


## Homework 3

### (Variational Inference, Markov Chain Monte Carlo)

For questions, please refer to Moodle.  
Released on 25/10/2023

#### GENERAL INSTRUCTIONS

- Submission of solutions is not mandatory but solving the exercises are highly recommended. The master solution will be released after the exercise deadline.
- Part of the exercises are available on Moodle as a quiz. These problems are marked with .

## Exercise 0: Reparameterizable Distributions

In this exercise, we will show that the reparameterization trick is not only useful for Gaussians, but also for other distributions.

(a) Let  $X \sim \text{Unif}([a, b])$  for any  $a \leq b$ . That is,

$$p_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{otherwise.} \end{cases}$$

Show that  $X$  can be reparameterized in terms of  $\text{Unif}([0, 1])$ .

*Hint: You may use that for any  $Y \sim \text{Unif}([a, b])$  and  $c \in \mathbb{R}$ ,*

- $Y + c \sim \text{Unif}([a + c, b + c])$  and
- $cY \sim \text{Unif}([c \cdot a, c \cdot b])$ .

(b) Let  $X$  be a random variable such that  $\log X \sim \mathcal{N}(\mu, \sigma^2)$ . Such a random variable  $X$  is said to be logarithmically normal distributed with parameters  $\mu$  and  $\sigma^2$ . Show that  $X$  can be reparameterized in terms of  $\mathcal{N}(0, 1)$ .

(c) Let  $X \sim \text{Cauchy}(\mu, \tau)$  be a random variable that follows a Cauchy distribution with location  $\mu$  and scale  $\tau$ . The CDF of a Cauchy distribution is defined as

$$P_X(x) = \frac{1}{\pi} \arctan\left(\frac{x - \mu}{\tau}\right) + \frac{1}{2}$$

Show that  $\text{Cauchy}(0, 1)$  can be reparameterized in terms of  $\text{Unif}([0, 1])$ .

## Exercise 1: Jensen & Gibbs Inequalities, Maximum Entropy Principle

As discussed in tutorial 2, one property of Gaussians that makes them omnipresent is the following: For a fixed mean  $\mu$  and variance  $\sigma^2$ , the distribution that has maximum entropy among all distributions that are supported on  $\mathbb{R}$  is the normal distribution. In this Exercise, we prove this property using the Kullback-Leibler (KL) divergence.

Let  $g(x) = \mathcal{N}(x; \mu, \sigma^2)$ , and  $f(x)$  a distribution with the same mean  $\mu$  and variance  $\sigma^2$ . We suppose that  $f(x)$  is supported on  $\mathbb{R}$ :  $f(x) > 0$  ( $\forall x \in \mathbb{R}$ ).

- (a) (Bonus) Prove Jensen's inequality in finite form. That is, given a convex function  $f$ , show that

$$f(\theta_1 x_1 + \dots + \theta_k x_k) \leq \theta_1 f(x_1) + \dots + \theta_k f(x_k)$$

for all  $x_1, \dots, x_k$  and  $\theta_1, \dots, \theta_k \geq 0$  with  $\theta_1 + \dots + \theta_k = 1$ .

- (b) Show that KL-divergence is non-negative.

- (c) Prove that:

$$KL(f||g) = H(g) - H(f).$$

Where  $H$  is the entropy function. ( $H(X) = \mathbb{E}[-\log(f(X))] = -\int_{\mathcal{X}} f(x) \log(f(X)) dx$ )

*Hint: Equivalently, show that  $H(f||g) = H(g)$ . That is, the expected surprise evaluated based on the Gaussian  $g$  is invariant to the true distribution  $f$  (as long as  $f$  has full support).*

- (d) Conclude that  $H(g) \geq H(f)$ .

## Exercise 2: Gaussian Process Classification

In the lecture you have learned about Gaussian processes, which are a generalization of Bayesian linear regression with a Gaussian prior and a Gaussian likelihood. Crucially, to perform closed-form inference with GPs, we exploited the closedness and conjugacy properties of Gaussian distributions.

