École des Ponts
ParisTech

Removing the mini-batching error in Bayesian inference using Adaptive Langevin dynamics

Inass SEKKAT
(CERMICS, Ecole des Ponts ParisTech)

JOURNÉES MAS 2022

August 2022

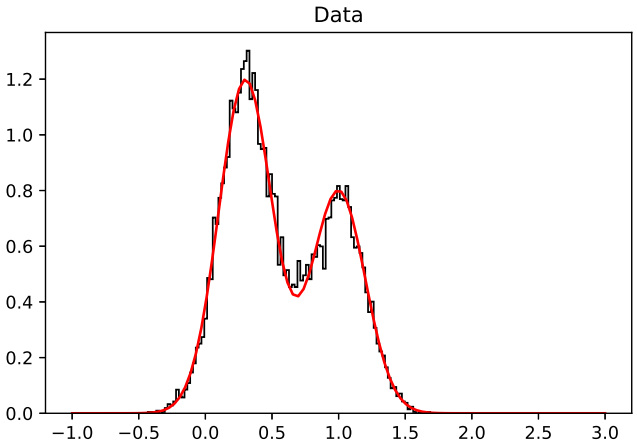
Bayesian inference: context

- **Computing averages w.r.t. the posterior distribution**

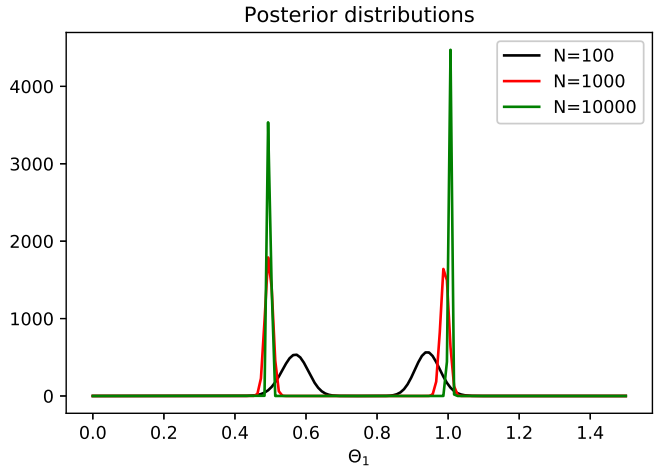
$$\pi(\theta|\mathbf{x}) \propto P_{\text{prior}}(\theta) \prod_{i=1}^{N_{\text{data}}} P_{\text{elem}}(x_i|\theta),$$

  - Data $\mathbf{x} = (x_i)_{i=1,\dots,N_{\text{data}}} \in \mathcal{X}^{N_{\text{data}}}$, iid w.r.t. $P_{\text{elem}}(x_i|\theta)$
  - $\theta \in \Theta \subset \mathbb{R}^d$ the vector of parameters to estimate
  - $P_{\text{prior}}(\cdot)$ is the prior distribution of the vector of parameters

- **Example: Mixture of Gaussians likelihood**
  - $P_{\text{elem}}(x_i|\theta) = \frac{w}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{(x_i-\mu_1)^2}{2\sigma_1^2}\right) + \frac{1-w}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{(x_i-\mu_2)^2}{2\sigma_2^2}\right),$
  - $\theta = (\mu_1, \mu_2)$ and $P_{\text{prior}}(\theta) = \mathcal{N}(0, \mathrm{I}_2)$



$\Longrightarrow$



Parameters:
$\mu_1 = 1,$
$\mu_2 = 0.5,$
$\sigma_1 =$
$\sigma_2 = 0.4,$
$w = 0.5.$

Overdamped Langevin dynamics

Overdamped Langevin dynamics

$$d\theta_t = F(\theta_t)\,dt + \sqrt{2}\,dW_t, \quad F(\theta) = \nabla_\theta \log(\pi(\theta|\mathbf{x}))$$

- π is invariant by the OL dynamics (+ ergodic):

$$\int_\Theta \phi(\theta)\,d\pi(\theta|\mathbf{x}) = \lim_{t\to\infty} \frac{1}{t}\int_0^t \phi(\theta_s)\,ds$$

