

Inria



European Research Council  
Established by the European Commission

# Quantifying the mini-batching error in Bayesian inference for Adaptive Langevin dynamics

**Gabriel STOLTZ**

(CERMICS, Ecole des Ponts & MATHEMATICALS team, Inria Paris)

*In collaboration with B. Leimkuhler, M. Sachs and I. Sekkat*

Séminaire parisien de statistiques (Mars 2023)

- **Sampling probability measures with Langevin dynamics**
- **Mini-batching for (overdamped) Langevin dynamics**
  - Structure of the covariance matrix of gradient estimator
  - Bias on the posterior distribution
- **Adaptive Langevin dynamics (AdL)**
  - Motivation for constant covariance matrix
  - Error estimates for generic covariance matrices
  - Theoretical convergence results<sup>1</sup>
- **Extended AdL (non-constant covariance matrices<sup>2</sup>)**

---

<sup>1</sup>B. Leimkuhler, M. Sachs and G. Stoltz, Hypocoercivity properties of adaptive Langevin dynamics, *SIAM J. Appl. Maths.* (2020)

<sup>2</sup>I. Sekkat and G. Stoltz, Removing the mini-batching error in Bayesian inference using Adaptive Langevin dynamics, *arXiv preprint* **2105.10347**

# Bayesian inference

- **Data**  $\{x_i\}_{i=1, \dots, N_{\text{data}}}$  **to be explained by a statistical model**
  - Parametrization by  $\theta \in \mathbb{R}^d$ : individual likelihoods  $P_{\text{elem}}(x_i|\theta)$
  - Prior  $P_{\text{prior}}(\theta)$  on the parameters
  - Sample  $\theta$  from  $\pi(\theta|\mathbf{x}) \propto P_{\text{prior}}(\theta) \prod_{i=1}^{N_{\text{data}}} P_{\text{elem}}(x_i|\theta)$
  - Usual MCMC (e.g. Metropolis): **each step costs  $O(N_{\text{data}})$**

- **Running example:** **Gaussian mixture** model  $\theta = (\mu_1, \mu_2)$

$$P_{\text{elem}}(x_i|\theta) = \frac{w}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{(x_i - \mu_1)^2}{2\sigma_1^2}\right) + \frac{1-w}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{(x_i - \mu_2)^2}{2\sigma_2^2}\right),$$

with  $\sigma_1 = \sigma_2 = 0.4$ ,  $w = 0.4$ , and Gaussian prior

Sample  $N_{\text{data}} = 200$  points from  $P_{\text{elem}}(\cdot|\theta^*)$  with  $\theta^* = (1, 0.5)$

(Overdamped)  
Langevin dynamics

# (Underdamped/kinetic) Langevin dynamics

Momenta  $p \in \mathbb{R}^d$ , conjugate to parameters  $\theta \in \mathbb{R}^d$

**Hamiltonian**  $H(\theta, p) = -\ln \pi(\theta|\mathbf{x}) + \frac{|p|^2}{2}$ , **friction**  $\gamma > 0$

$$\begin{cases} d\theta_t = p_t dt \\ dp_t = \nabla \ln \pi(\theta_t|\mathbf{x}) dt - \gamma p_t dt + \sqrt{2\gamma} dW_t \end{cases}$$

**Generator**  $\mathcal{L} = \mathcal{L}_{\text{ham}} + \gamma \mathcal{L}_{\text{FD}}$  with

$$\mathcal{L}_{\text{ham}} = p^T \nabla_{\theta} + [\nabla \ln \pi(\cdot|\mathbf{x})]^T \nabla_p, \quad \mathcal{L}_{\text{FD}} = -p^T \nabla_p + \Delta_p$$