In this exercise, we will study the use of Gaussian processes for classification tasks, commonly called *Gaussian process classification* (GPC). In the lecture, you have also been introduced to linear logistic regression, which corresponds to Bayesian linear regression with a Gaussian prior and a Bernoulli likelihood. Linear logistic regression is extended to GPC by replacing the Gaussian prior over weights with a GP prior on  $f$ ,

$$f \sim GP(0, k), \quad \pi(x) = \sigma(f(x)), \quad y | x, f \sim \text{Bern}(\pi(x))$$

where  $\sigma : \mathbb{R} \rightarrow (0, 1)$  is a logistic-type function. Note that Bayesian logistic regression is the special case where  $k$  is the linear kernel and  $\sigma$  is the logistic function. This is analogous to the relationship of Bayesian linear regression and Gaussian process regression which you have studied in exercise 2 of homework 2.

In the GP regression setting that you were introduced to in the lecture,  $y_i$  was assumed to be a noisy observation of  $f(x_i)$ . In the classification setting, we now have that  $y_i \in \{-1, 1\}$  is a binary class label and  $f(x_i) \in \mathbb{R}$  is a latent value.

We study the setting where  $\sigma(z) = \Phi(z; 0, \sigma_n^2)$  where  $\Phi(z; 0, \sigma_n^2)$  is the CDF of a univariate Gaussian with mean 0 and variance  $\sigma_n^2$ , also called a *probit likelihood*.

To make probabilistic predictions for a query point  $x^*$ , we first compute the distribution of the latent variable  $f^*$ ,

$$p(f^* | x_{1:n}, y_{1:n}, x^*) = \int p(f^* | x_{1:n}, x^*, f) p(f | x_{1:n}, y_{1:n}) df \quad (1)$$

where  $p(f | x_{1:n}, y_{1:n})$  is the posterior over the latent variables.

- (a) Assuming that we can efficiently compute  $p(f^* | x_{1:n}, y_{1:n}, x^*)$  (approximately), describe how we can produce the probabilistic prediction  $p(y^* = 1 | x_{1:n}, y_{1:n}, x^*)$ .
- (b) The integral of the latent predictive posterior (1) is analytically intractable, as we used a non-Gaussian likelihood. A common technique is to approximate the latent posterior  $p(f | x_{1:n}, y_{1:n})$  with a Gaussian using a Laplace approximation  $q := \mathcal{N}(\hat{f}, \Lambda^{-1})$ . It is not possible to obtain an analytical representation of the mode of the Laplace approximation  $\hat{f}$ . Instead,  $\hat{f}$  is commonly found using a second-order optimization scheme such as Newton's method.

- 1) Find the precision matrix  $\Lambda$  of the Laplace approximation.

*Hint: Observe that for a label  $y_i \in \{-1, 1\}$ , the probability of a correct classification given the latent value  $f_i$  is  $p(y_i | f_i) = \sigma(y_i f_i)$ , where we use the symmetry of the probit likelihood around 0.*

2) Assume that  $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{x}'$  is the linear kernel ( $\sigma_p = 1$ ) and that  $\sigma$  is the logistic function (cf. eq. (5.10) in the lecture notes). Show for this setting that the matrix  $\mathbf{\Lambda}$  derived in (1) is equivalent to the precision matrix  $\mathbf{\Lambda}'$  of the Laplace approximation of Bayesian logistic regression (cf. eq. (5.18) in the lecture notes).<sup>1</sup> You may assume that  $\hat{f}_i = \hat{\mathbf{w}}^\top \mathbf{x}_i$ .

*Hint 1:* Note that  $\mathbf{\Lambda}'$  is a precision matrix over weights  $\mathbf{w}$  and  $\mathbf{f} = \mathbf{X}\mathbf{w}$ , so the corresponding variance over latent values  $\mathbf{f}$  is  $\mathbf{X}\mathbf{\Lambda}'^{-1}\mathbf{X}^\top$ . The two precision matrices are therefore equivalent if  $\mathbf{\Lambda}^{-1} = \mathbf{X}\mathbf{\Lambda}'^{-1}\mathbf{X}^\top$ .

*Hint 2:* You may find the matrix inversion lemma (see below) and Woodbury's matrix identity useful.

3) Determine the mean and variance of the latent predictive posterior  $p(\mathbf{f}^* \mid \mathbf{x}_{1:n}, \mathbf{y}_{1:n}, \mathbf{x}^*)$  using the Laplace approximation of the latent posterior.

