

Inria



European Research Council
Established by the European Commission

Reducing the mini-batching error in Bayesian inference with Adaptive Langevin dynamics

Gabriel STOLTZ

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

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

MAC-MIGS tutorial “High Dimensional Sampling and Applications”

- **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 results¹
- **Extended AdL** (non-constant covariance matrices²)

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

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

Mini-batching for (overdamped) Langevin dynamics

Bayesian inference

- **Data** $\{x_i\}_{i=1, \dots, N_{\text{data}}}$ **to be explained by a statistical model**
 - Parametrization by $\theta \in \mathbb{R}^n$: individual likelihoods $P_{\text{elem}}(x_i|\theta)$
 - Prior $P_{\text{prior}}(\theta)$ on the parameters
 - Sample θ from $\pi(\theta|\mathbf{x}) \propto P_{\text{prior}}(\theta) \prod_{i=1}^{N_{\text{data}}} P_{\text{elem}}(x_i|\theta)$
 - Usual MCMC: **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 $\theta^* = (1, 0.5)$

Mini-batching procedure

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

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

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

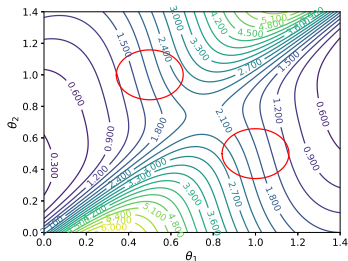
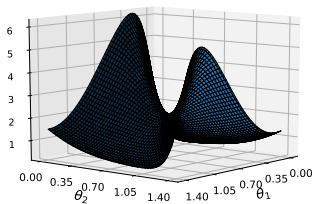
where $\varepsilon(n) \sim \frac{N_{\text{data}}^2}{n}$ for $n \ll N_{\text{data}}$

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

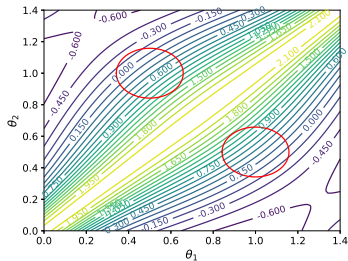
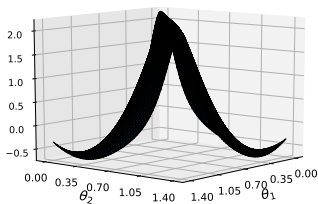
$Z_{\mathbf{x}, N_{\text{data}}, n}$ centered with identity covariance (**Non-Gaussian** for n small)

Covariance of gradient estimator for Gaussian mixture

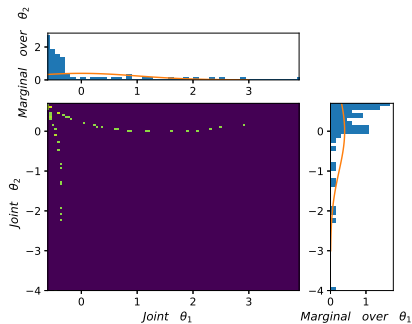
$\Sigma_{1,1}$



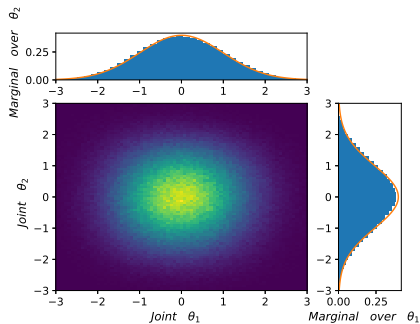
$\Sigma_{1,2}$



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 dynamics³)

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

- 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 measure⁴

³Welling/Teh, *ICML* (2011)

⁴S. 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)

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

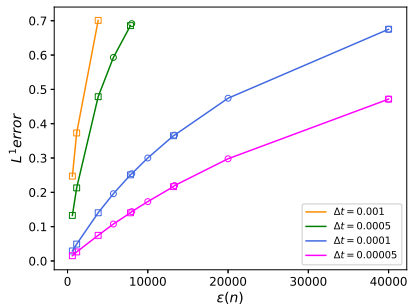
- Preserves the measure $\pi(\theta|\mathbf{x}) \times \mathcal{N}(0, \text{Id}_d)$
- Splitting scheme + mini-batching (C. Matthews and J. Weare (2018))

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

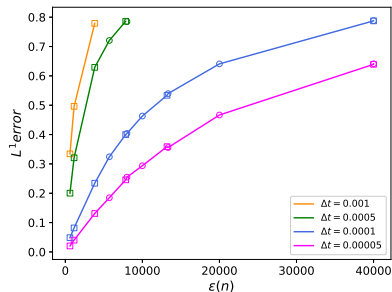
- Bias $\mathcal{O}(\varepsilon(n)\Delta t)$ since effective Langevin dynamics corresponds to

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

Numerical evidence of the bias



(a) SGLD



(b) Langevin dynamics

L^1 error on the θ_1 marginal of the posterior distribution for various values of Δt and n , when sampling with (○) and without replacement (□).

