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

Avetik Karagulyan

MMS 2020, “Luminy”, France

General setting of the problem

- **Problem:** sample from a given target distribution π defined in $\mathbb{R}^p$ with a large value of p.
- More precisely, for a given precision level ϵ, generate a random vector $\boldsymbol{\theta}$ with values in $\mathbb{R}^p$ such that its distribution μ satisfies

$$\text{distance}(\mu, \pi) \leq \epsilon. \tag{1}$$

- 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" $f : \mathbb{R}^p \to \mathbb{R}$.

General setting of the problem

Main assumption A(m, M):

- $f \in C^2(\mathbb{R}^p)$
- m-strong convexity: $\nabla^2 f(\boldsymbol{\theta}) \succeq m I_p$, with $m \geq 0$;
- M-Lipschitz gradients: $\nabla^2 f(\boldsymbol{\theta}) \preceq M I_p$, with $M > 0$.

Wasserstein distance:

$$W_2(\nu, \nu') = \inf \left\{ \mathbf{E}[\|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}'\|_2^2]^{1/2} : \boldsymbol{\vartheta} \sim \nu \text{ and } \boldsymbol{\vartheta}' \sim \nu' \right\}, \qquad (3)$$

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**:

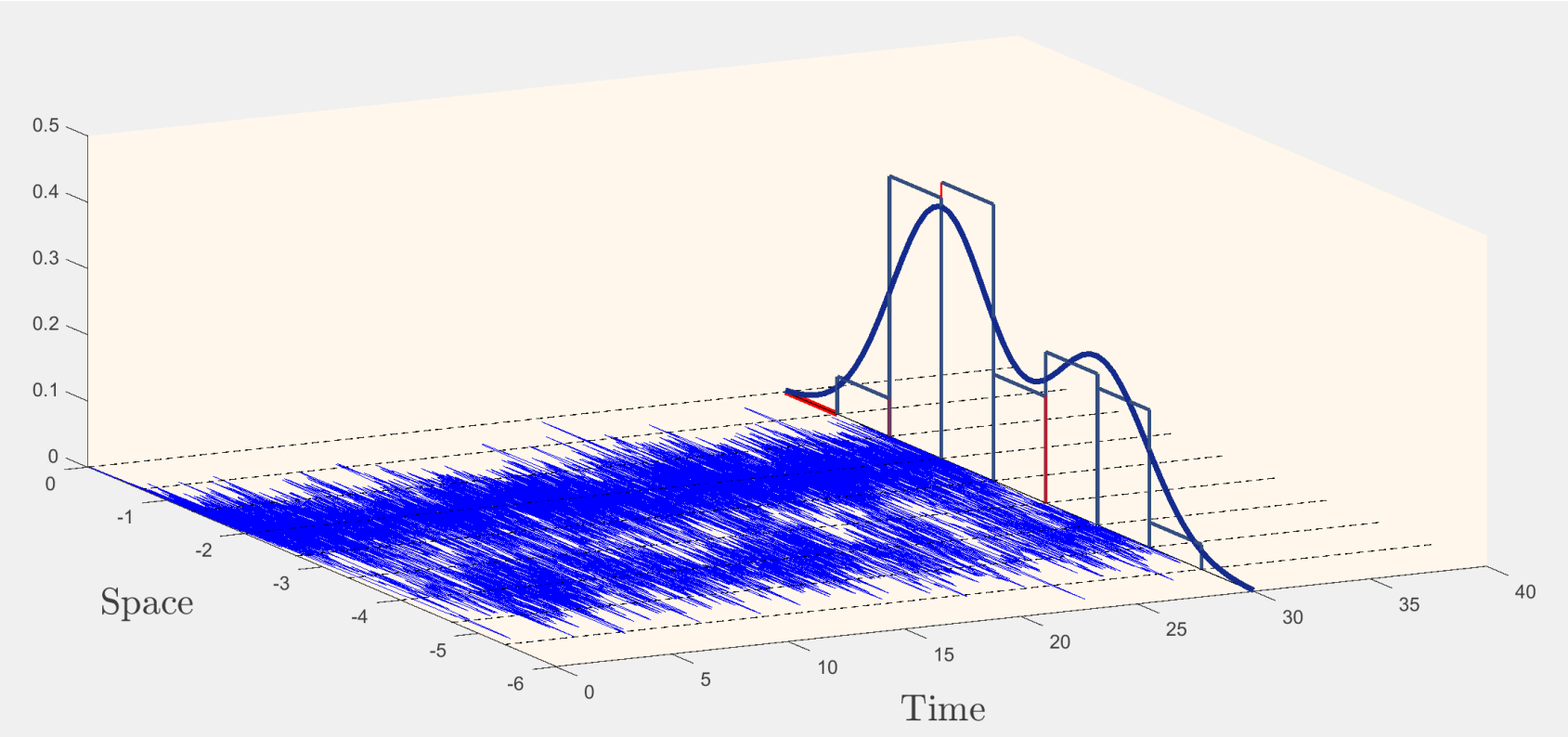
$$d\boldsymbol{L}_t^{\mathsf{LD}} = -\nabla f(\boldsymbol{L}_t^{\mathsf{LD}})dt + \sqrt{2}d\boldsymbol{W}_t. \tag{LD}$$

