Removing the mini-batching error in large scale Bayesian sampling

Gabriel STOLTZ

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

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

Gainesville, January 2020

Outline

- **Mini-batching for Langevin dynamics**
  - Motivation: Bayesian inference for large data sets
  - Bias in the sampled distributions
- **Adaptive Langevin dynamics**
  - Structure of the dynamics
  - Consistency (unbiasedness)
  - Central Limit theorem for trajectory averages
- **Current and future tracks**

B. Leimkuhler, M. Sachs and G. Stoltz, Hypocoercivity properties of adaptive Langevin dynamics, *arXiv preprint* **1908.09363**

Mini-batching for Langevin dynamics

Bayesian inference in the large data context

- **Data $\{x_i\}_{i=1,\dots,N}$ to be explained by a statistical model**
  - Parametrization by $q \in \mathbb{R}^n$: individual likelihoods $P(x_i|q)$
  - Prior $\rho(q)$ on the parameters
  - Sample q from $\nu(dq) = \mathrm{e}^{-V(q)}\,dq = Z_\nu^{-1}\rho(q)\prod_{i=1}^{N} P(x_i|q)\,dq$
  - Usual MCMC: each step costs $\mathrm{O}(N)$ $\rightarrow$ prohibitive for $N \gg 1$
- **Mini-batching**:
  - Sample $\mathcal{N}$ data points with replacement: $J_{\mathcal{N}} \in \{1,\dots,N\}^{\mathcal{N}}$
  - Unbiased stochastic estimator of ∇V

$$\nabla(\ln\rho)(q) + \frac{N}{\mathcal{N}}\sum_{j\in J_{\mathcal{N}}}\nabla_q(\ln P(x_j|q)) = -\nabla V(q) + \frac{N}{\sqrt{\mathcal{N}}}\mathcal{G}$$

  - Non-Gaussian noise statistics for $\mathcal{N}$ small
  - for $1 \ll \mathcal{N} \ll N$, it holds $\mathcal{G} \sim \mathcal{N}(0,\Sigma(q))$ with $\Sigma \in \mathbb{R}^{n\times n}$

Mini-batching and Langevin dynamics

- Overdamped Langevin $dq_t = -\nabla V(q_t)\,dt + \sqrt{2}\,dW_t$, discretization

$$q^{n+1} = q^n - \Delta t \nabla V(q^n) + \sqrt{2\Delta t}\, G^n$$

- With mini-batching (Stochastic gradient Langevin dynamics[1])

$$q^{n+1} = q^n - \Delta t \nabla V(q^n) + \frac{N\Delta t}{\sqrt{\mathcal{N}}}\mathcal{G}^n + \sqrt{2\Delta t}\, G^n$$

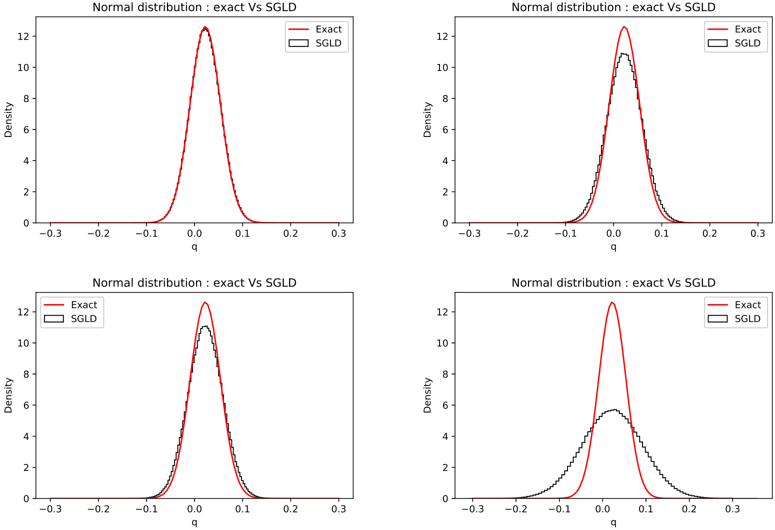
- Amounts to introducing an additional Brownian motion of unknown magnitude $\rightarrow$ bias

$$dq_t = -\nabla V(q_t)\,dt + \sqrt{2 + \frac{N^2\Delta t}{\mathcal{N}}\Sigma(q)}\, d\widetilde{W}_t$$