- Discretization: $\theta^{m+1} = \theta^m + \Delta t F(\theta^m) + \sqrt{2\Delta t}\,G^m$
- Using the estimator $\widehat{\phi}_{\Delta t, N_{\text{iter}}} = \frac{1}{N_{\text{iter}}}\sum_{m=1}^{N_{\text{iter}}} \phi(\theta^m)$, the total error

$$\widehat{\phi}_{\Delta t, N_{\text{iter}}} - \mathbb{E}_\pi(\phi) = \underbrace{\left(\mathbb{E}_{\pi_{\Delta t}}(\phi) - \mathbb{E}_\pi(\phi)\right)}_{\textit{bias}} + \underbrace{\left(\widehat{\phi}_{\Delta t, N_{\text{iter}}} - \mathbb{E}_{\pi_{\Delta t}}(\phi)\right)}_{\text{error coming from variance}}.$$

- Timestep bias $\mathcal{O}(\Delta t)$

Challenge: $F(\theta^m) = \nabla_\theta \log P_{\text{prior}}(\theta^m) + \sum_{i=1}^{N_{\text{data}}} \nabla_\theta \log P_{\text{elem}}(x_i|\theta^m)$,

$\rightarrow$ costs $\mathcal{O}(N_{\text{data}})$ per step, $N_{\text{data}} \gg 1$.

Minibatching: SGLD

- **Minibatching**[1]**:** stochastic estimator (SGLD)

$$\widehat{F_n}(\theta) = \nabla_\theta \log P_{\text{prior}}(\theta) + \frac{N_{\text{data}}}{n} \sum_{i \in I_n} \nabla_\theta \log P_{\text{elem}}(x_i|\theta),$$

$$= \nabla_\theta(\log \pi(\theta|\mathbf{x})) + \varepsilon^{\frac{1}{2}}(n) \Sigma_{\mathbf{x}}(\theta)^{\frac{1}{2}} Z_{\mathbf{x},N_{\text{data}},n},$$

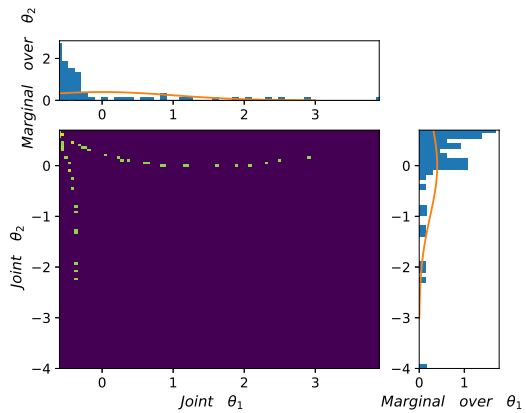
  - I_n a random subset of size n generated by sampling from $\{1, ..., N_{\text{data}}\}$
  - $Z_{\mathbf{x},N_{\text{data}},n}$ a centered random variable with variance I_d
  - $\Sigma_{\mathbf{x}}(\theta) = \mathrm{cov}_{\mathcal{I}}\left(\nabla_\theta \log P_{\text{elem}}(x_{\mathcal{I}}|\theta)\right)$
  - Noise magnitude

$$\varepsilon(n) = \begin{cases} \dfrac{N_{\text{data}}(N_{\text{data}} - 1)}{n}, & \text{for sampling with replacement,} \\ \dfrac{N_{\text{data}}(N_{\text{data}} - n)}{n}, & \text{for sampling without replacement.} \end{cases}$$

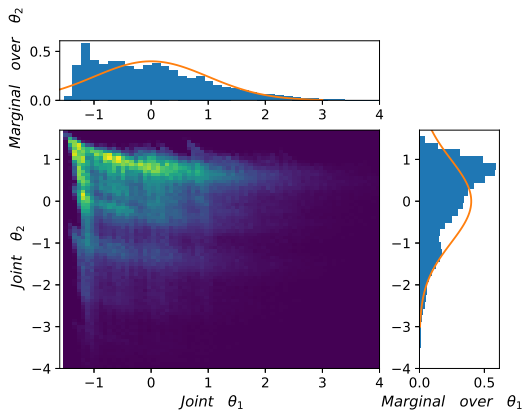
Goal: considering the case of extreme mini-batch size (of order 1)

[1] M. Welling and Y. W. Teh, ICML 2011.