  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):

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

Illustration of LD



The blue lines represent different paths of the Langevin process. We see that the histogram of the state at time $t = 30$ is close to the target density.

Langevin Monte-Carlo

Langevin Monte-Carlo (LMC) is defined as:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - h\nabla f(\boldsymbol{\theta}_k) + \sqrt{2h}\,\boldsymbol{\xi}_{k+1}; \qquad k = 0, 1, 2, \ldots \tag{5}$$

$(\boldsymbol{\xi}_k)_{k\in\mathbb{N}}$ is a sequence of i.i.d standard Gaussians ind. from $\boldsymbol{\theta}_k$.

Theorem (Durmus et al, 2018)

Suppose f satisfies the assumption A(m, M) with $m > 0$. If $h < 1/M$, then the following upper bound is satisfied:

$$W_2\left(\mu_K, \pi\right) \leq (1 - mh)^{K/2} W_2\left(\mu_0, \pi\right) + \{2hp\kappa\}^{1/2} \tag{6}$$

where μ_K is the law of $\boldsymbol{\theta}_K$ and $\kappa = M/m$ is the condition number.

Remarks on the theorem:

- The term $(1 - mh)^{K/2} W_2(\mu_0, \pi)$ can be viewed as an approximation error.
- The term $\{2hp\kappa\}^{1/2}$ is the error due to the discretization.
- We have at our disposal the step-size h and the time horizon $T = Kh$. Therefore choosing

$$h \leq \frac{\varepsilon^2}{p\kappa} \quad \text{and} \quad K = \frac{p\kappa}{m\varepsilon^2} \log(2/\varepsilon) \tag{7}$$

we have $W_2(\mu_K, \pi) < \varepsilon$.

Kinetic Langevin Dynamics

Kinetic (underdamped) Langevin diffusion:

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

- Here W_t is a Brownian motion and η is the friction parameter. (LD) is the limit of the rescaled kinetic diffusion $\bar{\boldsymbol{L}}_t = \boldsymbol{L}_{\eta t}^{\mathsf{KLD}}$ when the friction coefficient η tends to infinity (Nelson et al, '65).
- The Markov process $(\boldsymbol{L}_t^{\mathsf{KLD}}; \boldsymbol{V}_t^{\mathsf{KLD}})$ is positive recurrent and its invariant distribution is absolutely continuous wrt 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 $\exp(-t/\kappa)$. (Eberle et al., 2017)

Kinetic Langevin Monte-Carlo

KLMC is a discretization of Kinetic LD:

$$\boldsymbol{v}_{k+1} = \psi_0(h)\boldsymbol{v}_k - \psi_1(h)\nabla f(\boldsymbol{\theta}_k) + \sqrt{2\eta}\boldsymbol{\xi}_{k+1};$$
$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \psi_1(h)\boldsymbol{v}_k - \psi_2(h)\nabla f(\boldsymbol{\theta}_k) + \sqrt{2\eta}\boldsymbol{\xi}'_{k+1}.$$

- where $\psi_0(x) = \exp(-\eta x)$ and $\psi_{k+1}(x) = \int_0^x \psi_k(x)dx$.
- $(\boldsymbol{\xi}_k, \boldsymbol{\xi}'_k)_{k\in\mathbb{N}}$ is a sequence of standard Gaussians independent from the initial conditions.