**Unique inv. proba. meas.**  $\mu(d\theta dp) = \pi(\theta|\mathbf{x}) \mathcal{N}(0, \text{Id}_d) \propto e^{-H(\theta,p)} d\theta dp$

$$\forall \varphi, \quad \int_{\mathcal{E}} \mathcal{L} \varphi d\mu = 0 \quad \iff \quad \mathcal{L}^{\dagger} \mu = 0$$

# Hamiltonian and overdamped limits

- As  $\gamma \rightarrow 0$ , the **Hamiltonian** dynamics is recovered

$$\frac{d}{dt} \mathbb{E} [H(\theta_t, p_t)] = -\gamma (\mathbb{E} [|p_t|^2] - d) dt$$

Time  $\sim \gamma^{-1}$  to change energy levels in this limit<sup>3</sup>

- Overdamped** limit  $\gamma \rightarrow +\infty$ : rescaling of time  $\gamma t$

$$\begin{aligned} \theta_{\gamma t} - \theta_0 &= \frac{1}{\gamma} \int_0^{\gamma t} \nabla \ln \pi(\theta_s | \mathbf{x}) ds + \sqrt{\frac{2}{\gamma}} W_{\gamma t} - \frac{1}{\gamma} (p_{\gamma t} - p_0) \\ &= \int_0^t \nabla \ln \pi(\theta_{\gamma s} | \mathbf{x}) ds + \sqrt{2} B_t - \frac{1}{\gamma} (p_{\gamma t} - p_0) \end{aligned}$$

which converges to the solution of  $d\Theta_t = \nabla \ln \pi(\Theta_t | \mathbf{x}) dt + \sqrt{2} dB_t$

- In both cases, **slow convergence**, with rate scaling as  $\min(\gamma, \gamma^{-1})$

---

<sup>3</sup>Hairer and Pavliotis, *J. Stat. Phys.*, **131**(1), 175-202 (2008)

# Ergodicity results for Langevin dynamics (1)

Almost-sure convergence<sup>4</sup> of **ergodic averages**  $\widehat{\varphi}_t = \frac{1}{t} \int_0^t \varphi(\theta_s, p_s) ds$

**Asymptotic variance** of ergodic averages (with  $\Pi_0\varphi = \varphi - \mathbb{E}_\mu(\varphi)$ )

$$\sigma_\varphi^2 = \lim_{t \rightarrow +\infty} t \text{Var} [\widehat{\varphi}_t^2] = 2 \int_{\mathcal{E}} (-\mathcal{L}^{-1} \Pi_0 \varphi) \Pi_0 \varphi d\mu$$

**Central limit theorem**<sup>5</sup> when Poisson equation can be solved in  $L^2(\mu)$

$$-\mathcal{L}\Phi = \Pi_0\varphi$$

Well-posedness for  $\mathcal{L}$  invertible on subsets of  $L_0^2(\mu) = \Pi_0 L^2(\mu)$

$$-\mathcal{L}^{-1} = \int_0^{+\infty} e^{t\mathcal{L}} dt$$

<sup>4</sup>Kliemann, *Ann. Probab.* **15**(2), 690-707 (1987)

<sup>5</sup>Bhattacharya, *Z. Wahrsch. Verw. Gebiete* **60**, 185-201 (1982)

## Ergodicity results for Langevin dynamics (2)

Prove **exponential convergence** of the semigroup  $e^{t\mathcal{L}}$  on  $E \subset L_0^2(\mu)$

- Lyapunov techniques<sup>6</sup>  $L_{\mathcal{X}}^\infty(\mathcal{E}) = \left\{ \varphi \text{ measurable, } \sup \left| \frac{\varphi}{\mathcal{X}} \right| < +\infty \right\}$
- standard hypocoercive<sup>7</sup> setup  $H^1(\mu)$
- $L^2(\mu)$  after hypoelliptic regularization<sup>8</sup> from  $H^1(\mu)$
- direct transfer from  $H^1(\mu)$  to  $L^2(\mu)$  by spectral argument<sup>9</sup>
- directly<sup>10</sup>  $L^2(\mu)$  (recently<sup>11</sup> Poincaré using  $\partial_t - \mathcal{L}_{\text{ham}}$ )
- coupling arguments<sup>12</sup>
- direct estimates on the resolvent using Schur complements<sup>13</sup>