- Bias remains with underdamped/kinetic Langevin dynamics

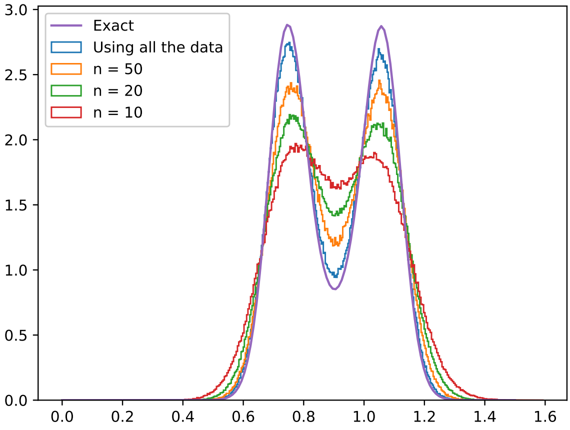
[1] Welling/Teh, *ICML* (2011)

Numerical evidence of the bias (1)



A posteriori distribution of the mean for a Gaussian distribution with Gaussian prior ($N = 1000$). Left: $\Delta t = 10^{-4}$. Right: $\Delta t = 10^{-3}$. Top: without mini-batching. Bottom: with.

Numerical evidence of the bias (2)



Mixture of two Gaussians, with $q = (\theta_1, \theta_2)$ (fixed variances and weights). Marginal distribution of the Gaussian centers ($N = 100$) for SGLD with $\Delta t = 10^{-3}$.

Adaptive Langevin dynamics

(Underdamped) Langevin dynamics

- Phase-space $\mathcal{E} = \mathcal{D} \times \mathbb{R}^d$, Hamiltonian $H(q,p) = V(q) + \frac{1}{2} p^T M^{-1} p$

Stochastic perturbation of the Hamiltonian dynamics

$$\begin{cases} dq_t = M^{-1} p_t \, dt \\ dp_t = -\nabla V(q_t)\, dt - \gamma M^{-1} p_t \, dt + \sqrt{\frac{2\gamma}{\beta}}\, dW_t \end{cases}$$

- Given (known) friction $\gamma > 0$ (could be a position-dependent matrix)
- Various ergodicity results (including exponential convergence of the law)
- Generator of the Langevin dynamics $\mathcal{L} = \mathcal{L}_{\mathrm{ham}} + \gamma \mathcal{L}_{\mathrm{FD}}$

$$\mathcal{L}_{\mathrm{ham}} = p^T M^{-1} \nabla_q - \nabla V^T \nabla_p, \qquad \mathcal{L}_{\mathrm{FD}} = -p^T M^{-1} \nabla_p + \frac{1}{\beta} \Delta_p$$

- Invariant proba. measure $\mu(dq\,dp) = Z^{-1} \mathrm{e}^{-\beta H(q,p)}\, dq\, dp = \nu(dq)\, \kappa(dp)$

Removing the mini-batching bias

- Assume constant $\Sigma(q)$ [see poster of Inass Sekkat...], variable friction ζ

Adaptive Langevin dynamics[1]: unknown σ (scalar, for simplicity)

$$\begin{aligned} dq &= M^{-1} p \, dt, \\ dp &= \left(-\nabla V(q) - \zeta M^{-1} p\right) dt + \sigma \, dW_t, \\ d\zeta &= \frac{1}{m} \left(p^T M^{-2} p - \beta^{-1} \mathrm{Tr}\left(M^{-1}\right)\right) dt \end{aligned}$$

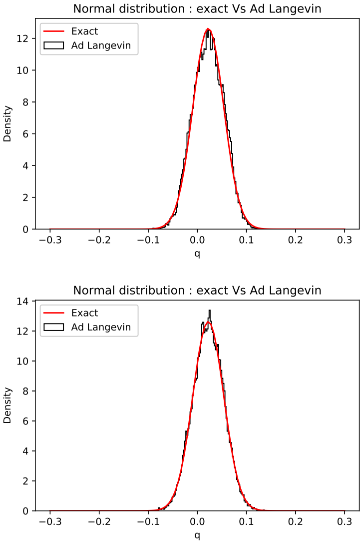
- Invariant measure π with density proportional to

$$\exp\left(-\beta\left[\frac{p^T M^{-1} p}{2} + V(q) + \frac{m}{2}\left(\zeta - \frac{\beta\sigma^2}{2}\right)^2\right]\right) dq \, dp \, d\zeta$$

- The marginal of π in q is indeed ν whatever σ... Prove convergence, in particular Central Limit Theorem?

