq \alpha\sqrt{p}\left(10\kappa K + \frac{K^2}{\sqrt{m}}\right) + \frac{4\delta\sqrt{p}}{m}.$$

For twice-differentiable U, adding the exact-RM bound gives

$$W_2(\mathcal{L}(X_k^{\mathrm{N}}), \pi) \leq \underbrace{1.11\exp(-mk\alpha/2)\, W_2(\mathcal{L}(X_0^{\mathrm{N}}), \pi)}_{\text{initialization}}$$

$$+ \underbrace{(2.4\sqrt{\kappa M\alpha} + 1.77)M\alpha\sqrt{p/m}}_{\text{discretization}} + \underbrace{\frac{4\delta\sqrt{p}}{m}}_{\text{bias}} + \underbrace{\alpha\sqrt{p}\left(10\kappa K + \frac{K^2}{\sqrt{m}}\right)}_{\text{centered noise}}.$$

RM sampling error

Start both chains from the same distribution. Take $\kappa = M/m$ and $\alpha \leq \min\{1/(8M\sqrt{\kappa}), 1/K^2\}$.

Noisy versus exact RM

$$W_2(\mathcal{L}(X_k^{\mathrm{N}}), \mathcal{L}(X_k)) \leq \alpha\sqrt{p}\left(10\kappa K + \frac{K^2}{\sqrt{m}}\right) + \frac{4\delta\sqrt{p}}{m}.$$

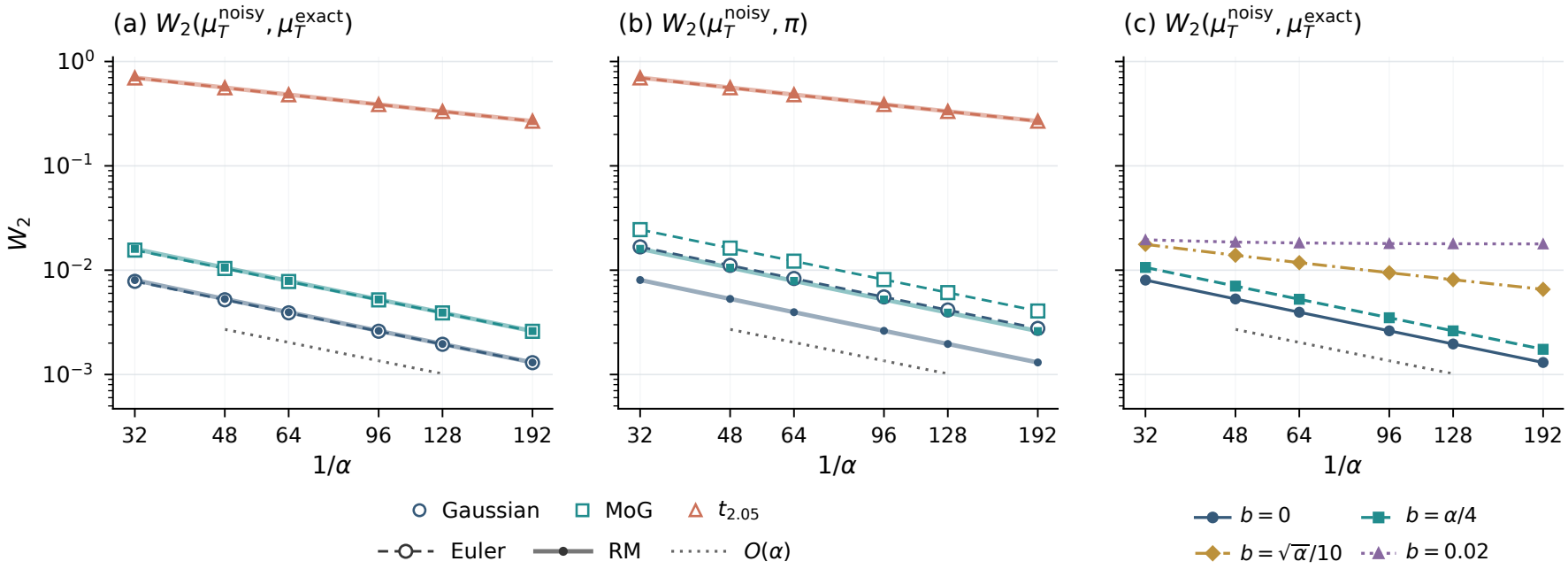
For twice-differentiable U, adding the exact-RM bound gives