Theorem (Dalalyan and Riou-Durand, '18)

Assume that f satisfies A(m, M), with $m > 0$. Then for $\eta > \sqrt{m+M}$ and $h < m/4\eta M$, the following bound is true:

$$W_2(\nu_k, \pi) \leq \sqrt{2}\left(1 - \frac{0.75mh}{\eta}\right)^k W_2(\nu_0, \pi) + \frac{Mh\sqrt{2p}}{m}.$$

Remarks on KLMC

- From the theorem we deduce that in order to have ε Wasserstein sampling error it suffices to do

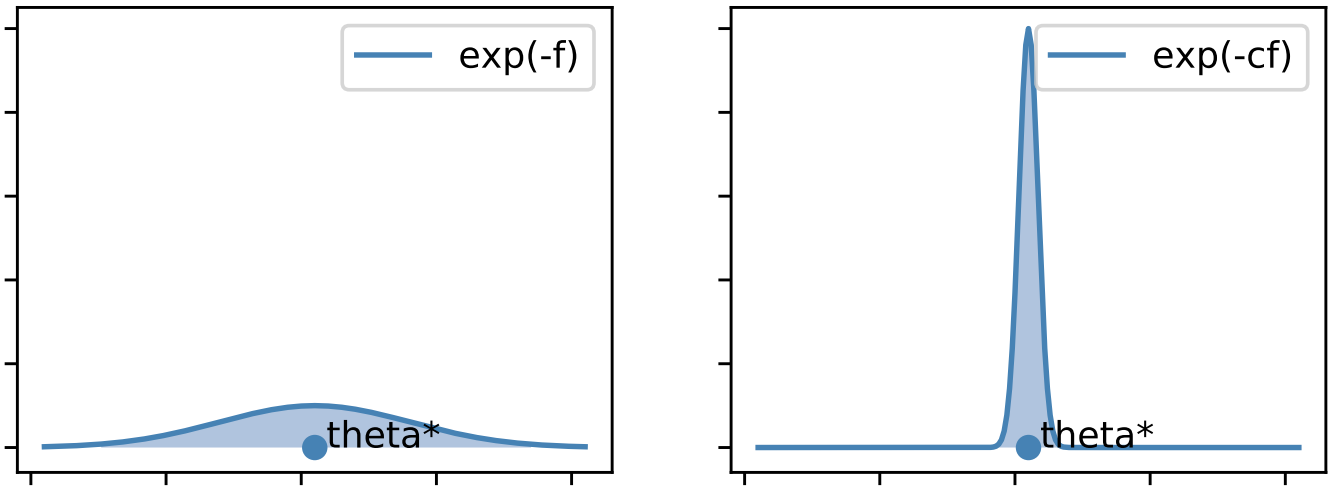
$$K_{\mathrm{KLMC}} = \tilde{O}\left(\kappa^{3/2}\left\{\kappa\bigvee\frac{p}{m\varepsilon^2}\right\}^{1/2}\right) \tag{9}$$

  iterations of KLMC.

- Compared to LMC, KLMC has better convergence rate in terms of p and ε. This is due to smoother trajectories of KLD. (Cheng et al, '17).

- Nevertheless, LMC and KLMC are not comparable. In the case when ∇f increases rapidly, LMC performs better.

Relation with optimization



- Suppose we can have exact samples from $\pi \propto \exp(-f)$. Then we can also sample from $\pi^c \propto \exp(-cf)$. When c is large, the samples are concentrated around the minimum point θ^*.

Relation with optimization

- In the case of non-convex f, one can use LMC for the minimization problem. The Gaussian noise term allows to escape local minima.
- The gradient flow of the function f is defined as the following differential equation:

$$d\boldsymbol{L}_t^{\mathsf{GF}} = -\nabla f(\boldsymbol{L}_t^{\mathsf{GF}})dt. \tag{10}$$

GF can be seen as a limit version of LD scaled over time axis, when the scaling parameter tends to infinity. (discussed later)

Relation with optimization

We define the following functionals for every $\mu \in \mathcal{P}_2(\mathbb{R}^p)$:

- **Energy functional:** $\mathcal{E}(\mu) := \mathbb{E}_\mu[f]$.
- **Gibbs-Boltzmann functional:** $\mathcal{H}(\mu) := \mathbb{E}_\mu[\log(d\mu/d\lambda)]$, where λ is the Lebesgue measure.

It is known that π minimizes $\mathcal{F} := \mathcal{H} + \mathcal{E}$, and that the Langevin diffusion is its gradient flux.

What happens when f is convex but not strongly?

Discrete case

- (Durmus et al, 2018) have shown that LMC with averaging converges:

$$\mathrm{KL}(\nu'_K|\pi) \leq \varepsilon, \qquad \text{when} \qquad K = O\left(\frac{\kappa p}{\varepsilon^2}\right), \qquad (11)$$

  where ν'_K is the mixture of the first K iterates of the LMC. **The idea:** leverage the connection between Langevin dynamics and optimization of KL divergence (Jordan et al. '98)(not Michael Jordan).
- (Dalalyan, K. Riou-Durand, 2019) proposed to use constant linear penalty in LMC and KLMC. **The idea:** sample from the adjusted density $\exp(-f(\boldsymbol{\theta}) - \alpha\|\boldsymbol{\theta}\|_2^2/2)$ and optimize over α. Convergence in Wasserstein-1 and Wasserstein-2 distance are obtained.