[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)

Numerical evidence of the absence of bias



Left: $\Delta t = 10^{-4}$. Right: $\Delta t = 10^{-2}$. Top: without mini-batching. Bottom: with.

Adaptive Langevin dynamics

• Normalization of the dynamics, for the invariant measure to be independent of m (take $M = \mathrm{Id}$ to simplify)

$$\begin{cases} dq_t = p_t \, dt, \\ dp_t = \left(-\nabla V(q_t) - \zeta_t p_t\right) dt + \sigma \, dW_t, \\ d\zeta_t = \dfrac{1}{m}\left(|p_t|^2 - \dfrac{n}{\beta}\right) dt \end{cases}$$

• Set $\varepsilon = \sqrt{m}$ and $\zeta = \gamma + \dfrac{\xi}{\varepsilon}$ with $\gamma = \beta\sigma^2/2$

Normalized Adaptive Langevin dynamics

$$\begin{cases} dq_t = p_t \, dt, \\ dp_t = \left(-\nabla V(q_t) - \dfrac{\xi_t}{\varepsilon} p_t - \gamma p_t\right) dt + \sqrt{\dfrac{2\gamma}{\beta}} dW_t, \\ d\xi_t = \dfrac{1}{\varepsilon}\left(|p_t|^2 - \dfrac{n}{\beta}\right) dt \end{cases}$$

Consistency of Adaptive Langevin dynamics (1)

• Invariant measure π with density $Z^{-1}\exp\left(-\beta\left[\frac{|p|^2}{2}+V(q)+\frac{\xi^2}{2}\right]\right)$

• The invariance of the probability measure π is expressed as: for all test function φ,

$$\int \mathcal{L}_{\mathrm{AdL}}\varphi\, d\pi = 0 = \int \varphi \mathcal{L}_{\mathrm{AdL}}^* 1\, d\pi$$

• Simple computations show that, with adjoints defined on $L^2(\pi)$, namely

$$\int (\partial_z \varphi)\,\phi\, d\pi = \int \varphi\left(\partial_z^* \phi\right)\, d\pi,$$

it holds

$$\begin{aligned}\partial_{q_i}^* &= -\partial_{q_i} + \beta \partial_{q_i} V,\\ \partial_{p_i}^* &= -\partial_{p_i} + \beta p_i,\\ \partial_\xi^* &= -\partial_\xi + \beta\xi\end{aligned}$$

Consistency of Adaptive Langevin dynamics (2)

- Generator $\mathcal{L}_{\mathrm{AdL}} = \mathcal{L}_{\mathrm{ham}} + \gamma \mathcal{L}_{\mathrm{FD}} + \varepsilon^{-1} \mathcal{L}_{\mathrm{NH}}$ with

$$\mathcal{L}_{\mathrm{ham}} = \frac{1}{\beta}\left(\nabla_p^* \nabla_q - \nabla_q^* \nabla_p\right) = \frac{1}{\beta} \sum_{i=1}^{n} \partial_{p_i}^* \partial_{q_i} - \partial_{q_i}^* \partial_{p_i},$$

$$\mathcal{L}_{\mathrm{FD}} = -\frac{1}{\beta} \nabla_p^* \nabla_p = -\frac{1}{\beta} \sum_{i=1}^{n} \partial_{p_i}^* \partial_{p_i},$$

$$\mathcal{L}_{\mathrm{NH}} = \left(|p|^2 - \frac{n}{\beta}\right) \partial_\xi - \xi \, p^T \nabla_p = \frac{1}{\beta^2}\Big((\partial_\xi - \partial_\xi^*) \nabla_p^* \nabla_p + \Delta_p^* \partial_\xi - \Delta_p \partial_\xi^*\Big)$$

