ITO DIFFUSION APPROXIMATION OF UNIVERSAL ITO CHAINS FOR SAMPLING, OPTIMIZATION AND BOOSTING

Aleksei Ustimenko
ShareChat, UK
aleksei.ustimenko@sharechat.co

Aleksandr Beznosikov
Innopolis University, Skoltech, RANEPA, Yandex
anbeznosikov@gmail.com

Aleksei Ustimenko, International Conference on Learning Representations 2024

What do we study?

- $$X_{k+1} = X_k + \eta\big(b\big(X_k\big) + \delta_k\big) + \sqrt{\eta^{1+\gamma}}\big(\sigma\big(X_k\big) + \Delta_k\big)\epsilon_k\big(X_k\big),$$
- On the first glance — a generic Markov chain with a drift function $b(x)$ and a covariance $\sigma(x)$
  - Assume those functions are Lipshitz-smooth, not even dissipative
- δ_k, Δ_k — a bias/covariance shift of an order $O(\eta^\alpha)$
- $\gamma \in \{0,1\}$ to cover SGD-like chains in the same analysis
- We want to bound Wasserstain-2 distance to the Itô-process:
- $$\mathrm{d}Z_t = b(Z_t)\mathrm{d}t + \sqrt{\eta^\gamma}\sigma\big(Z_t\big)\mathrm{d}W_t.$$

What do we study?

- $$X_{k+1} = X_k + \eta\big(b\big(X_k\big) + \delta_k\big) + \sqrt{\eta^{1+\gamma}}\big(\sigma\big(X_k\big) + \Delta_k\big)\epsilon_k\big(X_k\big),$$

- The noise-function $\epsilon_k(x)$ — different for every step k and covers non-gaussian state-dependent isotropic noise, possibly non-gaussian

- Assume it satisfies a weak version of the CLT in Wasserstein-2 metric

$$\mathbb{E}_\epsilon \epsilon_k\big(x\big) = 0_d, \quad \mathbb{E}_\epsilon \epsilon_k \epsilon_k^{\mathrm{T}}\big(x\big) = I_d, \quad \mathcal{W}_2^2\Big(\mathcal{N}\big(0_d, SI_d\big), \mathcal{L}\Big(\sum_{k=0}^{S-1} \epsilon_k\big(x\big)\Big)\Big) \leq M_\epsilon^2 \eta^\beta.$$

- Example: any noise of the form $\Sigma(x)\xi, \mathbb{E}\|\xi\|^4 < \infty$

Why such a generic form is of interest?

Table 1: Matching different methods and frameworks with the parameters of the Ito chain equation 1 and Assumption 1.

	Case	γ	α	β	$b(X_k)$	δ_k	$\sigma(X_k)$	Δ_k
Dynamics	GLD	0	∞ [(1)]	∞	$-\nabla_x f(X_k)$	0	$\sqrt{\frac{2}{\tau}} I_d$ [(2)]	0
	SGLD (Gelfand et al., 1992)	0	$\frac{1}{2}$	1	$-\nabla_x f(X_k)$	0	$\sqrt{\frac{2}{\tau}} I_d$	$\sqrt{\frac{2}{\tau} I_d + \eta \mathrm{Cov}(\widehat{\nabla})} - \sqrt{\frac{2}{\tau}} I_d$
	SGLD with smoothing (Chatterji et al., 2020)	0	$\frac{1}{2}$	1	$-\nabla_x f(X_k)$	$\nabla_x \left(f(X_k) - \mathbb{E}_\varepsilon f(X_k + \eta^{\frac{1}{2}} \varepsilon) \right)$	$\sqrt{\frac{2}{\tau}} I_d$	$\sqrt{\frac{2}{\tau} I_d + \eta \mathrm{Cov}(\widehat{\nabla})} - \sqrt{\frac{2}{\tau}} I_d$
Optimization	SGD (Robbins & Monro, 1951)	1	∞	0	$-\nabla_x f(X_k)$	0	$\sqrt{\mathrm{Cov}(\nabla)}$	0
	SGDA (Dem'yanov & Pevnyi, 1972)	1	∞	0	$\left(-\nabla_x f(X_k, Y_k), \nabla_{\mathbf{y}} f(X_k, Y_k) \right)$	0	$\sqrt{\mathrm{Cov}(\nabla)}$	0
	SA-FP (Bailion et al., 1978)	1	∞	0	$F(X_k) - X_k$	0	$\sqrt{\mathrm{Cov}(\widehat{F})}$	0
	SA (Bailion et al., 1978)	1	∞	0	$H(X_k) - a$ [(3)]	0	$\sqrt{\mathrm{Cov}(\widehat{H})}$	0
Boosting	SGB (Friedman, 2001)	1	∞	0	$-P(X_k)\nabla_x f(X_k)$	0	$\sqrt{\mathrm{Cov}(\widehat{\nabla})}$	0
	SGLB (Ustimenko & Prokhorenkova, 2021) [(4)]	0	$\frac{1}{2}$	0	$-P(X_k)\nabla_x f(X_k)$	0	$\sqrt{\frac{2}{\tau}} I_d$	$\sqrt{\eta \mathrm{Cov}(\widehat{\nabla})}$
	SGLB-O (Ustimenko & Prokhorenkova, 2021) [(5)]	0	$\frac{1}{4}$	0	$-P_\infty \nabla_x f(X_k)$	$(P_\infty - P(X_k))\nabla_x f(X_k)$	$\sqrt{\frac{2}{\tau}} I_d$	$\sqrt{\eta \mathrm{Cov}(\widehat{\nabla})}$

