Removing 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

CIRM, "On Future Synergies for Stochastic and Learning Algorithms" (Sept./Oct. 2021)

Outline

- **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[1]

- **Extended AdL** (non-constant covariance matrices[2])

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

[2] I. 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 $\{x_i\}_{i=1,\dots,N_{\mathrm{data}}}$ to be explained by a statistical model**
  - Parametrization by $\theta \in \mathbb{R}^n$: individual likelihoods $P_{\mathrm{elem}}(x_i|\theta)$
  - Prior $P_{\mathrm{prior}}(\theta)$ on the parameters
  - Sample θ from $\pi(\theta|\mathbf{x}) \propto P_{\mathrm{prior}}(\theta) \displaystyle\prod_{i=1}^{N_{\mathrm{data}}} P_{\mathrm{elem}}(x_i|\theta)$
  - Usual MCMC: each step costs $\mathrm{O}(N_{\mathrm{data}})$

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

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

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

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

Mini-batching procedure

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

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

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

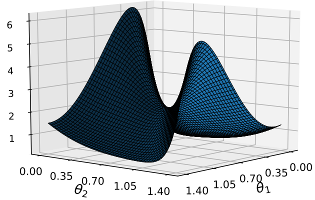
where $\varepsilon(n) \sim \dfrac{N_{\text{data}}^2}{n}$ for $n \ll N_{\text{data}}$

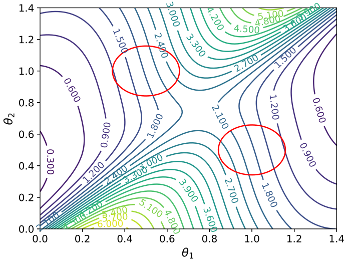
$$\Sigma_{\mathbf{x}}(\theta) = \frac{1}{N_{\text{data}}-1} \sum_{i=1}^{N_{\text{data}}} \left[\nabla_\theta(\log P_{\text{elem}}(x_i|\theta)) - \text{average}\right] [\dots]^T \in \mathbb{R}^{d\times d}$$

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

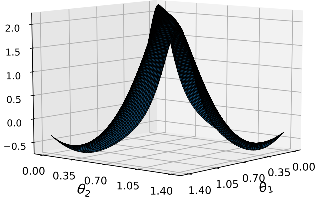
Covariance of gradient estimator for Gaussian mixture

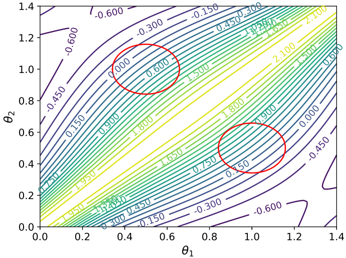
$\Sigma_{1,1}$



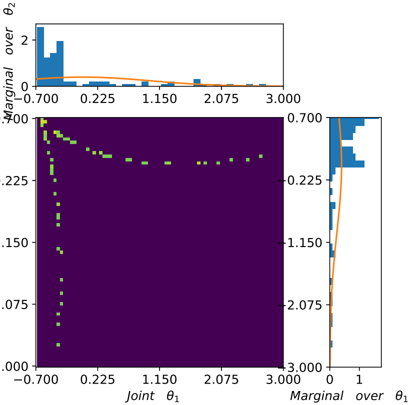


$\Sigma_{1,2}$

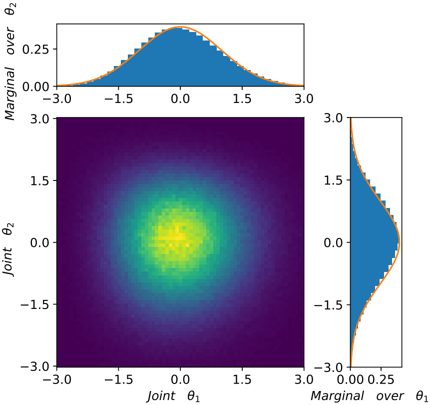




Nature of the random variable $Z_{\mathbf{x},N_{\mathrm{data}},n}$



(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[3])
$$\theta^{m+1} = \theta^m + \Delta t\widehat{F_n}(\theta^m) + \sqrt{2\Delta t}\,G^m.$$
- Amounts to adding additional Brownian motion of unknown magnitude

Effective Langevin dynamics

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

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

- Bias of order $\varepsilon(n)\Delta t$ on the invariant measure[4]