$$W_2(\mathcal{L}(X_k^{\mathrm{N}}), \pi) \leq \underbrace{1.11\exp(-mk\alpha/2)\, W_2(\mathcal{L}(X_0^{\mathrm{N}}), \pi)}_{\text{initialization}}$$

$$+ \underbrace{(2.4\sqrt{\kappa M\alpha} + 1.77)M\alpha\sqrt{p/m}}_{\text{discretization}} + \underbrace{\frac{4\delta\sqrt{p}}{m}}_{\text{bias}} + \underbrace{\alpha\sqrt{p}\left(10\kappa K + \frac{K^2}{\sqrt{m}}\right)}_{\text{centered noise}}.$$

Numerical results

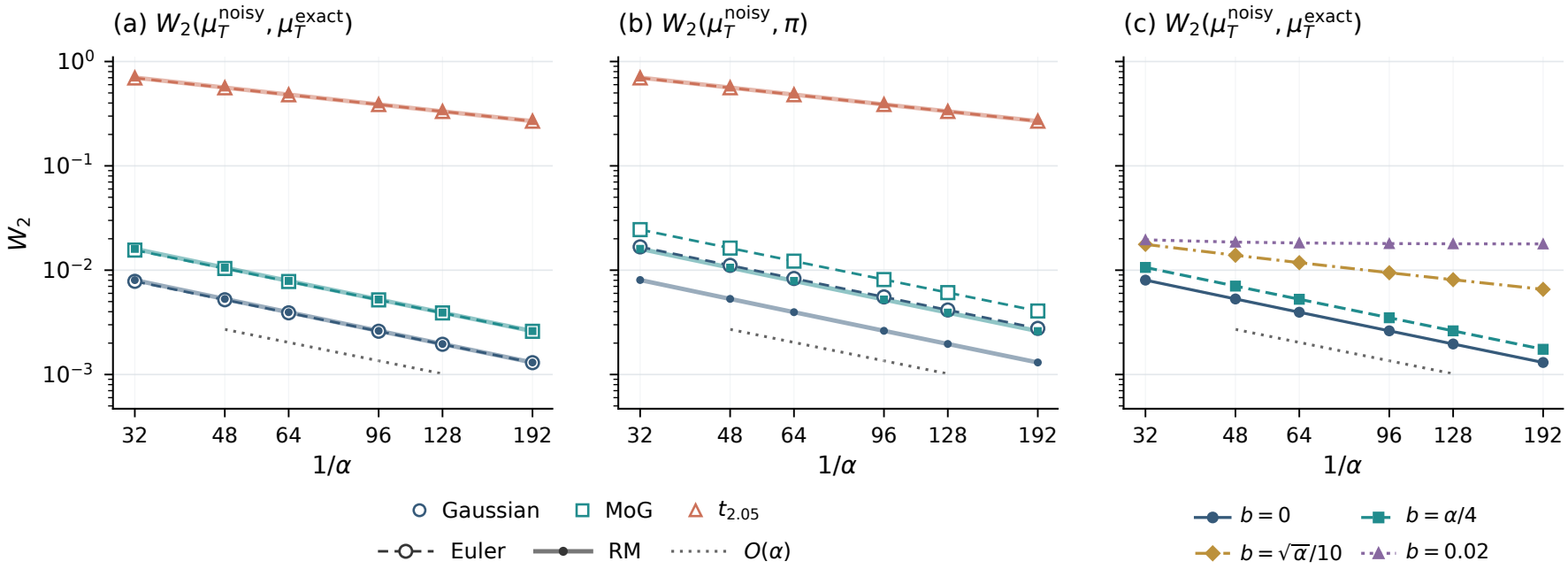
Figure 2. Strongly convex target, $X_0 \sim \pi$, terminal time $T = 8$.



- In panel (a), the error from gradient noise is nearly linear in α for both methods with Gaussian or Gaussian-mixture (MoG) noise.
- With Student t noise (2.05 degrees of freedom), the error decays more slowly over these step sizes.