LD + Poincaré inequality

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

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

Chewi et al (2020) have shown that if π satisifes (P), then

$$W_2^2(\mu_t, \pi) \leq 2C_{\mathsf{P}} e^{-\frac{2t}{C_{\mathsf{P}}}} \chi^2(\mu_0 \| \pi). \tag{12}$$

The convergence is exponential but the constants can be large.

KLS conjecture

C_{p} of any distribution π is bounded by some universal constant C_{KLS} times the largest eigenvalue of the covariance matrix of π:

$$C_{\mathsf{p}} \leq C_{\mathsf{KLS}} \times \mathbb{E}_{\pi} \left[\| \boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} \|_{\mathrm{op}} \right]. \tag{13}$$

Penalized Langevin Dynamics

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

$$d\boldsymbol{L}_t^{\mathsf{PLD}} = -\big(\nabla f(\boldsymbol{L}_t^{\mathsf{PLD}}) + \alpha(t)\boldsymbol{L}_t^{\mathsf{PLD}}\big)\,dt + \sqrt{2}\,d\boldsymbol{W}_t, \tag{14}$$

where $\alpha : [0, \infty) \to [0, \infty)$ is a positive time-dependent penalty factor converging to zero as $t \to \infty$.

- For every fixed time instant t the potential is equal to $f(\cdot) + \alpha(t)\|\cdot\|^2/2$, which is $\alpha(t)$-strongly convex.

Convergence of PLD

Define $\mu_2 := \mathbb{E}_\pi[\|\boldsymbol{\theta}\|^2]$.

Proposition

Suppose f satisfies A(0, M). If $\alpha(t) = 1/(2\mu_2 + 2t)$, then

$$W_2(\nu_t^{\mathsf{PLD}}, \pi) \leq \frac{10\mu_2\left[1 + \log\left(1 + (t/\mu_2)\right)\right]}{\sqrt{t + \mu_2}}, \tag{15}$$

Remark:

- We obtain polynomial convergence.
- The result only depends on the second order moment.
- In the m-strongly convex case $\mu_2 \leq p/m$.

Main theorem

Theorem

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

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

- The error bound stated in the previous proposition is obtained by "optimizing" this upper bound w.r.t. α.

Sketch of the proof

We apply the triangle inequality to $W_2(\nu_t^{\mathsf{PLD}}, \pi)$:

$$W_2\left(\nu_t^{\mathsf{PLD}}, \pi\right) \leq W_2\left(\nu_t^{\mathsf{PLD}}, \pi_{\alpha(t)}\right) + W_2\left(\pi_{\alpha(t)}, \pi\right), \tag{16}$$

where $\pi_\gamma(\boldsymbol{\theta}) \propto \exp(-f(\boldsymbol{\theta}) - \gamma\|\boldsymbol{\theta}\|^2/2)$, for $\forall\gamma > 0$. We then define $\phi(t) := W_2\left(\nu_t^{\mathrm{PLD}}, \pi_{\alpha(t)}\right)$ and prove the following bound:

$$\phi'(t) \leq -\alpha(t)\phi(t) - \frac{11\alpha'(t)}{\sqrt{\alpha(t)}} \cdot \mu_2\left(\pi_{\alpha(t)}\right). \tag{17}$$

Applying Gronwall inequality to ϕ and transportation inequality to the term $W_2\left(\pi_{\alpha(t)}, \pi\right)$ we obtain the proof.

Tempered PLD

- Define $f_\tau(\cdot) := f(\cdot)/\tau$, $\pi^\tau \propto \exp(-f_\tau)$ and

$$\boldsymbol{L}_t^\tau = \boldsymbol{L}_0^\tau - \frac{1}{\tau}\int_0^t \left(\nabla f\left(\boldsymbol{L}_s^\tau\right) + \alpha(s/\tau)\boldsymbol{L}_s^\tau\right) ds + \sqrt{2}\boldsymbol{W}_t.$$

- We know that $\boldsymbol{L}_t^\tau \to \pi^\tau$. Tempered PLD is defined as $\boldsymbol{X}^{\mathsf{TPLD}} := \boldsymbol{L}_{\tau t}^\tau$ and it satisfies

$$d\boldsymbol{X}_t^{\mathsf{TPLD}} = -\left(\nabla f\left(\boldsymbol{X}_t^{\mathsf{TPLD}}\right) + \alpha(t)\boldsymbol{X}_t^{\mathsf{TPLD}}\right) dt + \sqrt{2\tau} d\boldsymbol{W}_t. \tag{TPLD}$$