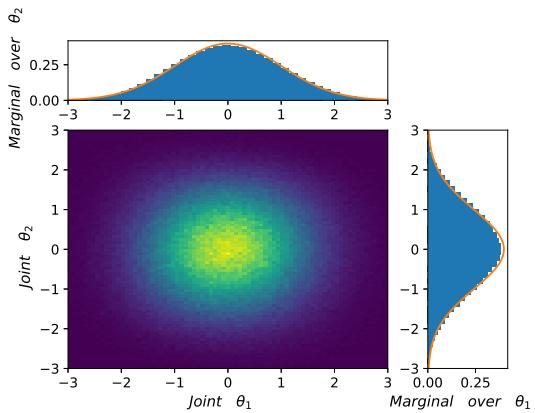
Minibatching: SGLD



$n = 1$



$n = 5$



$n = 30$

$Z_{\mathbf{x},N_{\text{data}},n}$ is not Gaussian for small values of n

$\rightarrow$ **Stochastic Gradient Langevin Dynamics (SGLD)** [2]

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

- Bias on the invariant probability measure[3] $\mathcal{O}((1+\varepsilon(n))\Delta t)$
- Control variable[4] or preferential sampling: reduce the covariance $\Sigma_{\mathbf{x}}$
- modified SGLD[5] renormalize the magnitude of the injected noise using $\Sigma_{\mathbf{x}}(\theta)$

[2] M. Welling and Y. W. Teh, ICML 2011

[3] S. J. Vollmer, K. C. Zygalakis, Y. W. Teh, JMLR 2016

[4] N. Brosse, A. Durmus, and E. Moulines, NeurIPS 2018

[5] S. J. Vollmer, K. C. Zygalakis, and Y. W. Teh, JMLR 2016

Langevin dynamics

Better way to sample from $\pi(\theta|\mathbf{x})$:

Langevin dynamics ($\Gamma \in \mathbb{R}^{d\times d}$ a positive definite symmetric matrix)

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

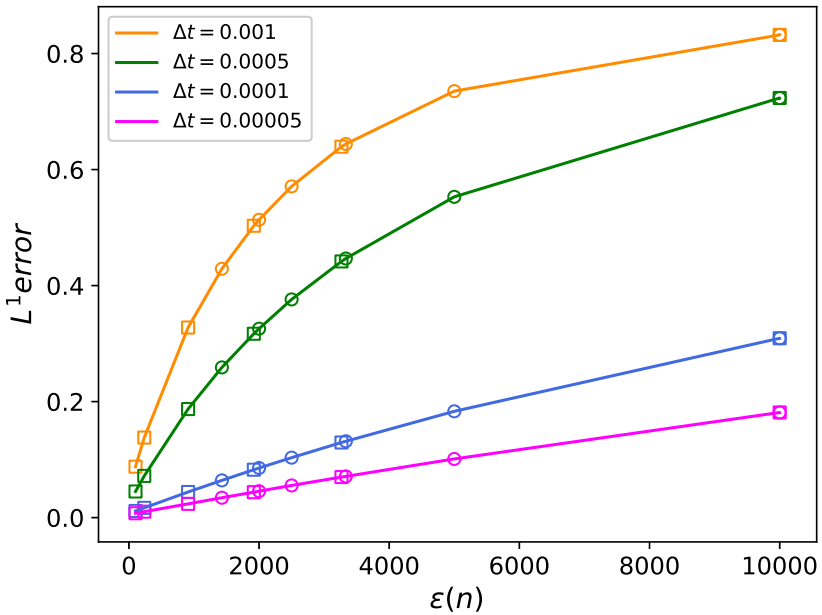
- Generator $\mathcal{L}_{\text{lan}} = \mathcal{L}_{\text{ham}} + \mathcal{L}_{\text{FD},\Gamma}$.
- Invariant probability measure $\mu(d\theta\, dp|\mathbf{x}) = \pi(\theta|\mathbf{x}) \dfrac{\exp(-|p|^2/2)}{(2\pi)^{d/2}}\, d\theta\, dp$
- Bias when minibatching $\mathcal{O}(\varepsilon(n)\Delta t) \gg$ timestep bias $\mathcal{O}(\Delta t^2)$ (numerical scheme by splitting)