[3] Welling/Teh, *ICML* (2011)
[4] 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 $\pi(\theta|\mathbf{x}) \times \mathcal{N}(0, \mathrm{Id}_d)$

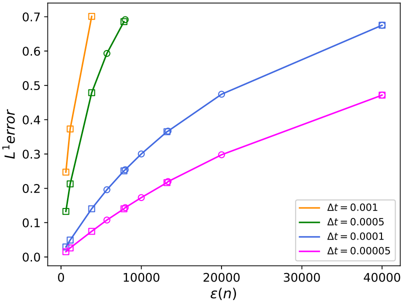
- Splitting scheme + mini-batching

$$\begin{cases} p^{m+1/3} = \alpha_{\Delta t/2}\,p^m + (\mathrm{I}_d - \alpha_{\Delta t})^{1/2}\,G^m, \qquad \alpha_t = \mathrm{e}^{-\Gamma t} \\ \theta^{m+1/2} = \theta^m + \Delta t\,p^{m+1/3}/2 \\ p^{m+2/3} = p^{m+1/3} + \Delta t\widehat{F_n}(\theta^{m+1/2}) \\ \theta^{m+1} = \theta^{m+1/2} + \Delta t\,p^{m+2/3}/2 \\ p^{m+1} = \alpha_{\Delta t/2}\,p^{m+2/3} + (\mathrm{I}_d - \alpha_{\Delta t})^{1/2}\,G^{m+1/2} \end{cases}$$

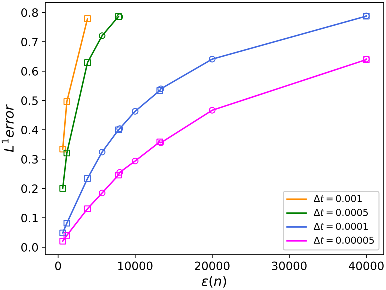
- Bias $\mathrm{O}(\varepsilon(n)\Delta t)$ since effective Langevin dynamics corresponds to

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

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 ($\circ$) 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[1]: unknown A

$$
\begin{aligned}
d\theta_t &= p_t \, dt, \\
dp_t &= \left(\nabla\left(\log \pi(\theta_t|\mathbf{x})\right) - \xi_t p_t\right) dt + \sqrt{2} A^{1/2} \, dW_t, \\
d[\xi_t]_{i,j} &= \frac{1}{\eta} \left(p_{i,t} p_{j,t} - \delta_{i,j}\right) dt, \quad 1 \leqslant i, j \leqslant d,
\end{aligned}
$$

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

[1] 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)

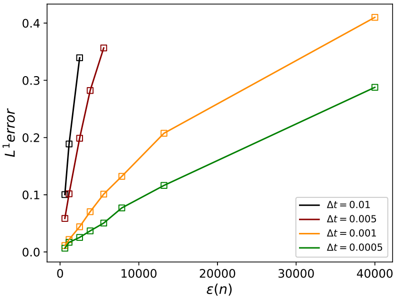
- effective dynamics of Strang splitting $\rightarrow$ AdL for $A = \gamma \mathrm{Id}_d + \frac{\varepsilon(n)\Delta t}{2}\Sigma_{\mathbf{x}}$

$$
\begin{cases}
p^{m+1/2} = \mathrm{e}^{-\Delta t \xi^m/2} p^m + \left[\gamma (\xi^m)^{-1}\left(\mathrm{I}_d - \mathrm{e}^{-\Delta t \xi^m}\right)\right]^{1/2} G^m, \\
\xi^{m+1/2} = \xi^m + \dfrac{\Delta t}{2\eta}\left(p^{m+1/2}\left(p^{m+1/2}\right)^T - \mathrm{I}_d\right), \\
\theta^{m+1/2} = \theta^m + \dfrac{\Delta t}{2} p^{m+1/2}, \\
\widetilde{p}^{m+1/2} = p^{m+1/2} + \Delta t \widehat{F_n}\left(\theta^{m+1/2}\right), \\
\qquad \vdots
\end{cases}
$$

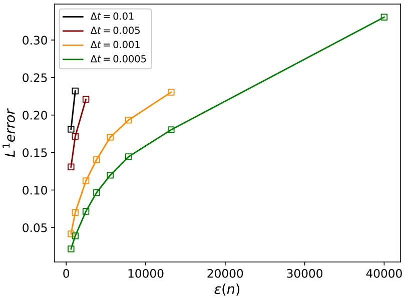
- When $\Sigma_{\mathbf{x}}$ is constant, bias on the invariant measure is $\mathrm{O}(\varepsilon(n)^{3/2}\Delta t^2)$
- When $\Sigma_{\mathbf{x}}$ is not constant, bias of order $\varepsilon(n)\Delta t \left\| \Sigma_{\mathbf{x}} - \int_\Theta \Sigma_{\mathbf{x}}\, \pi(\cdot|\mathbf{x}) \right\|_{L^2(\pi)}$