*Hint 1:* You may use the matrix inversion lemma, stating that for matrices  $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$ ,

$$(\mathbf{A} + \mathbf{B})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}(\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1}\mathbf{A}^{-1}. \quad (2)$$

*Hint 2:* Condition on the latent variables  $\mathbf{f}$  using the laws of total expectation and variance.

4) Are the prediction  $p(y^* = 1 \mid \mathbf{x}_{1:n}, \mathbf{y}_{1:n}, \mathbf{x}^*)$  you obtained in (a) (but now using the Laplace-approximated latent predictive posterior) and the prediction  $\sigma(\mathbb{E}_{q(\mathbf{f}^* \mid \mathbf{x}_{1:n}, \mathbf{y}_{1:n}, \mathbf{x}^*)}[\mathbf{f}^*])$  identical? If not, describe in words how are they different.

(c) So far, we have assumed that the latent Gaussian process is noise-free and we have combined it with the probit likelihood function.

Show that using a noise-free latent process combined with the probit likelihood  $\Phi(z; 0, \sigma_n^2)$  is equivalent (in expectation) to using a latent process with Gaussian noise  $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$  combined with the step-function likelihood

$$h(z) := \begin{cases} 1 & \text{if } z \geq 0 \\ 0 & \text{if } z < 0. \end{cases} \quad (3)$$

*Hint:* Introduce explicitly noisy latent variables  $\tilde{f}_i$ , which differ from  $f_i$  by the Gaussian noise  $\epsilon_i$ . Write down the step-function likelihood for a single case as a function of  $\tilde{f}_i$ .

---

<sup>1</sup>This should not be surprising since — as already mentioned — Gaussian process classification is a generalization of Bayesian logistic regression.

### Exercise 3: Markov Chain, Stationary Distribution

- (a) [✓] TheSpoon is a restaurant rating company that annually rates restaurants around the world. It classifies restaurants into three categories: "poor", "satisfactory" and "good". No restaurant moved from "poor" to "good" in one year. However, a 5% of the restaurants get downgraded from "good" to "poor". 20% of the restaurants in the "poor" category become "satisfactory". While 10% of those in the "satisfactory" category get upgraded to "good", 30% become "poor"; 25% of those in the "good" category are downgraded to "satisfactory". In the long run, what percentage of restaurants will be classified as "good" by TheSpoon?
- (b) [✓] Alice gambles such that their winning probability in each round of gambling is  $p$ ;  $p \in (0, 1)$ . Assume that Alice has wealth  $w_t$  at time  $t$ . Winning increases and losing decreases Alice's wealth by 1 each. They continue to gamble until their wealth reaches 0 or  $k$ ; for a fixed  $k \in \mathbb{N}$ . We construct a Markov chain by considering  $\{w_t\}_{t \in \mathbb{N}_0}$  over state space  $[0, k]$  and with the described transition probabilities. Is this chain ergodic?
- (c) [✓] Consider the state space  $S = \{00, 01, 10, 11\}$  of binary strings having length 2. Let  $p(i|j) = 0.5$  if  $i$  differs from  $j$  in exactly one bit, and  $p(i|j) = 0$  otherwise. Provide the transition matrix  $P$  for this distribution. We consider a Markov chain with transition probability matrix  $P$ . Is this chain ergodic?

## Exercise 4: Gibbs Sampling

- (a) Let  $p(x)$  be a probability density over  $\mathbb{R}^d$ , which we want to sample from. Assume that  $p$  is a Gibbs distribution, that is, its PDF can be expressed as

$$p(x) = \frac{1}{Z} \exp(-f(x))$$

where  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  is some energy function.

In this exercise, we will study a single round of Gibbs sampling with initial state  $x$  and final state  $x'$  where

$$x'_j = \begin{cases} x_j & \text{if } j \neq i \\ x'_i & \text{otherwise.} \end{cases}$$

for some fixed index  $i$  and  $x'_i \sim p(\cdot \mid x_{-i})$ .