Effective dynamics associated with numerical scheme

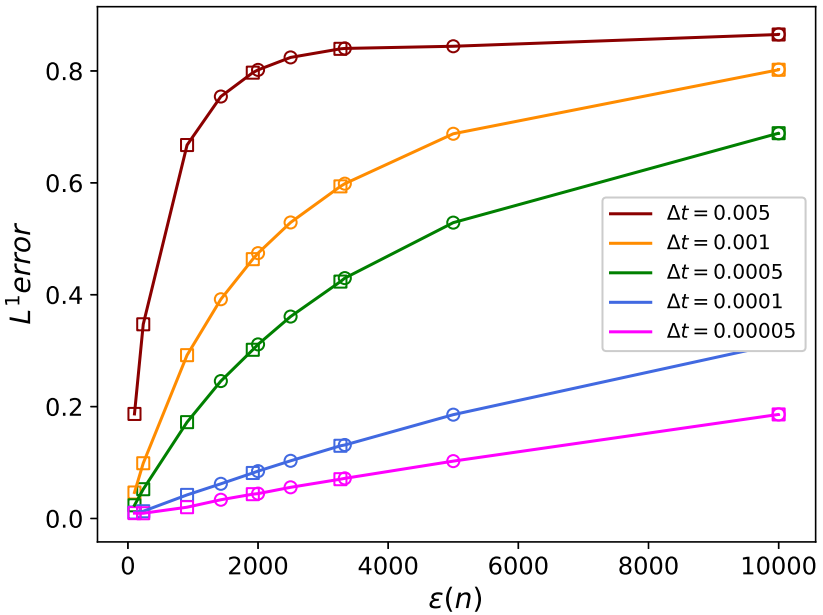
$$\begin{cases} d\tilde{\theta}_t = \tilde{p}_t\, dt, \\ d\tilde{p}_t = \nabla_{\tilde{\theta}}(\log \pi(\tilde{\theta}_t|\mathbf{x})\, dt - \Gamma \tilde{p}_t\, dt + \left(2\Gamma + \Delta t \varepsilon(n) \Sigma_{\mathbf{x}}(\tilde{\theta}_t)\right)^{1/2} dW_t \end{cases}$$

Numerical illustration

$$\varepsilon(n) = \begin{cases} \dfrac{N_{\text{data}}(N_{\text{data}} - 1)}{n}, & \text{for sampling with replacement,} \\ \dfrac{N_{\text{data}}(N_{\text{data}} - n)}{n}, & \text{for sampling without replacement.} \end{cases}$$



Using SGLD



Using a discretization of Langevin dynamics

L^1 error on the θ_1 marginal of the posterior distribution when the elementary likelihoods are mixtures of Gaussians, when sampling with ($\circ$) and without replacement ($\square$).

Adaptive Langevin dynamics

- Effective fluctuation in Langevin dynamics $A_{\Delta t}(\theta) = \Gamma + \frac{\varepsilon(n)\Delta t}{2}\Sigma_{\mathbf{x}}(\theta)$.

Fundamental assumption: constant covariance (H)

$$\Sigma_{\mathbf{x}}(\theta) = \Sigma_{\mathbf{x}}, \text{ constant}$$

- The Adaptive Langevin dynamics automatically corrects for the extra noise[6,7]:

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

- Invariant probability measure $\pi(\theta|\mathbf{x})\,\mathcal{N}_p(0, \mathrm{I_d})\,\mathcal{N}_\xi\left(A_{\Delta t}, \eta^{-1}\mathrm{I_d}\right)$.