Numerical results

Figure 2. Strongly convex target, $X_0 \sim \pi$, terminal time $T = 8$.



- In panel (a), the error from gradient noise is nearly linear in α for both methods with Gaussian or Gaussian-mixture (MoG) noise.
- With Student t noise (2.05 degrees of freedom), the error decays more slowly over these step sizes.

Summary

- **Goal.** A sharper one-step W_2 bound when we replace the gradient in an Euler or RM update by an estimate.
- **Euler.** The centered-noise bound is $O(\sqrt{\alpha})$ under bounded variance and $O(\alpha)$ under sub-Gaussian tails.
- **Randomized midpoint.** To our knowledge, these are the first such W_2 bounds for noisy RM.

Thank you!

References

Y. Xiong and V. Rao (2026). Wasserstein bounds for Euler and randomized midpoint Langevin sampling with noisy gradients. *arXiv:2610.37858*.

- A. Dalalyan and A. Karagulyan (2019). User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes and their Applications*.
- M. Talagrand (1996). Transportation cost for Gaussian and other product measures. *Geometric and Functional Analysis*.
- R. Shen and Y. T. Lee (2019). The randomized midpoint method for log-concave sampling. *Advances in Neural Information Processing Systems*.
- L. Yu, A. Karagulyan and A. Dalalyan (2024). Langevin Monte Carlo for strongly log-concave distributions: Randomized midpoint revisited. *International Conference on Learning Representations*.
- F. Pedrotti and P. A. Whalley (2026). Wasserstein mixing time of the unadjusted Langevin algorithm. *arXiv:2608.02430*.

Proof of Lemma 1

Fix X and write $\xi = \xi(X)$. Let $t = \sqrt{\alpha/2}$, $\gamma = \mathcal{N}(0, I_p)$ and $\mu_t = \mathcal{L}(G - t\xi)$, with $G \sim \gamma$ independent of ξ.

$$W_2^2(\mu_t, \gamma) \overset{\text{Talagrand}}{\leq} 2D_{\mathrm{KL}}(\mu_t \| \gamma) \overset{\text{Jensen}}{\leq} 2 \log \int r_t^2 \, \mathrm{d}\gamma, \quad r_t = \frac{\mathrm{d}\mu_t}{\mathrm{d}\gamma}.$$

For an independent copy ζ of ξ, Gaussian integration and the sub-Gaussian bound give

$$\int r_t^2 \, \mathrm{d}\gamma = \mathbb{E} \exp(t^2 \zeta^\top \xi) \leq (1 - K^4 t^4)^{-p/2}.$$

Since $\alpha K^2 \leq 1$ implies $K^4 t^4 \leq 1/4$, rescaling gives

$$2\alpha W_2^2(\mu_t, \gamma) \leq 2\alpha p[-\log(1 - K^4 t^4)] \leq \tfrac{2}{3} \alpha^3 K^4 p.$$

RM local error

Add back the centered error from stage two to obtain

$$\bar{X}_k = X_k^{\mathrm{N}} + \alpha \xi_k^{(2)}.$$

An exact RM step from $\bar{X}_k$ gives $\breve{X}_{k+1}$.

Local error (Lemma 3.8)

In $\bar{X}_{k+1} - \breve{X}_{k+1}$, pairing the previous error with the fresh Gaussian gives

$$\mathbb{E}\left\| -\alpha \xi_k^{(2)} + \sqrt{2\alpha}\,(Z_{k+1} - \widetilde{Z}_{k+1}) \right\|^2 \leq \tfrac{2}{3}\alpha^3 K^4 p.$$

The remaining drift and bias terms have L^2 norm $O(\alpha^2 + \alpha\delta)$.

RM contraction

From $\bar{X}_k$ and X_k, take exact RM steps with the same random draws.

Contraction (Lemma 3.9)

For $\alpha \le 1/(8M\sqrt{\kappa})$,

$$\|\breve{X}_{k+1} - X_{k+1}\|^2 \le (1 - \tfrac{47}{64} m\alpha)\|\bar{X}_k - X_k\|^2.$$

For chains with the same initial distribution, Theorem 3.10 gives

$$\sup_{k \ge 0} W_2(\mathcal{L}(X_k^{\mathrm{N}}), \mathcal{L}(X_k)) = O(\alpha + \delta).$$