- Antisymmetric parts $\mathcal{L}_{\mathrm{ham}}$, $\mathcal{L}_{\mathrm{NH}}$ and symmetric one $\mathcal{L}_{\mathrm{FD}}$

- Invariance follows from $\mathcal{L}_{\mathrm{AdL}}^* 1 = (-\mathcal{L}_{\mathrm{ham}} + \gamma \mathcal{L}_{\mathrm{FD}} - \varepsilon^{-1} \mathcal{L}_{\mathrm{NH}}) 1 = 0$

- Expects $\left(\mathrm{e}^{t \mathcal{L}_{\mathrm{AdL}}} \varphi\right)(q_0, p_0, \xi_0) = \mathbb{E}^{(q_0, p_0, \xi_0)}(\varphi(q_t, p_t, \xi_t)) \xrightarrow[t \to +\infty]{} \mathbb{E}_\pi(\varphi)$

Expected scalings for the convergence of the law

- Generator $\simeq$ superposition of $\mathcal{L}_{\mathrm{ham}} + \gamma \mathcal{L}_{\mathrm{FD}}$ and $\varepsilon^{-1}\mathcal{L}_{\mathrm{NH}} + \gamma \mathcal{L}_{\mathrm{FD}}$
  - Exponential rate of decay $\sim \min(\gamma, \gamma^{-1})$ for the Langevin part
  - Nosé–Hoover-like part rewritten as $\varepsilon^{-1}(\mathcal{L}_{\mathrm{NH}} + \gamma\varepsilon \mathcal{L}_{\mathrm{FD}})$
    $\rightarrow$ suggests rate of decay $\sim \varepsilon^{-1} \min(\gamma\varepsilon, (\gamma\varepsilon)^{-1})$

Exponential convergence of the semigroup

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

$$\forall t \geqslant 0,\ \forall \varphi \in L^2(\pi), \quad \left\| \mathrm{e}^{t\mathcal{L}_{\mathrm{AdL}}} \varphi - \int \varphi \, d\pi \right\|_{L^2(\pi)} \leqslant C \mathrm{e}^{-\lambda_{\varepsilon,\gamma}} \left\| \varphi - \int \varphi \, d\pi \right\|_{L^2(\pi)}$$

with the lower bound $\lambda_{\varepsilon,\gamma} \geqslant \overline{\lambda} \min\left(\gamma, \gamma\varepsilon^2, \frac{1}{\gamma}, \frac{1}{\gamma\varepsilon^2}\right)$. As a consequence,

$$\mathcal{L}_{\mathrm{AdL}}^{-1} = -\int_0^{\infty} \mathrm{e}^{t\mathcal{L}_{\mathrm{AdL}}} \, dt, \quad \left\| \mathcal{L}_{\mathrm{AdL}}^{-1} \right\|_{\mathcal{B}(L_0^2(\pi))} \leqslant \frac{C}{\overline{\lambda}} \max\left(\gamma, \gamma^{-1}, \gamma\varepsilon^2, \gamma^{-1}\varepsilon^{-2}\right).$$

Sharpness of the scaling and elements of proof

- Scaling of resolvent norm sharp in view of specific solutions, *e.g.*

$$\mathcal{L}_{\mathrm{AdL}}\left(\gamma\varepsilon\xi + \frac{|p|^2}{2} - \frac{p^T\nabla V}{\gamma}\right) = -\frac{\xi|p|^2}{\varepsilon} + \frac{p^T\nabla V}{\gamma\varepsilon} - \frac{1}{\gamma}\left(p^T\nabla^2 V p - |\nabla V|^2\right),$$

which shows that $\left\|\mathcal{L}_{\mathrm{AdL}}^{-1}\right\|_{\mathcal{B}(L_0^2(\pi))} \geqslant c\gamma\varepsilon^2$ by choosing $\gamma \gg \varepsilon \gg 1$