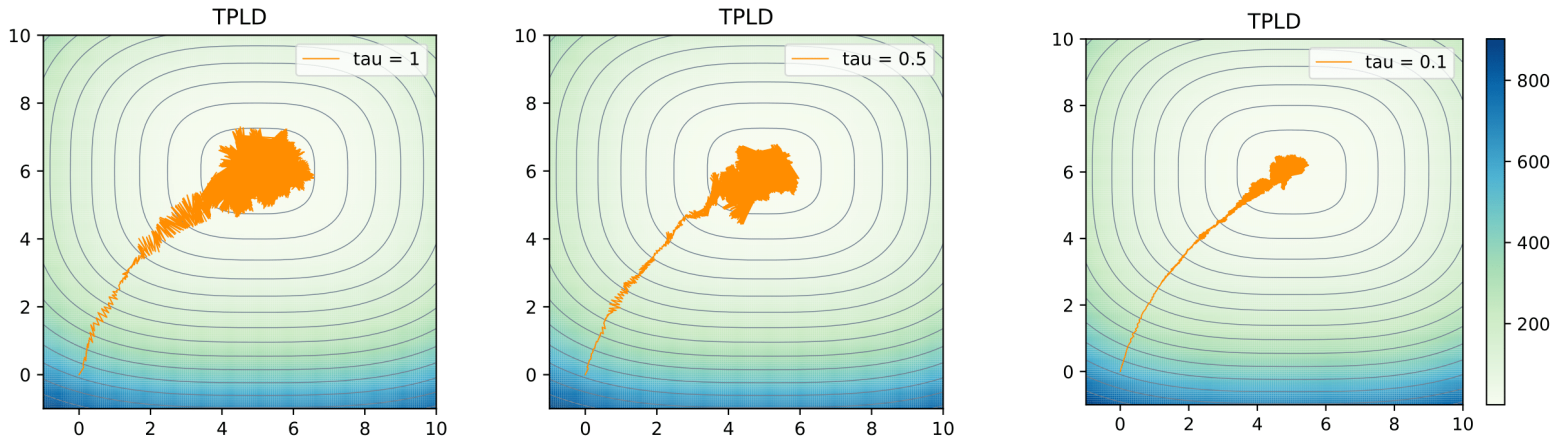


Figure: Plots of TPLD for different temperature parameter.

- Sampling $\xrightarrow{\tau \to 0}$ optimization and (TPLD) $\xrightarrow{\tau \to 0}$ (PGF):

$$d\boldsymbol{X}_u^{\mathsf{PGF}} = -\left(\nabla f\left(\boldsymbol{X}_u^{\mathsf{PGF}}\right) + \alpha(u)\boldsymbol{X}_u^{\mathsf{PGF}}\right) du. \qquad \text{(PGF)}$$

The convergence of PGF

Technical assumption B(D,q)**:**

$$\|\boldsymbol{x}_\gamma - \boldsymbol{x}_{\tilde{\gamma}}\|_2 \leq \frac{\mathrm{D}}{\tilde{\gamma}^{\mathrm{q}}}(\tilde{\gamma} - \gamma)\,\|\boldsymbol{x}_*\|_2\,, \quad \forall \gamma < \tilde{\gamma} \tag{18}$$

where $\boldsymbol{x}_\gamma := \arg\min\{f(\boldsymbol{x}) + \gamma\|\boldsymbol{x}\|^2/2\}$.

Theorem

If f *satisfies* $\mathrm{A}(0, \mathrm{M})$ *and* $\mathrm{B}(\mathrm{D}, \mathrm{q})$*, then taking* $\alpha(t) = \frac{(1-\mathrm{q})}{(t+A)}$ *we have*

$$\left\|\boldsymbol{X}_t^{\mathrm{PGF}} - \boldsymbol{\theta}_*\right\|_2 \leq \frac{A^{1-\mathrm{q}} + \mathrm{D} + \mathrm{D}\log(1 + (\mathrm{t}/\mathrm{A}))}{(t+A)^{1-\mathrm{q}}}\,\|\boldsymbol{\theta}_*\|_2\,. \tag{19}$$

Conclusion and outlook

Take home message

- Penalized Langevin Dynamics gives explicit non-asymptotic bounds on the Wasserstein error of sampling.

Future work

- Assessing the error of the penalized kinetic Langevin diffusion in continuous-time.
- Discretization of the time-continuous dynamics.
- Gain understanding on the condition required in the theorem for the PGF.

This is the last slide.
Thank you!