[(1)] η^∞ means that the terms multiplied by it vanish, i.e., we can take α as large as we desire when calculating overall approximation error.

[(2)] τ refers to inverse diffusion temperature.

[(3)] a is any constant. Stochastic Approximation tries to solve $H(x) = a$.

[(4)] SGLB here is defined as in the original paper, but here we ignore smoothing applied to the trees selection algorithm.

[(5)] "O" stands for "original", i.e., as presented in the original paper. In that case, such coefficients appear if we take the distribution of trees as in the paper.

How it compares with prior works?

Table 2: Comparison of the theoretical setups and results on Markov chains and diffusions analysis.

	Noise		Generator, i.e. $b(\cdot)$			
Reference	Distribution	**Dependence**	**Non-convex**	**Non-dissipative**	**Non-uniformly elliptic**	$\mathcal{W}_2$
(Raginsky et al., 2017)	$\mathcal{N}$+SG	✓	✓	✗	✓	✓
(Dalalyan, 2017)	$\mathcal{N}$	✗	✗	✗	✗	✓
(Cheng et al., 2018)	$\mathcal{N}$	✗	✓	✗	✗	✗
(Erdogdu et al., 2018)	$\mathcal{N}$+SG	✓	✓	✗	✗	✗
(Durmus & Moulines, 2019)	$\mathcal{N}$	✗	✗	✗	✗	✓
(Ma et al., 2019)	$\mathcal{N}$	✗	✓	✗	✗	✗
(Li et al., 2019b)	$\mathcal{N}$	✓	✗	✗	✓	✓
(Chatterji et al., 2020)	$\mathcal{N}$	✗	✗	✗	✗	✓
(Feng et al., 2019)	$\|\epsilon\| \leq$ const a.s.	✗	✓	✗	✗	✗
(Orvieto & Lucchi, 2018)	$\mathcal{N}$	✗	✓	✗	✗	✗
(Ankirchner & Perko, 2021)	$\mathcal{N}$	✗	✓	✗	✓	✗
(Hu et al., 2017)	$\mathcal{N}$	✗	✓	✗	✗	✗
(Xie et al., 2021)	$\mathcal{N}$	✗	✗	✗	✗	✗
(Ustimenko & Prokhorenkova, 2021)	Mixture $\mathcal{N}$	✓	✓	✗	✓	✗
(Cheng et al., 2020)	$\mathcal{N}$+SG	✓	✓	✗	✗	✗
(Li et al., 2019a)	$\mathbb{E}\|\epsilon\|^4 \leq$ const	✓	✓	✗	✓	✗
Ours	$\mathbb{E}\|\epsilon\|^4 \leq$ const	✓	✓	✓	✗	✓

Table 3: Comparison of guarantees on discretization error in our work with the literature.

	Noise	**Reference**	**Rate**
SGLD	Gaussian	(Muzellec et al., 2020)	$\mathcal{O}(\eta^{\frac{1}{4}})$
		Ours	$\mathcal{O}(\eta^{\frac{1}{4}})$
	any	Ours	$\mathcal{O}(\eta^{\frac{1}{4}})$
SGD	Gaussian	(Cheng et al., 2020) (1)	$\mathcal{O}(\eta^{\frac{1}{8}})$
		Ours	$\mathcal{O}(\eta^{\frac{3}{4}})$
	any	(Cheng et al., 2020) (1)	$\mathcal{O}(\eta^{\frac{1}{8}})$
		Ours	$\mathcal{O}(\eta^{\frac{1}{2}})$
SGB	any	Ours	$\mathcal{O}(\eta^{\frac{1}{2}})$
SGLB	any	Ours	$\mathcal{O}(\eta^{\frac{1}{4}})$