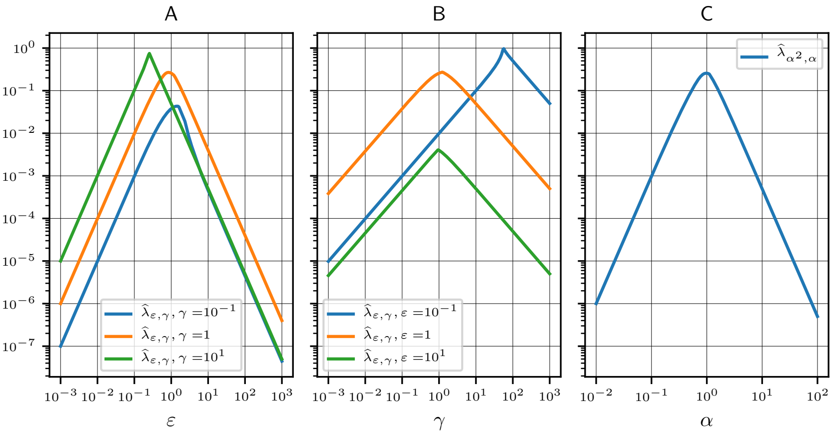
- Proof based on hypocoercive estimates[2,3] with a careful construction of the modified scalar $L^2(\pi)$ product

- Complements proof of exponential decay using Lyapunov techniques[4] (for which convergence rates are not explicit in terms of the parameters)

---

[2] Dolbeault, Mouhot and Schmeiser, *C. R. Math. Acad. Sci. Paris* (2009)
[3] Dolbeault, Mouhot and Schmeiser, *Trans. AMS*, **367**, 3807–3828 (2015)
[4] D. Herzog, *Commun. Math. Sci.* (2018)

Spectral gap in a simple case



Spectral gap computed with a Galerkin method for V quadratic

A: Scaling $\min(\varepsilon^2, \varepsilon^{-2})$ for γ fixed.

B: Scaling $\min(\gamma, \gamma^{-1})$ for ε fixed.

C: Scaling $\min(\alpha^3, \alpha^{-3})$ for $\alpha = \gamma = \varepsilon$.

Central Limit Theorem

- Consider $\varphi \in L^2(\pi)$ and $\overline{\varphi}_t := \dfrac{1}{t}\displaystyle\int_0^t \varphi(q_s, p_s, \xi_s)\, ds$

Central Limit Theorem

$$\sqrt{t}\,(\widehat{\varphi}_t - \mathbb{E}_\pi \varphi) \xrightarrow[t\to+\infty]{\text{law}} \mathcal{N}(0, \sigma^2_{\varepsilon,\gamma}(\varphi)),$$

with the asymptotic variance (with $\Pi_0\varphi = \varphi - \mathbb{E}_\pi(\varphi)$)

$$\sigma^2_{\varepsilon,\gamma}(\varphi) = 2\int \left(-\mathcal{L}_{\mathrm{AdL}}^{-1}\Pi_0\varphi\right)\Pi_0\varphi\, d\pi \leqslant \frac{2C\|\varphi\|^2_{L^2(\pi)}}{\overline{\lambda}} \max\left(\gamma, \gamma^{-1}, \gamma\varepsilon^2, \gamma^{-1}\varepsilon^{-2}\right)$$

- **Suggests taking $\gamma = 1$ and $\varepsilon \sim 1$**
- Langevin type limit $\varepsilon \to +\infty$ for a function $\varphi(q,p)$ (independent of ξ)

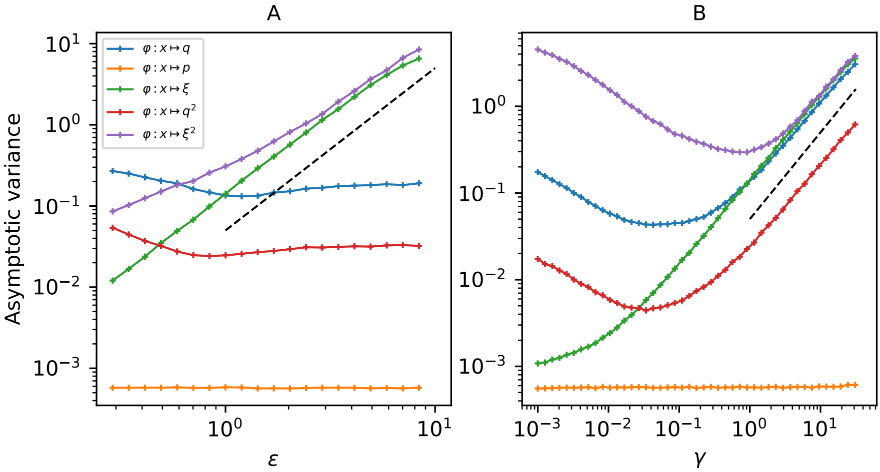
$$\left|\sigma^2_{\varepsilon,\gamma}(\varphi) - \sigma^2_{\infty,\gamma}(\varphi)\right| \leqslant \frac{K}{\varepsilon}$$