Show that

$$\mathbb{E}_{x'_i \sim p(\cdot \mid x_{-i})} [f(x')] \leq f(x) - S[p(x_i \mid x_{-i})] + H[p(\cdot \mid x_{-i})]$$

where  $S[u] = -\log u$  denotes the surprise of an event of probability  $u$ . That is, the energy is expected to decrease if the surprise of  $x_i$  given  $x_{-i}$  is larger than the expected surprise of the new  $x'_i$  given  $x_{-i}$ , i.e.,  $S[p(x_i \mid x_{-i})] \geq H[p(\cdot \mid x_{-i})]$ .<sup>2</sup>

(Hint: Recall the framing of Gibbs sampling as a variant of Metropolis-Hastings and relate this to the acceptance distribution of Metropolis-Hastings when  $p$  is a Gibbs distribution.)

In the remaining two questions, we look at some examples where Gibbs sampling is useful.

- (b) [✓] Consider the following generative model  $p(\mu, \lambda, x_{1:n})$  given by the likelihood  $x_{1:n} \mid \mu, \lambda \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \lambda^{-1})$  and the priors

$$\mu \sim \mathcal{N}(\mu_0, \lambda_0^{-1}) \quad \text{and} \quad \lambda \sim \text{Gamma}(\alpha, \beta).$$

We would like to sample from the posterior  $p(\mu, \lambda \mid x_{1:n})$ . Show that

$$\mu \mid \lambda, x_{1:n} \sim \mathcal{N}(m_\lambda, l_\lambda^{-1}) \quad \text{and} \quad \lambda \mid \mu, x_{1:n} \sim \text{Gamma}(a_\mu, b_\mu),$$

and derive  $m_\lambda, l_\lambda, a_\mu, b_\mu$ . Such a prior is called a *semi-conjugate prior* to the likelihood, as the prior on  $\mu$  is conjugate for any fixed value of  $\lambda$  and vice-versa.

- (c) The *Pareto distribution* was originally used to model the distribution of wealth in a society, but is also used to model many other phenomena such as the size of cities, the frequency of words, and the returns on stocks. Formally, the Pareto distribution is defined by the following PDF,

$$\text{Pareto}(x; \alpha, c) = \frac{\alpha c^\alpha}{x^{\alpha+1}} \mathbb{1}\{x \geq c\}, \quad x \in \mathbb{R} \quad (5)$$

where the *tail index*  $\alpha > 0$  models the “weight” of the right tail of the distribution and  $c > 0$  corresponds to a cutoff threshold. The distribution is supported on  $[c, \infty)$ , and as  $\alpha \rightarrow \infty$  it approaches a point density at  $c$ .

Let us assume that  $x_{1:n} \mid \alpha, c \stackrel{\text{iid}}{\sim} \text{Pareto}(\alpha, c)$  and assume the improper prior  $p(\alpha, c) \propto \mathbb{1}\{\alpha, c > 0\}$  which essentially corresponds to a uniform prior (i.e., “no prior”). Derive the posterior  $p(\alpha, c \mid x_{1:n})$ . Then, also derive the conditional distributions  $p(\alpha \mid c, x_{1:n})$  and  $p(c \mid \alpha, x_{1:n})$ , and observe that they correspond to known distributions / are easy to sample from.

<sup>2</sup>Recall that the entropy  $H(p)$  is the expected surprise of samples from  $p$ ,  $H(p) = \mathbb{E}_{x \sim p}[S[p(x)]]$ .

## Exercise 5: Mixing Time of Langevin Dynamics (\*)

(\*): This exercise goes beyond the technical tools covered in the lecture. We introduce all required concepts here, so that this exercise is solvable with knowledge from the lecture. Yet, the topics such as stochastic differential equations and vector calculus will not be tested in the exam and are not preliminaries for the successful completion of this course.

In the lecture, you have been introduced to the *unadjusted Langevin algorithm* (ULA) which for a given learning rate  $\eta > 0$  selects points

$$\theta_{k+1} = \theta_k - \eta \nabla f(\theta_k) + \epsilon \quad (6)$$

where  $\epsilon \sim \mathcal{N}(\mathbf{0}, 2\eta I)$ . It can be seen that ULA is the discrete-time (Euler) approximation with step size  $\eta$  of the following stochastic differential equation (SDE)

$$d\theta_t = -\nabla f(\theta_t) dt + \sqrt{2} d\mathbf{W}_t \quad (7)$$

where  $\{\mathbf{W}_t\}_{t \geq 0}$  denotes the Brownian motion in  $\mathbb{R}^n$  with  $\mathbf{W}_0 = \mathbf{0}$ .<sup>3</sup> The continuous-time process described by this SDE is called *Langevin dynamics*.

