Penalized Langevin dynamics with vanishing penalty for smooth and log-concave targets

Avetik Karagulyan Arnak Dalalyan

CREST, IP Paris

We study the problem of sampling from a probability distribution on $\mathbb{R}^p$ defined via a convex and smooth potential function. We first consider a continuous-time diffusion-type process, termed Penalized Langevin dynamics (PLD), the drift of which is the negative gradient of the potential plus a linear penalty that vanishes when time goes to infinity. An upper bound on the Wasserstein-2 distance between the distribution of the PLD at time t and the target is established. This upper bound highlights the influence of the speed of decay of the penalty on the accuracy of approximation. As a consequence, considering the low-temperature limit we infer a new non-asymptotic guarantee of convergence of the penalized gradient flow for the optimization problem.

Introduction

- Our goal is to sample from a given target distribution π defined on $\mathbb{R}^p$ with a large value of p. The latter means, that for a given precision level ε, we want to generate a random vector $\boldsymbol{\theta}$ with values in $\mathbb{R}^p$ such that its distribution μ satisfies

$$\text{distance}(\mu, \pi) \le \varepsilon, \tag{1}$$

for some probability measure distance.
- Important particular case: π has a density (w.r.t. the Lebesgue measure) given by

$$\pi(\theta) \propto \exp(-f(\theta)), \tag{2}$$

with a ``potential'' $f : \mathbb{R}^p \to \mathbb{R}$.

Notation

Main assumption $A(m, M)$. We say that f satisfies $A(m, M)$, if it is twice differentiable and the following matrix inequalities are true:

- m-strong convexity: $\nabla^2 f(\boldsymbol{\theta}) \succeq m I_p$, with $m \ge 0$; (π is m-strongly log-concave)
- M-Lipschitz gradients: $\nabla^2 f(\boldsymbol{\theta}) \preceq M I_p$, with $M > 0$.

Wasserstein distance:

$$W_q(\nu, \nu') = \inf \left\{ \mathbb{E}[\|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}'\|_2^q]^{1/q} : \boldsymbol{\vartheta} \sim \nu \text{ and } \boldsymbol{\vartheta}' \sim \nu' \right\}, \tag{3}$$

where the infimum is over all joint distributions having ν and ν' as the first and the second marginal distributions.

Langevin diffusion

- Vanilla Langevin diffusion:

$$d\boldsymbol{L}_t^{\text{LD}} = -\nabla f(\boldsymbol{L}_t^{\text{LD}})dt + \sqrt{2} d\boldsymbol{W}_t. \tag{LD}$$

The solution of this equation is a Markov process having π as an invariant distribution.
- When the potential function f is m-strongly convex, the Markov process is ergodic and it converges to π exponentially (Villani 2008):

$$W_2(\nu_t^{\text{LD}}, \pi) \le e^{-mt} W_2(\nu_0^{\text{LD}}, \pi). \tag{4}$$

- In the non-strongly convex case this classical result does not provide convergence for LD.

LD + Poincaré inequality

Poincaré inequality: We say π satisfies the Poincaré inequality, if for $\forall g \in L^2(\pi)$ locally-Lipschitz, we have

$$\text{var}_\pi[g] \le C_{\mathsf{P}} \mathbb{E}_\pi \left[\|\nabla g\|^2 \right]. \tag{P}$$

Example: Log-concave distributions satisfy this inequality. Furthermore, $C_{\mathsf{P}} = 1/m$, for m-strongly log-concave distributions.
Chewi et al (2020) have shown that if π satisfies (P), then

$$W_2^2(\mu_t, \pi) \le 2C_{\mathsf{P}} e^{-\frac{2t}{C_{\mathsf{P}}}} \chi^2(\mu_0 \| \pi). \tag{5}$$

The convergence is exponential but the result is not explicit, as the Poincaré constant is unknown. Kannan, Lovász and Simonovits (1995) conjectured that C_{P} is bounded universally in terms of the dimension p. Chen (2020) almost proved it by showing that $C_{\mathsf{P}} = O(p^{o(1)})$, but the result remains asymptotic.