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

Normalized Adaptive Langevin dynamics (a unknown)

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

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

Precise convergence result

Hypocoercive estimates[5] in $L^2(\nu)$; complements Lyapunov estimates[6]

Exponential convergence of the semigroup

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

$$\forall t \geqslant 0,\ \forall \varphi \in L^2(\nu), \qquad \left\| \mathrm{e}^{t\mathcal{L}_{\mathrm{AdL}}}\varphi - \int \varphi \, d\nu \right\|_{L^2(\nu)} \leqslant C \mathrm{e}^{-\lambda_{a,\eta}} \left\| \varphi - \int \varphi \, d\nu \right\|_{L^2(\nu)}$$

with the lower bound $\lambda_{a,\eta} \geqslant \overline{\lambda} \min\left(a, a\eta, \dfrac{1}{a}, \dfrac{1}{a\eta}\right)$. As a consequence,

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

Bounds on the resolvent hence on the asymptotic variance[7] and CLT

---

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

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

[7] E. Bernard, M. Fathi, A. Levitt and G. Stoltz, *arxiv preprint* **2003.00726**

Extended adaptive Langevin dynamics

Construction of extended AdL

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

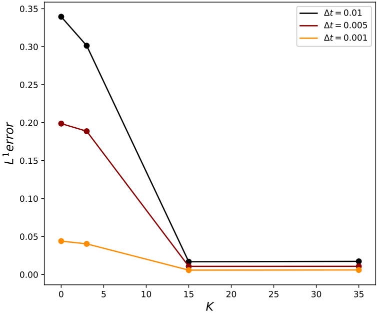
Extended Adaptive Langevin dynamics for $A = \gamma \mathrm{Id}_d + \varepsilon(n)\Delta t \Sigma_{\mathbf{x}}/2$

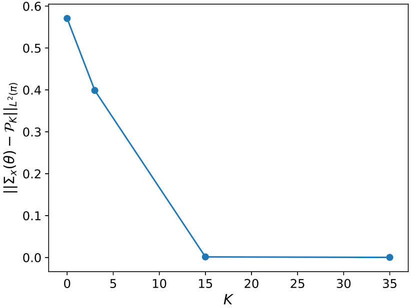
$$
\begin{aligned}
d\theta_t &= p_t \, dt, \\
dp_t &= \nabla_\theta(\log \pi(\theta_t|\mathbf{x}))\, dt - \xi_t(\theta_t) p_t \, dt + \sqrt{2} A(\theta_t)^{1/2} \, dW_t, \\
d[\xi_{k,t}]_{i,j} &= \frac{f_k(\theta_t)}{\eta_k} \left(p_{i,t} p_{j,t} - \delta_{i,j} \right), \qquad 1 \leqslant i,j \leqslant d, \quad 0 \leqslant k \leqslant K,
\end{aligned}
$$

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

Error on the posterior for Gaussian mixture

Basis functions: spatial decomposition of the domain + polynomials





Left: L^1 error on the θ_1-marginal for $n = 15$.
Right: $L^2(\pi)$ projection error of $\Sigma_{\mathbf{x}}$ onto $\mathrm{Span}(f_0, \dots, f_K)$.

Conclusion and perspectives

Main messages

- Bias on posterior for underdamped-like Langevin dynamics

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

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

- $\mathcal{P}_K = 0$ for standard Langevin
- $\mathcal{P}_K = \overline{\Sigma}_{\mathbf{x}} = \displaystyle\int_\Theta \Sigma_{\mathbf{x}}\, \pi(\cdot|\mathbf{x})$ for matrix AdL
- $\mathcal{P}_K = \dfrac{1}{d} \mathrm{Tr}\left(\overline{\Sigma}_{\mathbf{x}}\right) \mathrm{Id}_d$ for scalar AdL

- Scalar AdL sufficient when $\overline{\Sigma}_{\mathbf{x}}$ almost isotropic (ex. MNIST logistic regression)

- Need to better understand the structure of $\Sigma_{\mathbf{x}}$ (low rank?)