**Rate of convergence**  $\min(\gamma, \gamma^{-1})$  in all cases

---

<sup>6</sup>Wu ('01); Mattingly/Stuart/Higham ('02); Rey-Bellet ('06); Hairer/Mattingly ('11)

<sup>7</sup>Villani (2009) and before Talay (2002), Eckmann/Hairer (2003), Hérau/Nier (2004),...

<sup>8</sup>Hérau, *J. Funct. Anal.* (2007)

<sup>9</sup>Deligiannidis/Paulin/Doucet, *Ann. Appl. Probab.* (2020)

<sup>10</sup>Hérau (2006), Dolbeaut/Mouhot/Schmeiser (2009, 2015)

<sup>11</sup>Armstrong/Mourrat (2019), Cao/Lu/Wang (2019), Brigati (2021)

<sup>12</sup>Eberle/Guillin/Zimmer, *Ann. Probab.* (2019)

<sup>13</sup>Bernard/Fathi/Levitt/Stoltz, *Annales Henri Lebesgue* (2022)



# Mini-batching for (overdamped) Langevin dynamics

# Mini-batching procedure

Sample  $n$  data points **with(out)** replacement: random set  $I_n$

$$\nabla_{\theta}(\log \pi(\theta|\mathbf{x})) = \nabla_{\theta}(\log P_{\text{prior}}(\theta)) + \sum_{j=1}^{N_{\text{data}}} \nabla_{\theta}(\log P_{\text{elem}}(x_j|\theta))$$

Unbiased stochastic estimator of  $\nabla_{\theta}(\log \pi(\theta|\mathbf{x}))$

$$\begin{aligned}\widehat{F}_n(\theta) &= \nabla_{\theta}(\log P_{\text{prior}}(\theta)) + \frac{N_{\text{data}}}{n} \sum_{i \in I_n} \nabla_{\theta}(\log P_{\text{elem}}(x_i|\theta)) \\ &= \nabla_{\theta}(\log \pi(\theta|\mathbf{x})) + \sqrt{\varepsilon(n)} \Sigma_{\mathbf{x}}(\theta)^{1/2} Z_{\mathbf{x}, N_{\text{data}}, n}\end{aligned}$$

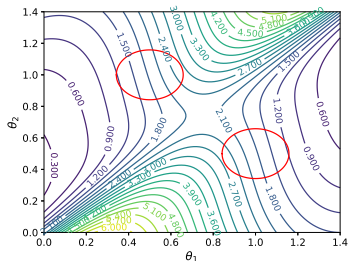
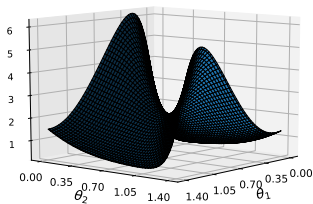
where  $\varepsilon(n) \sim \frac{N_{\text{data}}^2}{n}$  for  $n \ll N_{\text{data}}$

$$\Sigma_{\mathbf{x}}(\theta) = \frac{1}{N_{\text{data}} - 1} \sum_{i=1}^{N_{\text{data}}} [\nabla_{\theta}(\log P_{\text{elem}}(x_i|\theta)) - \text{average}] [\dots]^T \in \mathbb{R}^{d \times d}$$

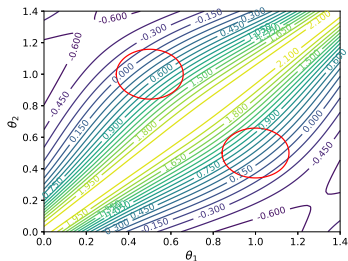
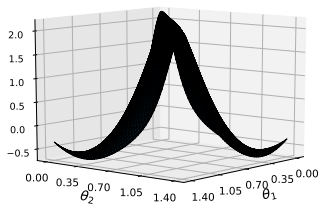
$Z_{\mathbf{x}, N_{\text{data}}, n}$  centered with identity covariance (**Non-Gaussian** for  $n$  small)