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 dynamics¹: **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 θ is indeed $\pi(\theta | \mathbf{x})$** whatever $A \dots$ Prove **convergence/CLT?**

¹A. 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)

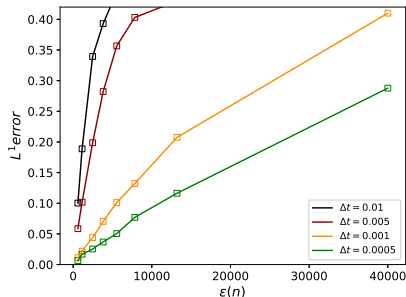
- **effective dynamics** of Strang splitting \rightarrow AdL for $A = \gamma \text{Id}_d + \frac{\varepsilon(n)\Delta t}{2} \Sigma_{\mathbf{x}}$

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

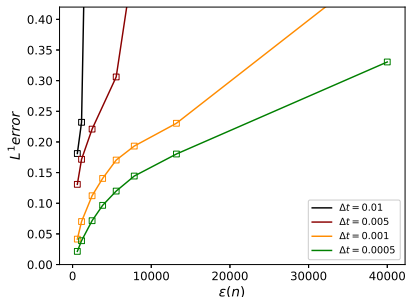
- When $\Sigma_{\mathbf{x}}$ is constant, bias on the invariant measure is $O(\varepsilon(n)^{3/2} \Delta t^2)$
- When $\Sigma_{\mathbf{x}}$ is not constant, bias of order $\varepsilon(n)\Delta t \left\| \Sigma_{\mathbf{x}} - \int_{\Theta} \Sigma_{\mathbf{x}} \pi(\cdot | \mathbf{x}) \right\|_{L^2(\pi)}$

Reduction of the mini-batching error with AdL

L^1 error on θ_1 marginal of posterior; sampling without replacement



(a) ξ matrix



(b) ξ scalar

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

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

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}_{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 estimates⁵ in $L^2(\nu)$; complements Lyapunov estimates⁶

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

$$\|\mathcal{L}_{\text{AdL}}^{-1}\|_{\mathcal{B}(L^2_0(\nu))} \leq \frac{C}{\bar{\lambda}} \max\left(a, a\eta, \frac{1}{a}, \frac{1}{a\eta}\right)$$

Bounds on the resolvent hence on the asymptotic variance⁷ and CLT

⁵F. Hérau (2006); J. Dolbeault, C. Mouhot and C. Schmeiser (2009, 2015)

⁶D. Herzog, *Commun. Math. Sci.* (2018)

⁷E. 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 $A = \gamma \text{Id}_d + \varepsilon(n) \Delta t \Sigma_{\mathbf{x}} / 2$

$$d\theta_t = p_t dt,$$

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

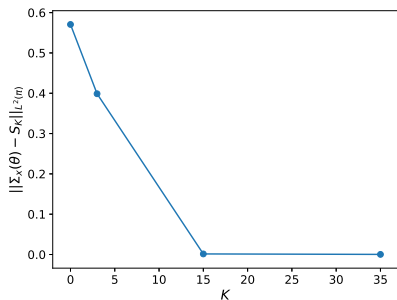
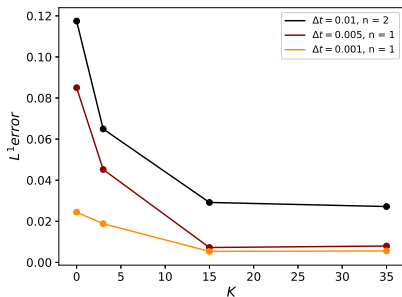
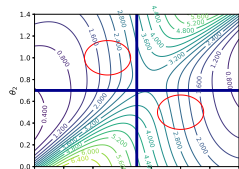
$$d[\xi_{k,t}]_{i,j} = \frac{f_k(\theta_t)}{\eta_k} (p_{i,t} p_{j,t} - \delta_{i,j}), \quad 1 \leq i, j \leq d, \quad 0 \leq k \leq K,$$

- **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 θ_1 marginal posterior for $n = 1$ (AdL for $K = 1$)

Right: $L^2(\pi)$ projection error of Σ_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...