$\rightarrow$ diagonal case (assuming $\Sigma_{\mathbf{x}}$ is diagonal): $d[\xi_t]_i = \frac{1}{\eta}\left(p_i^2 - 1\right) dt, \qquad 1 \leqslant i \leqslant d$

$\rightarrow$ scalar case(assuming that $\Sigma_{\mathbf{x}} = \sigma \mathrm{I}_d$): $d\xi_t = \frac{1}{\eta}\left(p^T p - d\right) dt$

[6] A. Jones and B. Leimkuhler, The Journal of Chemical Physics, 2011

[7] Ding & al., Advances in Neural Information Processing Systems 27, 2014

Adaptive Langevin dynamics

- Numerical scheme by Strang splitting, with $A_{\Delta t} = \gamma \mathrm{I}_d + \frac{\varepsilon(n)\Delta t}{2}\Sigma_{\mathbf{x}}(\theta)$

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

- When $\Sigma_{\mathbf{x}}$ is constant, bias of order $\mathcal{O}(\varepsilon(n)^{3/2}\Delta t^2)$.

Adaptive Langevin dynamics

If the Fundamental assumption (H) is not satisfied ...

- When $\Sigma_{\mathbf{x}}$ is not constant, bias of order

$$\varepsilon(n)\Delta t \min_{M \in \mathscr{M}_d} \|\Sigma_{\mathbf{x}} - M\|_{L^2(\pi)},$$

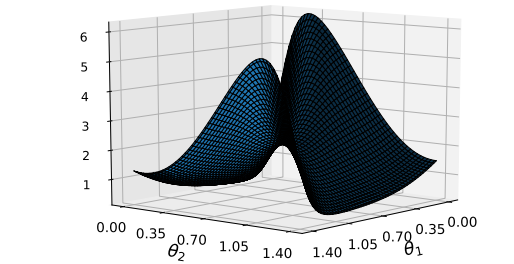
where $\mathscr{M}_d$ depends on representation of ξ.

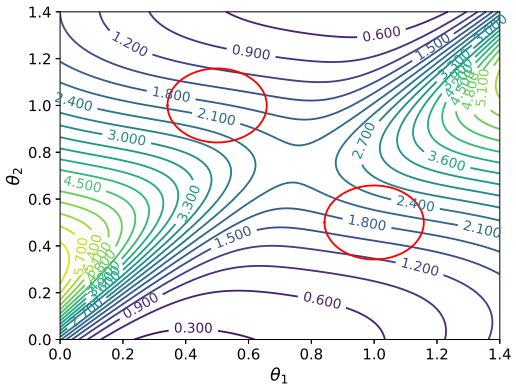
$\rightarrow$ full matrix: $M^* = \int_\Theta \Sigma_{\mathbf{x}}(\theta)\pi(\theta|\mathbf{x})\, d\theta$

$\rightarrow$ diagonal matrix: M^* with entries $\int_\Theta [\Sigma_{\mathbf{x}}(\theta)]_{i,i}\, \pi(\theta|\mathbf{x})\, d\theta, \qquad 1 \leqslant i \leqslant d$

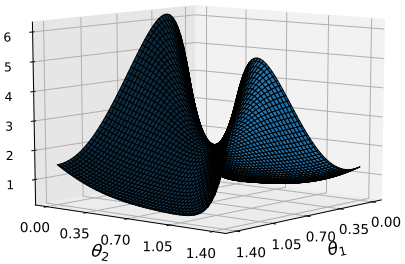
$\rightarrow$ scalar: $M^* = \frac{1}{d} \int_\Theta \mathrm{Tr}\left(\Sigma_{\mathbf{x}}(\theta)\right) \pi(\theta|\mathbf{x})\, d\theta.$