# Covariance of gradient estimator for Gaussian mixture

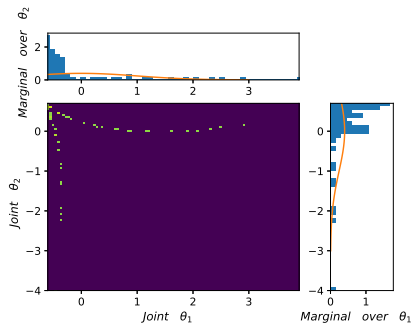
$\Sigma_{1,1}$



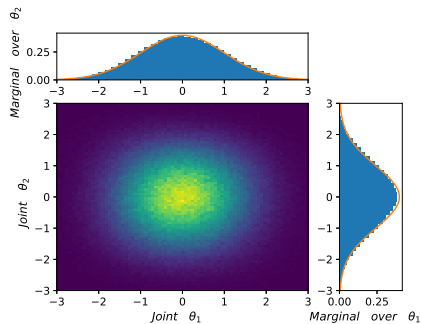
$\Sigma_{1,2}$



# Nature of the random variable $Z_{\mathbf{x}, N_{\text{data}}, n}$



(a)  $n = 1$



(b)  $n = 30$

# Mini-batching and overdamped Langevin dynamics

- Overdamped Langevin  $d\theta_t = \nabla_{\theta}(\log \pi(\theta_t|\mathbf{x})) dt + \sqrt{2} dW_t$ , discretization

$$\theta^{m+1} = \theta^m + \Delta t \nabla_{\theta}(\log \pi(\theta^m|\mathbf{x})) + \sqrt{2\Delta t} G^m$$

- With **mini-batching** (Stochastic gradient Langevin dynamics<sup>14</sup>)

$$\theta^{m+1} = \theta^m + \Delta t \widehat{F}_n(\theta^m) + \sqrt{2\Delta t} G^m.$$

- Amounts to adding **additional Brownian motion of unknown** magnitude

## Effective Langevin dynamics

$$d\tilde{\theta}_t = \nabla_{\theta}(\log \pi(\tilde{\theta}_t|\mathbf{x})) dt + \sqrt{2 \left( 1 + \frac{\varepsilon(n)\Delta t}{2} \Sigma_{\mathbf{x}}(\tilde{\theta}_t) \right)} d\tilde{W}_t$$

**Key point:**  $\mathbb{E}^{\theta_0}(\varphi(\theta^1)) = \mathbb{E}^{\theta_0}(\varphi(\tilde{\theta}_{\Delta t})) + O(\Delta t^3) = \mathbb{E}^{\theta_0}(\varphi(\theta_{\Delta t})) + O(\Delta t^2)$

- **Bias** of order  $\varepsilon(n)\Delta t$  on the invariant measure<sup>15</sup>

<sup>14</sup>Welling/Teh, *ICML* (2011)

<sup>15</sup>S. Vollmer, K. Zygalakis, Y. Teh, *JMLR* (2016)

# Mini-batching and underdamped Langevin dynamics

- Underdamped Langevin dynamics ( $\Gamma \in \mathbb{R}^{d \times d}$  symm. positive definite)

$$\begin{cases} d\theta_t = p_t dt \\ dp_t = \nabla_{\theta}(\log \pi(\theta_t|\mathbf{x})) dt - \Gamma p_t dt + \sqrt{2}\Gamma^{1/2} dW_t \end{cases}$$