Proof: asymptotic analysis and fine estimates[5] of $\mathcal{L}_{\mathrm{Lang}} = \mathcal{L}_{\mathrm{ham}} + \gamma\mathcal{L}_{\mathrm{FD}}$

[5] Talay, *Markov Proc. Rel. Fields* (2002); Kopec, *BIT* (2015)

Scaling of the asymptotic variance

One-dimensional system with simple skewed double-well potential

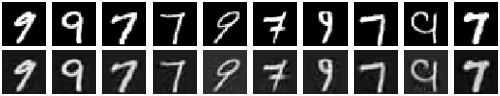


Letft: scaling $\max(1, \varepsilon^2)$ of the variance (γ fixed).
Right: scaling $\max(\gamma, \gamma^{-1})$ of the variance (ε fixed).

Illustration of CLT for MNIST data (1)

• Bayesian logistic regression trained on a subset of the MNIST benchmark data: classify 7 and 9

• Preprocess data: whitening, keep first 100 PCA components ($x^j \in \mathbb{R}^{100}$)



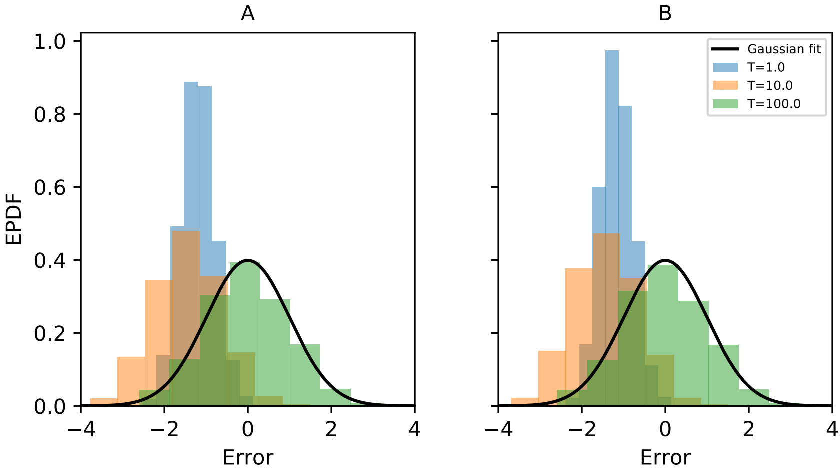
• Weakly informative Gaussian prior on $q \in \mathbb{R}^{100}$, elementary likelihood

$$P(y^j, x^j \mid q) = \frac{\exp\left(y^j (x^j)^T q\right)}{1 + \exp\left((x^j)^T q\right)},$$

where $N = 12,251$ and $y^j \in \{0,1\}$ (0 for 7 and 1 for 9)

• Minibatches of size $\mathcal{N} = 100$, no additional noise, numerical integration by a splitting scheme

Illustration of CLT for MNIST data (2)



Empirical pdf of the rescaled residual error $\sqrt{\frac{K\Delta t}{\widehat{\sigma}^2(\varphi)}}(\widehat{\varphi}_K - \mathbb{E}_\pi(\varphi))$, for $\varphi(q) = q_{65}$ (Left) and $\varphi(q) = q_{65}^2$ (Right).

Current and future tracks

Current and future tracks

- Extension to matrix-valued, q-dependent noises[6]

$$\xi(q) = \sum_{k=0}^{K} A_k f_k(q), \qquad A_k \in \mathbb{R}^{n \times n}$$

for some basis of functions f_k and truncature level K

- High-dimensional space of parameters $n \gg 1$: low-rank representation

[6] I. Sekkat and G. Stoltz, in preparation