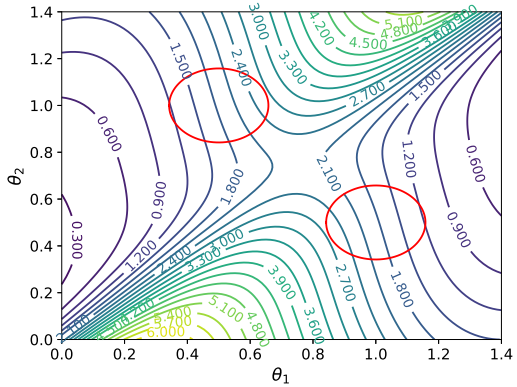
Covariance of the force not constant



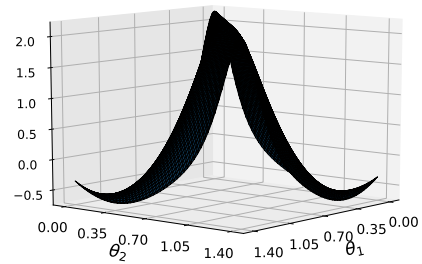


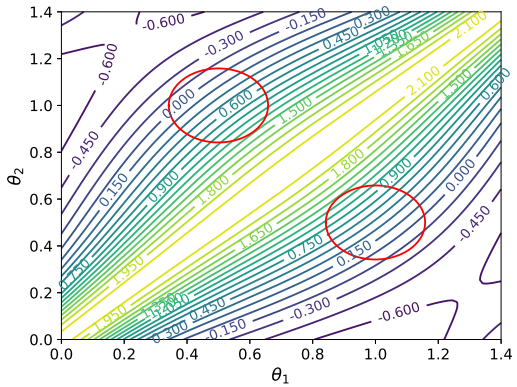
$\Sigma_{\mathbf{x},1,1}(\theta)$





$\Sigma_{\mathbf{x},1,2}(\theta)$



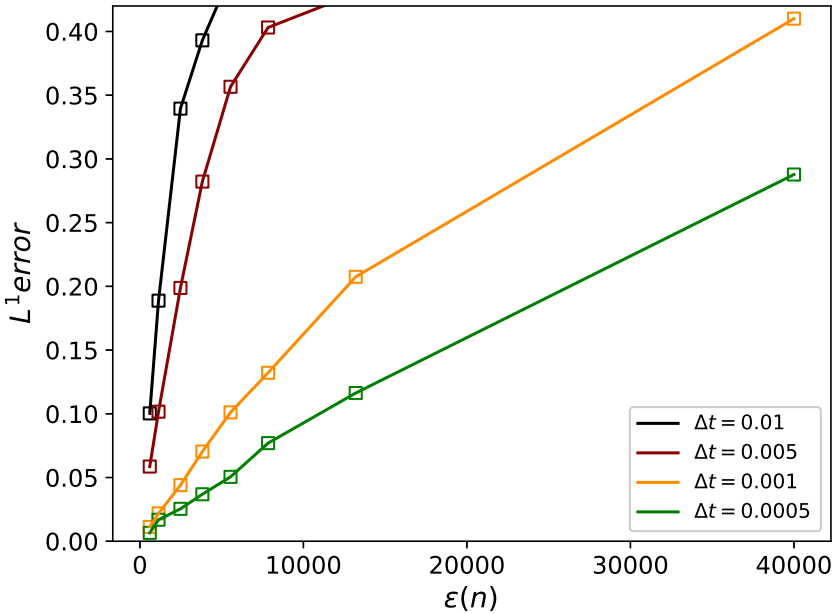


$\Sigma_{\mathbf{x},2,2}(\theta)$

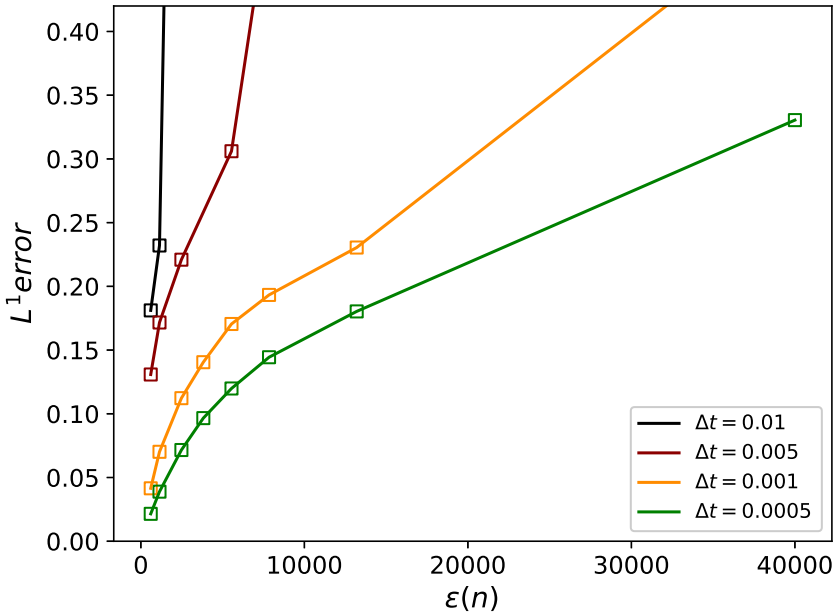
$\rightarrow$ Assumption (H) does not hold: bias $\mathcal{O}(\varepsilon(n)\Delta t)$ as for Langevin dynamics, but smaller prefactor, proportional to

$$\left\| \Sigma_{\mathbf{x}} - \int_{\Theta} \Sigma_{\mathbf{x}}(\theta)\pi(\theta|\mathbf{x})\, d\theta \right\|^2_{L^2(\pi)}$$

Numerical illustration



ξ matrix



ξ scalar

L^1 error on the posterior distribution, when sampling from the posterior distribution for the mixture of Gaussians case using AdL (sampling without replacement).

AdL leads to smaller bias than SGLD but bias remains significant for smaller n (for which the target computational gains are obtained).

Extended Adaptive Langevin Dynamics