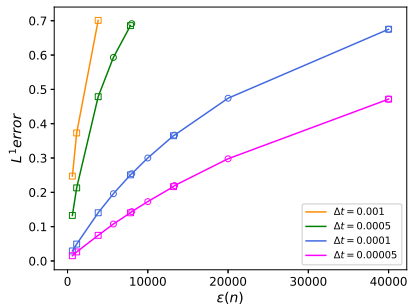
- Preserves the measure  $\pi(\theta|\mathbf{x}) \times \mathcal{N}(0, \text{Id}_d)$
- Splitting scheme + mini-batching (C. Matthews and J. Weare (2018))

$$\begin{cases} p^{m+1/3} = \alpha_{\Delta t/2} p^m + (\text{Id} - \alpha_{\Delta t})^{1/2} G^m, & \alpha_t = e^{-\Gamma t} \\ \theta^{m+1/2} = \theta^m + \Delta t p^{m+1/3} / 2 \\ p^{m+2/3} = p^{m+1/3} + \Delta t \widehat{F}_n(\theta^{m+1/2}) \\ \theta^{m+1} = \theta^{m+1/2} + \Delta t p^{m+2/3} / 2 \\ p^{m+1} = \alpha_{\Delta t/2} p^{m+2/3} + (\text{Id} - \alpha_{\Delta t})^{1/2} G^{m+1/2} \end{cases}$$

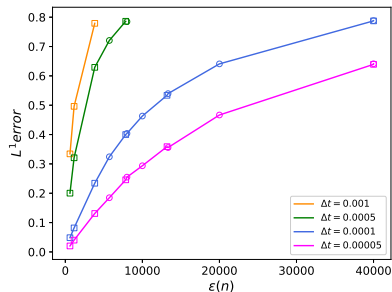
- Bias  $\mathcal{O}(\varepsilon(n)\Delta t)$  since effective Langevin dynamics corresponds to

$$\sqrt{2}\Gamma^{1/2} dW_t \longleftarrow \left(2\Gamma + \varepsilon(n)\Delta t \Sigma_{\mathbf{x}}(\tilde{\theta}_t)\right)^{1/2} d\widetilde{W}_t$$

# Numerical evidence of the bias



(a) SGLD



(b) Langevin dynamics

$L^1$  error on the  $\theta_1$  marginal of the posterior distribution for various values of  $\Delta t$  and  $n$ , when sampling with  $(\circ)$  and without replacement  $(\square)$ .

# Adaptive Langevin dynamics



# Motivation for Adaptive Langevin dynamics (1/2)

**Key assumption:**  $\Sigma_{\mathbf{x}}(\theta)$  is constant (not realistic)

- **Variable friction**  $\xi \in \mathbb{R}^{d \times d}$ , Nosé–Hoover type feedback

Adaptive Langevin dynamics<sup>1</sup>: **unknown**  $A$

$$d\theta_t = p_t dt,$$

$$dp_t = (\nabla(\log \pi(\theta_t | \mathbf{x})) - \xi_t p_t) dt + \sqrt{2} A^{1/2} dW_t,$$

$$d[\xi_t]_{i,j} = \frac{1}{\eta} (p_{i,t} p_{j,t} - \delta_{i,j}) dt, \quad 1 \leq i, j \leq d,$$

- Invariant measure  $\pi(\theta | \mathbf{x}) \times \mathcal{N}(0, \text{Id}_d) \times \prod_{i,j=1}^d \mathcal{N}(A_{ij}, \eta^{-1})$
- **Marginal in  $\theta$  is indeed  $\pi(\theta | \mathbf{x})$**  whatever  $A \dots$  Prove **convergence/CLT?**

<sup>1</sup>A. Jones and B. Leimkuhler, *J. Chem. Phys.* (2011); Ding et al., *NIPS* (2014);  
B. Leimkuhler and X. Shang, *SIAM J. Sci. Comput.* (2015)