In this exercise, we will show that for certain Gibbs distributions  $p(\theta) \propto \exp(-f(\theta))$ , Langevin dynamics is rapidly mixing. That is, Langevin dynamics converges “quickly” to the distribution  $p$ . To do this, we will observe that Langevin dynamics can be seen as a continuous-time optimization algorithm in the space of distributions. The same techniques that we study in this exercise can then also be used to derive analogous convergence guarantees for ULA.

First, we consider a simpler and more widely-known optimization algorithm, namely the *gradient flow*

$$dx_t = -\nabla f(x_t) dt. \quad (8)$$

Note that gradient descent is simply the discrete-time approximation of gradient flow just as ULA is the discrete-time approximation of Langevin dynamics. In the analysis of ordinary differential equations (ODEs) such as the gradient flow, so-called *Lyapunov functions* are commonly used to prove convergence of  $x_t$  to a fixed point (also called an *equilibrium*).

Let us assume that  $f$  is  $\alpha$ -strongly convex for some  $\alpha > 0$ , that is,

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\alpha}{2} \|y - x\|_2^2 \quad \forall x, y \in \mathbb{R}^n. \quad (9)$$

In words,  $f$  is lower bounded by a quadratic function with curvature  $\alpha$ . Moreover, assume w.l.o.g. that  $f$  is minimized at  $f(\mathbf{0}) = 0$ .<sup>4</sup>

(a) Show that  $f$  satisfies the *Polyak-Łojasiewicz (PL) inequality*, that is,

$$f(x) \leq \frac{1}{2\alpha} \|\nabla f(x)\|_2^2 \quad \forall x \in \mathbb{R}^n. \quad (10)$$

(b) Prove  $\frac{d}{dt} f(x_t) \leq -2\alpha f(x_t)$ .

Thus,  $\mathbf{0}$  is the fixed point of eq. (8) and the Lyapunov function  $f$  is monotonically decreasing along the trajectory of  $x_t$ . We recall *Grönwall's inequality* which states that for any real-valued continuous functions  $g(t)$  and  $\beta(t)$  on the interval  $[0, T] \subset \mathbb{R}$  such that  $\frac{d}{dt} g(t) \leq \beta(t)g(t)$  for all  $t \in [0, T]$  we have

$$g(t) \leq g(0) \exp\left(\int_0^t \beta(s) ds\right) \quad \forall t \in [0, T]. \quad (11)$$

<sup>3</sup>For more details, see remark 6.29 in the lecture notes.

<sup>4</sup>This can always be achieved by shifting the coordinate system and subtracting a constant from  $f$ .

(c) Conclude that  $f(x_t) \leq e^{-2\alpha t} f(x_0)$ .

Now that we have proven the convergence of gradient flow using  $f$  as Lyapunov function, we will follow the same template to prove the convergence of Langevin dynamics to the distribution  $p(\theta) \propto \exp(-f(\theta))$ . We will use that the evolution of  $\{\theta_t\}_{t \geq 0}$  following the SDE from eq. (7) is equivalently characterized by their densities  $\{q_t\}_{t \geq 0}$  following the *Fokker-Planck equation*

$$\frac{\partial q_t}{\partial t} = \nabla \cdot (q_t \nabla f) + \Delta q_t. \quad (12)$$

Here,  $\nabla \cdot$  and  $\Delta$  are the divergence and Laplacian operators, respectively.<sup>5</sup> Intuitively, the first term of the Fokker-Planck equation corresponds to the drift and its second term corresponds to the diffusion (i.e., the Gaussian noise).

**Intuition:** Recall that the divergence  $\nabla \cdot \mathbf{F}$  of a vector field  $\mathbf{F}$  measures the change of volume under the flow of  $\mathbf{F}$ . That is, if in the small neighborhood of a point  $x$ ,  $\mathbf{F}$  points towards  $x$ , then the divergence at  $x$  is negative as the volume shrinks. If  $\mathbf{F}$  points away from  $x$ , then the divergence at  $x$  is positive as the volume increases.

The Laplacian  $\Delta \varphi = \nabla \cdot (\nabla \varphi)$  of a scalar field  $\varphi$  can be understood intuitively as measuring “heat dissipation”. That is, if  $\varphi(x)$  is smaller than the average value of  $\varphi$  in a small neighborhood of  $x$ , then the Laplacian at  $x$  is positive.