Penalized Langevin diffusion

We propose to modify the Langevin equation by adding a vanishing linear penalty:

$$d\boldsymbol{L}_t^{\text{PLD}} = -\left(\nabla f(\boldsymbol{L}_t^{\text{PLD}}) + \alpha(t)\boldsymbol{L}_t^{\text{PLD}}\right) dt + \sqrt{2}\, d\boldsymbol{W}_t, \tag{6}$$

where $\alpha : [0, \infty) \to [0, \infty)$ is a positive time-dependent penalty factor converging to zero as $t \to \infty$. The idea is to have a $\alpha(t)$-strongly convex potential function $f(\cdot) + \alpha(t)\|\cdot\|^2/2$, for every fixed time instant t.

Main theorem

Theorem. Suppose f satisfies $\mathrm{A}(0, M)$. Then, for every positive number t and for $\beta(t) = \int_0^t \alpha(u)\, du$, we have

$$W_2(\nu_t^{\text{PLD}}, \pi) \le \sqrt{\mu_2}\, e^{-\beta(t)} + 11\mu_2 e^{-\beta(t)} \int_0^t \frac{|\alpha'(s)| e^{\beta(s)}}{\sqrt{\alpha(s)}} ds + \sqrt{\alpha(t)} \mu_2,$$

where $\mu_2 = \mathbb{E}_\pi[\|\boldsymbol{\vartheta}\|^2]$.

Remarks:

- We obtain polynomial convergence.
- The result only depends on the second order moment.
- The almost optimal upper bound that this result can provide, is obtained when $\alpha(t) = O(1/(\mu_2 + t))$. In this case,

$$W_2(\nu_t^{\text{PLD}}, \pi) \le \frac{10\mu_2 \left[1 + \log\left(1 + (t/\mu_2)\right)\right]}{\sqrt{t + \mu_2}}. \tag{7}$$

Kinetic Langevin diffusion

Kinetic Langevin diffusion is a system of two SDEs and it is the origin of LD. It was first proposed to model the movement of a particle in an environment with friction.

$$d\boldsymbol{L}_t^{\text{KLD}} = \boldsymbol{V}_t^{\text{KLD}} dt;$$
$$d\boldsymbol{V}_t^{\text{KLD}} = -(\eta \boldsymbol{V}_t^{\text{KLD}} + \nabla f(\boldsymbol{L}_t^{\text{KLD}}))dt + \sqrt{2\eta} d\boldsymbol{W}_t.$$

- Here W_t is a Brownian motion and η is the friction parameter. LD is the limit of the rescaled kinetic diffusion $\bar{\boldsymbol{L}}_t = \boldsymbol{L}_{\eta t}^{\text{KLD}}$ when the friction coefficient η tends to infinity (Nelson et al, '65).
- The Markov process $(\boldsymbol{L}_t^{\text{KLD}}; \boldsymbol{V}_t^{\text{KLD}})$ is positive recurrent and its invariant distribution is absolutely continuous w.r.t. the Lebesgue measure on $\mathbb{R}^{2p}$. The corresponding invariant density is

$$p_*(\boldsymbol{\theta}, \boldsymbol{v}) \propto \exp(-f(\boldsymbol{\theta}) - \|\boldsymbol{v}\|_2^2/2). \tag{8}$$

- Mixing-time in Wasserstein distance is of order $\exp(-t/\kappa)$, with $\kappa = M/m$. (Eberle et al., 2017)

Penalized Kinetic Langevin diffusion

Similar to the Vanilla Langevin diffusion, the Kinetic Langevin diffusion also encounters the issue of convergence in the non-strongly convex case. We propose to use the same penalization for KLD, which results the following system of SDEs:

$$d\boldsymbol{L}_t^{\text{PKLD}} = \boldsymbol{V}_t^{\text{PKLD}} dt;$$
$$d\boldsymbol{V}_t^{\text{PKLD}} = -(\eta \boldsymbol{V}_t^{\text{PKLD}} + \nabla f(\boldsymbol{L}_t^{\text{PKLD}}) + \alpha(t)\boldsymbol{L}_t^{\text{PKLD}})dt + \sqrt{2\eta} d\boldsymbol{W}_t.$$

Remark. We prove that PKLD converges to the target distribution π with $\tilde{O}(1/\sqrt{\mu_2 + t})$ rate.

Contributions and future work

Our main contributions can be summarized as follows:

- We propose a general penalization method that can be applied to LD, KLD and many other SDEs. In particular, for LD and KLD we prove explicit non-asymptotic upper bounds for the Wasserstein-2 error.
- In the paper, we also analyze the analogical problem of gradient flows. Leveraging the similarity of the sampling and optimization problems, we prove the convergence of the gradient flow to the minimum point of f under some technical assumptions.