## Motivation for Adaptive Langevin dynamics (2/2)

- **effective dynamics** of Strang splitting  $\rightarrow$  AdL for  $A = \gamma \text{Id}_d + \frac{\varepsilon(n)\Delta t}{2} \Sigma_{\mathbf{x}}$

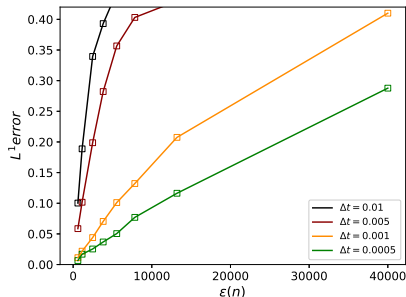
$$\left\{ \begin{array}{l} p^{m+1/2} = e^{-\Delta t \xi^m / 2} p^m + \left[ \gamma (\xi^m)^{-1} \left( \text{Id} - e^{-\Delta t \xi^m} \right) \right]^{1/2} G^m, \\ \xi^{m+1/2} = \xi^m + \frac{\Delta t}{2\eta} \left( p^{m+1/2} \left( p^{m+1/2} \right)^T - \text{Id} \right), \\ \theta^{m+1/2} = \theta^m + \frac{\Delta t}{2} p^{m+1/2}, \\ \tilde{p}^{m+1/2} = p^{m+1/2} + \Delta t \widehat{F}_n \left( \theta^{m+1/2} \right), \\ \vdots \end{array} \right.$$

- When  $\Sigma_{\mathbf{x}}$  is constant, bias on the invariant measure is  $O(\varepsilon(n)^{3/2} \Delta t^2)$

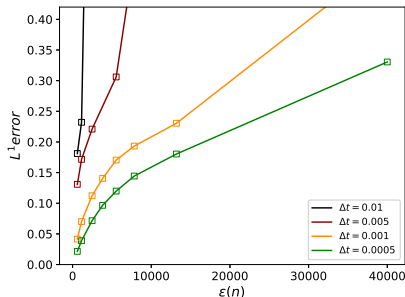
- When  $\Sigma_{\mathbf{x}}$  is not constant, bias of order  $\varepsilon(n)\Delta t \left\| \Sigma_{\mathbf{x}} - \int_{\Theta} \Sigma_{\mathbf{x}} \pi(\cdot | \mathbf{x}) \right\|_{L^2(\pi)}$

# Reduction of the mini-batching error with AdL

$L^1$  error on  $\theta_1$  marginal of posterior; sampling without replacement



(a)  $\xi$  matrix



(b)  $\xi$  scalar

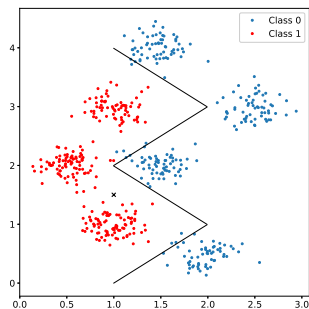
**Linear (asymptotic) regime:** error  $\sim \epsilon(n)\Delta t \min_{M \in \mathcal{M}_d} \|\Sigma_{\mathbf{x}} - M\|_{L^2(\pi)}$

$\mathcal{M}_d$  depends on representation of  $\xi$  (full matrices, diagonal, isotropic...)

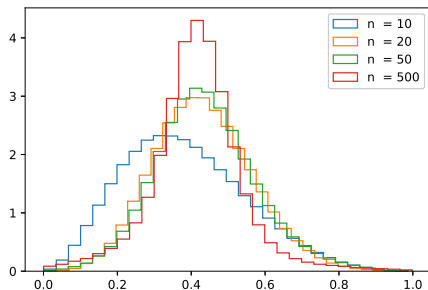
# Bayesian neural networks

Loss: binary cross entropy, prediction  $N_{\theta}(x') = \sigma(W^2 r(W^1 x' + b))$