- **Assumption:** there exists $f_0, \dots, f_K$ a finite basis of functions such that

$$A_{\Delta t}(\theta) = \sum_{k=0}^{K} A_k f_k(\theta).$$

The Extended Adaptive Langevin[8]:

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

- **Invariant probability measure:**

$$\pi_K(d\theta\, dp\, d\xi_0 \dots d\xi_N) \propto \pi(\theta|\mathbf{x}) d\theta \mu(dp) \prod_{k=0}^{K} \prod_{i,j=1}^{d} \exp\left(-\frac{\eta_k}{2}\left(\xi_{k,i,j} - A_{k,i,j}\right)^2\right) d\xi_{k,i,j}.$$

[8] I. Sekkat, G. Stoltz, arXiv preprint 2105.10347

Extended Adaptive Langevin Dynamics

- Bias on the invariant measure $\varepsilon(n)\Delta t \left\| \Sigma_{\mathbf{x}} - \overline{\Sigma}_{\mathbf{x}}^{K} \right\|_{L^2(\pi)}$, where

$$\left\| \Sigma_{\mathbf{x}} - \overline{\Sigma}_{\mathbf{x}}^{K} \right\|_{L^2(\pi)} = \inf_{M_0,\ldots,M_K \in \mathbb{R}^d} \left\| \Sigma_{\mathbf{x}} - \sum_{k=0}^{K} M_k f_k \right\|_{L^2(\pi)},$$

$\rightarrow L^2(\pi)$ projection of $\Sigma_{\mathbf{x}}$ onto the vector space of symmetric matrices generated by the basis

Choice of basis functions:

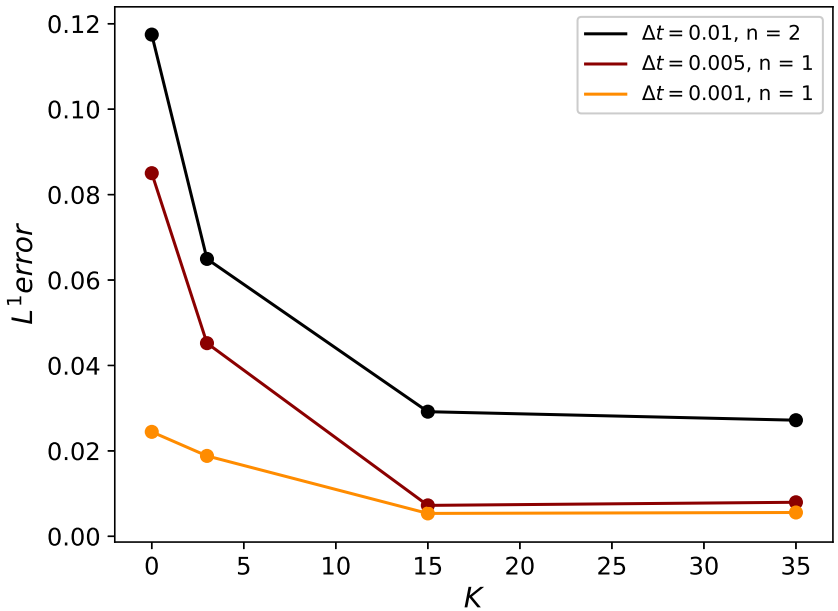
- Spatial decomposition: piecewise constant approximation (geometric decompositions or Voronoi tessellation)

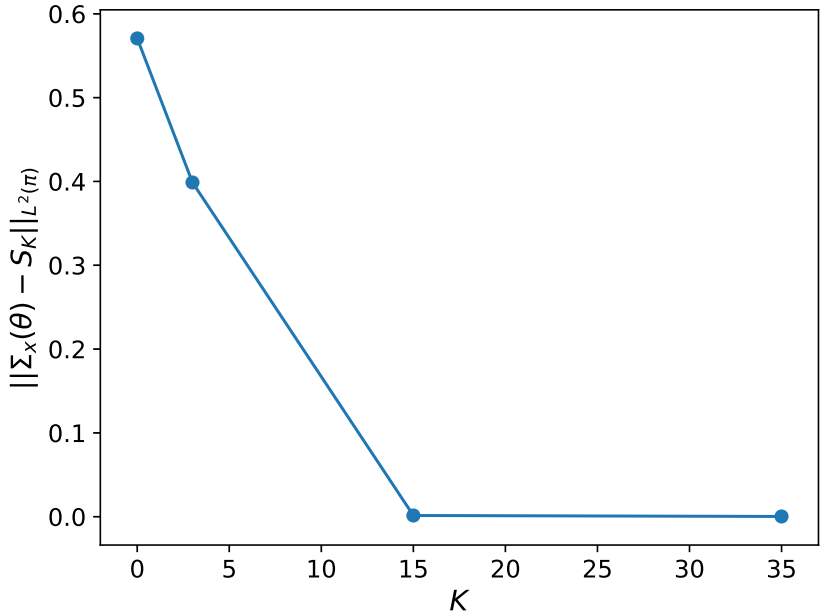
- Polynomial approximation: couple the decomposition with spectral approximation (polynomial basis for example)

Numerical example

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

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





Left: L^1 error on the posterior distribution when sampling using eAdL for various values of K and Δt when $n = 1$ ($K = 1$ corresponds to AdL), Right: $\|\Sigma_{\mathbf{x}} - S_K\|_{L^2(\pi)}$

Summary

- **Main messages**
  - Bias on posterior for Langevin like dynamics

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

    where $\mathcal{P}_K$ depends on the dynamics:
    - $\mathcal{P}_K = 0$ for the Langevin dynamics
    - $\mathcal{P}_K = \int_\Theta \Sigma_{\mathbf{x}}(\theta)\pi(\theta|\mathbf{x})\, d\theta$ for AdL dynamics (full matricial case)
  - Scalar AdL sufficient when $\Sigma_{\mathbf{x}}$ almost isotropic (ex. MNIST logistic regression).

- **Perspectives**
  - Need to better understand the structure of $\Sigma_{\mathbf{x}}$[9]
  - Bayesian neural network.

[9] P. Chaudhari & S. Soatto, Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks, ICLR 2018