Regarding the Fokker-Planck equation (12), the second term  $\Delta q_t$  can therefore be understood as locally dissipating the probability mass of  $q_t$  (which is due to the diffusion term in the SDE).

On the other hand, the term  $\nabla \cdot (q_t \nabla f)$  can be understood as a Laplacian of  $f$  “weighted” by  $q_t$ . Intuitively, the vector field  $\nabla f$  moves flow in the direction of high energy, and hence, its divergence is larger in regions of lower energy and smaller in regions of higher energy. This term therefore corresponds to a drift from regions of high energy to regions of low energy.

(d) Show that  $\Delta q_t = \nabla \cdot (q_t \nabla \log q_t)$ , implying that the Fokker-Planck equation simplifies to

$$\frac{\partial q_t}{\partial t} = \nabla \cdot \left( q_t \nabla \log \frac{q_t}{p} \right). \quad (13)$$

*Hint: The Laplacian of a scalar field  $\varphi$  is  $\Delta \varphi \doteq \nabla \cdot (\nabla \varphi)$ .*

Observe that the Fokker-Planck equation already implies that  $p$  is indeed a stationary distribution, as if  $q_t = p$  then  $\frac{\partial q_t}{\partial t} = 0$ . Moreover, note the similarity of the integrand of  $\text{KL}(q_t \| p)$ ,  $q_t \log \frac{q_t}{p}$ , to eq. (13). We will therefore use the KL-divergence with respect to  $p$  as the Lyapunov function.

(e) Prove  $\frac{d}{dt} \text{KL}(q_t \| p) = -J(q_t \| p)$ . Here,

$$J(q_t \| p) \doteq \mathbb{E}_{\theta \sim q_t} \left[ \left\| \nabla \log \frac{q_t(\theta)}{p(\theta)} \right\|_2^2 \right] \quad (14)$$

denotes the *relative Fisher information* of  $q_t$  with respect to  $p$ .

*Hint: For any distribution  $q$  on  $\mathbb{R}^n$ ,*

$$\int_{\mathbb{R}^n} (\nabla \cdot q \mathbf{F}) \varphi \, d\mathbf{x} = - \int_{\mathbb{R}^n} q \nabla \varphi \cdot \mathbf{F} \, d\mathbf{x} \quad (15)$$

*follows for any vector field  $\mathbf{F}$  and scalar field  $\varphi$  from the divergence theorem and the product rule of the divergence operator.*

<sup>5</sup>For ease of notation, we omit the explicit dependence of  $q_t$ ,  $p$ , and  $f$  on  $\theta$ .

Thus, the relative Fisher information can be seen as the negated time-derivative of the KL-divergence, and as  $J(q_t \| p) \geq 0$  it follows that the KL-divergence is decreasing along the trajectory.

The *log-Sobolev inequality* (LSI) is satisfied by a distribution  $p$  with a constant  $\alpha > 0$  if for all  $q$ :

$$\text{KL}(q \| p) \leq \frac{1}{2\alpha} J(q \| p). \quad (16)$$

It is a classical result that if  $f$  is  $\alpha$ -strongly convex then  $p$  satisfies the LSI with constant  $\alpha$ .<sup>6</sup>

- (f) Show that if  $f$  is  $\alpha$ -strongly convex for some  $\alpha > 0$  (we say that  $p$  is “ $\alpha$ -strongly log-concave”), then  $\text{KL}(q_t \| p) \leq e^{-2\alpha t} \text{KL}(q_0 \| p)$ .
- (g) Conclude that under the same assumption on  $f$ , Langevin dynamics is rapidly mixing, that is,  $\tau_{\text{TV}}(\epsilon) \in \mathcal{O}(\text{poly}(n, \log(1/\epsilon)))$ .<sup>7</sup>

## References

- [1] Dominique Bakry and Michel Émery. Diffusions hypercontractives. In *Séminaire de Probabilités XIX 1983/84: Proceedings*, pages 177–206. Springer, 2006.

---

<sup>6</sup>Refer to [1]

<sup>7</sup>Refer to remark 6.11 in the lecture notes for the definitions of total variation distance and mixing time  $\tau_{\text{TV}}(\epsilon)$ .