Hidden layer of size 8



(a) Data to classify



(b)  $N_{\theta}(x')$  for  $\xi$  diagonal

Reduction of minibatching error compared to scalar case

# Convergence of adaptive Langevin dynamics

# Convergence of Adaptive Langevin dynamics

- Case  $A = a\text{Id}_d$  and **scalar friction** (otherwise ergodicity issues)
- Change of variable  $\xi = A + \eta^{-1/2}\zeta\text{Id}_d$  with  $\zeta \in \mathbb{R}$

## Normalized Adaptive Langevin dynamics ( $a$ unknown)

$$\begin{cases} d\theta_t = p_t dt \\ dp_t = \left( \nabla(\log \pi(\theta_t|\mathbf{x})) - ap_t - \frac{\zeta_t}{\sqrt{\eta}} p_t \right) dt + \sqrt{2a} dW_t \\ d\zeta_t = \eta^{-1/2} (|p_t|^2 - d) dt \end{cases}$$

- **Generator**  $\mathcal{L}_{\text{AdL}} = \mathcal{L}_{\text{ham}} + a\mathcal{L}_{\text{FD}} + \eta^{-1/2}\mathcal{L}_{\text{NH}}$ 
  - $\mathcal{L}_{\text{ham}}, \mathcal{L}_{\text{NH}}$  **antisymmetric** and  $\mathcal{L}_{\text{FD}}$  symmetric
  - exponential rate of decay  $\sim \min(a, a^{-1})$  for  $\mathcal{L}_{\text{ham}} + a\mathcal{L}_{\text{FD}}$
  - **Nosé–Hoover**-like part rewritten as  $\eta^{-1/2}(\mathcal{L}_{\text{NH}} + a\sqrt{\eta}\mathcal{L}_{\text{FD}})$   
→ suggests rate of decay  $\sim \eta^{-1/2} \min(a\sqrt{\eta}, (a\sqrt{\eta})^{-1})$

# Precise convergence result

Hypo-coercive estimates<sup>16</sup> in  $L^2(\nu)$ ; complements Lyapunov estimates<sup>17</sup>

## Exponential convergence of the semigroup

There exist  $C, \bar{\lambda}$  such that, for any  $a, \eta > 0$ , there is  $\lambda_{a,\eta} > 0$  for which

$$\forall t \geq 0, \forall \varphi \in L^2(\nu), \quad \left\| e^{t\mathcal{L}_{\text{AdL}}} \varphi - \int \varphi d\nu \right\|_{L^2(\nu)} \leq C e^{-\lambda_{a,\eta} t} \left\| \varphi - \int \varphi d\nu \right\|_{L^2(\nu)}$$

with the lower bound  $\lambda_{a,\eta} \geq \bar{\lambda} \min \left( a, a\eta, \frac{a}{\eta}, \frac{1}{a}, \frac{1}{a\eta} \right)$ . As a consequence,

$$\left\| \mathcal{L}_{\text{AdL}}^{-1} \right\|_{\mathcal{B}(L_0^2(\nu))} \leq \frac{C}{\bar{\lambda}} \max \left( a, a\eta, \frac{1}{a}, \frac{\eta}{a}, \frac{1}{a\eta} \right)$$

Bounds on the resolvent hence on the asymptotic variance<sup>18</sup> and CLT

<sup>16</sup>F. Hérau (2006); J. Dolbeault, C. Mouhot and C. Schmeiser (2009, 2015)

<sup>17</sup>D. Herzog, *Commun. Math. Sci.* (2018)

<sup>18</sup>E. Bernard, M. Fathi, A. Levitt and G. Stoltz, *Ann. Henri Lebesgue* (2022)

# Extended adaptive Langevin dynamics



# Construction of extended AdL

$$\text{Key assumption: } \Sigma_{\mathbf{x}}(\theta) = \sum_{k=0}^K S_k f_k(\theta) \text{ with } S_k \in \mathbb{R}^{d \times d}$$