(1) $\mathcal{W}_1$ - distance

High level idea

- optimal transport $(\sqrt{S}\zeta_k^S(x), \epsilon_k^S(x)) \sim \Pi_*\Big(\mathcal{N}(0_d, SI_d),\, \mathcal{L}\big(\epsilon_k^S(x)\big)\Big),$
- chain with Gaussian noise $Y^X_{\overline{\eta}(k+1)} = Y^X_{\overline{\eta}k} + \overline{\eta}\, b\big(Y^X_{\overline{\eta}k}\big) + \underbrace{\sqrt{\overline{\eta}\,\eta^\gamma}\,\sigma\big(Y^X_{\overline{\eta}k}\big)\zeta_k^S\big(X_{Sk}\big)}_{\text{coupling via noise}},$
  - **it won't work as we don't assume dissipativity** of $b(x)$!
  - "regularize it": $Y^X_{\overline{\eta}(k+1)} = Y^X_{\overline{\eta}k} + \overline{\eta} b(Y^X_{\overline{\eta}k}) + \sqrt{\overline{\eta}\eta^\gamma}\sigma(Y^X_{\overline{\eta}k})\zeta_k^S\big(X_{Sk}\big) \underbrace{-L\overline{\eta}(Y^X_{\overline{\eta}k} - X_{Sk})}_{\text{Window coupling}}.$
  - Interpolate $\mathrm{d}Y_t = \big(b(Y_t) + (g_t^* + g_t^S)\big)\mathrm{d}t + \sqrt{\eta^\gamma}\big(\sigma(Y_t) + \Sigma_t^*\big)\mathrm{d}W_t,$
  - Yet another process to prepare for the Girsanov theorem:

$$\mathrm{d}Z_t^Y = (b(Z_t^Y) + g_t^S)\mathrm{d}t + \underbrace{L_1(Y_t - Z_t^Y)\mathrm{d}t}_{\text{Window coupling}} + \sqrt{\eta^\gamma}\sigma(Z_t^Y)\mathrm{d}W_t,$$

  - Use a weaker version of the Girsanov theorem, which we proved in the work

Main results

Theorem 1 (One-time Girsanov formula for mixed Ito/adapted coefficients)**.** *Assume that* $(G_t)_{t\geq 0}$ *is a* $(W_t)_{t\geq 0}$*-adapted process with an integratable by* $t \geq 0$ *second moment. Consider two SDEs ran using two independent Brownian processes:*

$$\begin{aligned} \mathrm{d}Z_t &= b(Z_t)\mathrm{d}t + G_t\mathrm{d}t + \sigma(Z_t)\mathrm{d}W_t, \\ \mathrm{d}Z_t^* &= b(Z_t^*)\mathrm{d}t \qquad\quad + \sigma(Z_t^*)\mathrm{d}\widetilde{W}_t. \end{aligned}$$

Let $\sigma_0 > 0$ *be the minimal possible eigenvalue of* $\sigma(x)$*. Then we have that for any time horizon* $T \geq 0$ *the following bound holds:*

$$D_{\mathrm{KL}}\big(\mathcal{L}(Z_T)\big|\big|\mathcal{L}(Z_T^*)\big) \leq {\sigma_0}^{-2} \int_0^{\mathrm{T}} \mathbb{E}\big\|G_t\big\|^2 \mathrm{d}t < \infty.$$

Theorem 2. *Let Assumption 1 hold. Then for all* $k \in \mathbb{N}_0$*:*

$$\mathcal{W}_2\big(\mathcal{L}(X_{k'}), \mathcal{L}(Z_{k'\eta})\big) = \mathcal{O}\Big(\big(1 + (k'\eta)^{\frac{1}{2}}\big)e^{\mathcal{O}(k'\eta)}\eta^{\theta} + (k'\eta)^{\frac{1}{4}}e^{\mathcal{O}(k'\eta)}\eta^{\frac{\theta}{2}+\frac{\gamma}{4}}\Big),$$

where $\theta = \min\left\{\alpha; \frac{(\gamma+1)(1+\chi_0)+(\gamma+\beta)(1-\chi_0)}{4}\right\}$.

Corollary 5 (SGD with Gaussian noise)**.** *Under the conditions of Theorem 2, if* $\chi_0 = 1$*,* $\gamma = 1$*,* $\alpha \geq 1$*, then* $\theta = 1$ *and* $\mathcal{W}_2\big(\mathcal{L}(X_k), \mathcal{L}(Z_{k\eta})\big) = \mathcal{O}\left(\eta^{\frac{3}{4}}\right)$.

Thank you