- **Position dependent** friction  $\xi_t(\theta) = \sum_{k=0}^K \xi_{k,t} f_k(\theta)$  with  $\xi_{k,t} \in \mathbb{R}^{d \times d}$

Extended Adaptive Langevin dynamics for  $A = \gamma \text{Id}_d + \varepsilon(n) \Delta t \Sigma_{\mathbf{x}} / 2$

$$d\theta_t = p_t dt,$$

$$dp_t = \nabla_{\theta}(\log \pi(\theta_t | \mathbf{x})) dt - \xi_t(\theta_t) p_t dt + \sqrt{2} A(\theta_t)^{1/2} dW_t,$$

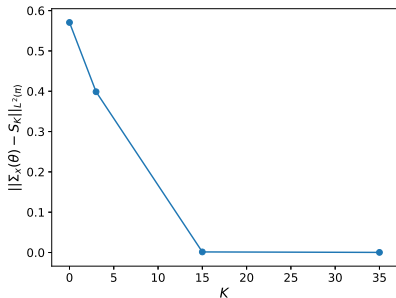
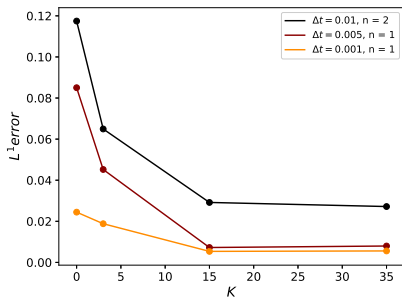
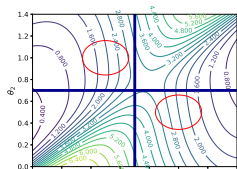
$$d[\xi_{k,t}]_{i,j} = \frac{f_k(\theta_t)}{\eta_k} (p_{i,t} p_{j,t} - \delta_{i,j}), \quad 1 \leq i, j \leq d, \quad 0 \leq k \leq K,$$

- **Bias** on invariant measure  $\sim \varepsilon(n) \Delta t \min_{\mathcal{M}_0, \dots, \mathcal{M}_K} \left\| \Sigma_{\mathbf{x}} - \sum_{k=0}^K \mathcal{M}_k f_k \right\|_{L^2(\pi)}$

# Error on the posterior for Gaussian mixture

Basis functions: partition domain into 4 rectangles  $\mathcal{D}_i$

$$f_i(\theta) = \mathbf{1}_{\mathcal{D}_i}(\theta) \text{ polynomial}(\theta)$$



Left:  $L^1$  error on the  $\theta_1$  marginal posterior for  $n = 1$  (AdL for  $K = 1$ )

Right:  $L^2(\pi)$  projection error of  $\Sigma_x$  onto  $\text{Span}(f_0, \dots, f_K)$

# Conclusion and perspectives

# Main messages

- **Bias on posterior** for underdamped-like Langevin dynamics

$$\sim \frac{N_{\text{data}}^2 \Delta t}{n} \|\Sigma_{\mathbf{x}} - \mathcal{P}_K\|_{L^2(\pi)}$$

where  $\mathcal{P}_K$  depends on the dynamics which is considered

- $\mathcal{P}_K = 0$  for standard Langevin
  - $\mathcal{P}_K = \bar{\Sigma}_{\mathbf{x}} = \int_{\Theta} \Sigma_{\mathbf{x}} \pi(\cdot | \mathbf{x})$  for matrix AdL
  - $\mathcal{P}_K = \frac{1}{d} \text{Tr}(\bar{\Sigma}_{\mathbf{x}}) \text{Id}_d$  for scalar AdL
- Scalar AdL sufficient when  $\bar{\Sigma}_{\mathbf{x}}$  almost isotropic (ex. MNIST logistic regression)
  - Need to **better understand the structure of  $\Sigma_{\mathbf{x}}$**  (low rank?)
- Current investigations on Bayesian neural networks...