Statistical Estimation of the Poincaré constant and Application to Sampling Multimodal Distributions

Loucas Pillaud-Vivien[1], **Francis Bach**[1], **Tony Lelièvre**[2], **Alessandro Rudi**[1], **Gabriel Stoltz**[2]
[1]INRIA - Ecole Normale Supérieure - PSL Research University,
[2]Université Paris-Est - CERMICS (ENPC) - INRIA

Abstract

Poincaré inequalities are ubiquitous in probability and analysis and have various applications in statistics (concentration of measure, rate of convergence of Markov chains). The Poincaré constant, for which the inequality is tight, is related to the typical convergence rate of diffusions to their equilibrium measure. In this paper, we show both theoretically and experimentally that, given sufficiently many samples of a measure, we can estimate its Poincaré constant. As a by-product of the estimation of the Poincaré constant, we derive an algorithm that captures a low dimensional representation of the data by finding directions which are difficult to sample. These directions are of crucial importance for sampling or in fields like molecular dynamics, where they are called reaction coordinates. Their knowledge can leverage, with a simple conditioning step, computational bottlenecks by using importance sampling techniques.

1 Introduction

Sampling is a cornerstone of probabilistic modelling, in particular in the Bayesian framework where statistical inference is rephrased as the estimation of the posterior distribution given the data [Robert, 2007, Murphy, 2012]: the representation of this distribution through samples is both flexible, as most interesting quantities can be computed from them (e.g., various moments or quantiles), and practical, as there are many sampling algorithms available depending on the various structural assumptions made on the model. Beyond one-dimensional distributions, a large class of these algorithms are iterative and update samples with a Markov chain which eventually converges to the desired distribution, such as Gibbs sampling or Metropolis-Hastings (or more general Markov chain Monte-Carlo algorithms [Gamerman and Lopes, 2006, Gilks et al., 1995, Durmus and Moulines, 2017]) which are adapted to most situations, or Langevin's algorithm [Durmus and Moulines, 2017, Raginsky et al., 2017, Welling and Teh, 2011, Mandt et al., 2017, Lelièvre and Stoltz, 2016, Bakry et al., 2014], which is adapted to sampling from densities in $\mathbb{R}^d$.

While these sampling algorithms are provably converging in general settings when the number of iterations tends to infinity, obtaining good explicit convergence rates has been a central focus of study, and is often related to the mixing time of the underlying Markov chain [Meyn and Tweedie, 2012]. In particular, for sampling from positive densities in $\mathbb{R}^d$, the Markov chain used in Langevin's algorithm can classically be related to a diffusion process, thus allowing links with other communities such as molecular dynamics [Lelièvre and Stoltz, 2016]. The main objective of molecular dynamics is to infer macroscopic properties of matter from atomistic models via averages with respect to probability measures dictated by the principles of statistical physics. Hence, it relies on high dimensional and highly multimodal probabilistic models.

When the density is log-concave, sampling can be done in polynomial time with respect to the dimension [Ma et al., 2018, Durmus et al., 2017, Durmus and Moulines, 2017]. However, in general, sampling with generic algorithms does not scale well with respect to the dimension. Furthermore, the multimodality of the objective measure can trap the iterates of the algorithm in some regions for long durations: this phenomenon is known as metastability. To accelerate the sampling procedure, a common technique in molecular dynamics is to resort to importance sampling strategies where the target probability measure is biased using the image law of the process for some low-dimensional function, known as

Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

"reaction coordinate" or "collective variable". Biasing by this low-dimensional probability measure can improve the convergence rate of the algorithms by several orders of magnitude [Lelièvre et al., 2008, Lelièvre, 2013]. Usually, in molecular dynamics, the choice of a good reaction coordinate is based on physical intuition on the model but this approach has limitations, particularly in the Bayesian context [Chopin et al., 2012]. There have been efforts to numerically find these reaction coordinates Gkeka [2019]. Computations of spectral gaps by approximating directly the diffusion operator work well in low-dimensional settings but scale poorly with the dimension. One popular method is based on diffusion maps [Coifman and Lafon, 2006, Coifman et al., 2006, Rohrdanz et al., 2011], for which reaction coordinates are built by approximating the entire infinite-dimensional diffusion operator and selecting its first eigenvectors.

In order to assess or find a reaction coordinate, it is necessary to understand the convergence rate of diffusion processes. We first introduce in Section 2 Poincaré inequalities and Poincaré constants that control the convergence rate of diffusions to their equilibrium. We then derive in Section 3 a kernel method to estimate it and optimize over it to find good low dimensional representation of the data for sampling in Section 4. Finally we present in Section 5 synthetic examples for which our procedure is able to find good reaction coordinates.

Contributions. In this paper, we make the following contributions:

- We show both theoretically and experimentally that, given sufficiently many samples of a measure, we can estimate its Poincaré constant and thus quantify the rate of convergence of Langevin dynamics.
- By finding projections whose marginal laws have the largest Poincaré constant, we derive an algorithm that captures a low dimensional representation of the data. This knowledge of "difficult to sample directions" can be then used to accelerate dynamics to their equilibrium measure.

2 Poincaré Inequalities

2.1 Definition

We introduce in this part the main object of this paper which is the Poincaré inequality [Bakry et al., 2014]. Let us consider a probability measure $d\mu$ on $\mathbb{R}^d$ which has a density with respect to the Lebesgue measure. Consider $H^1(\mu)$ the space of functions in $L^2(\mu)$ (i.e., which are square integrable) that also have all their first order derivatives in L^2, that is, $H^1(\mu) = \{f \in L^2(\mu),\ \int_{\mathbb{R}^d} f^2 d\mu + \int_{\mathbb{R}^d} \|\nabla f\|^2 d\mu < \infty\}$.

Definition 1 (Poincaré inequality and Poincaré constant)**.** *The Poincaré constant of the probability measure $d\mu$ is the smallest constant $\mathcal{P}_\mu$ such that for all $f \in H^1(\mu)$ the following Poincaré inequality **(PI)** holds:*

$$\int_{\mathbb{R}^d} f(x)^2 d\mu(x) - \left(\int_{\mathbb{R}^d} f(x) d\mu(x)\right)^2 \leqslant \mathcal{P}_\mu \int_{\mathbb{R}^d} \|\nabla f(x)\|^2 d\mu(x). \quad (1)$$

In Definition 1 we took the largest possible and the most natural functional space $H^1(\mu)$ for which all terms make sense, but Poincaré inequalities can be equivalently defined for subspaces of test functions $\mathcal{H}$ which are dense in $H^1(\mu)$. This will be the case when we derive the estimator of the Poincaré constant in Section 3.

Remark 1 (A probabilistic formulation of the Poincaré inequality.)**.** *Let X be a random variable distributed according to the probability measure $d\mu$. **(PI)** can be reformulated as: for all $f \in H^1(\mu)$,*

$$\mathrm{Var}_\mu\left(f(X)\right) \leqslant \mathcal{P}_\mu\, \mathbb{E}_\mu\left[\|\nabla f(X)\|^2\right]. \quad (2)$$

Poincaré inequalities are hence a way to bound the variance from above by the so-called Dirichlet energy $\mathbb{E}\left[\|\nabla f(X)\|^2\right]$ (see [Bakry et al., 2014]).

2.2 Consequences of (PI): convergence rate of diffusions

Poincaré inequalities are ubiquitous in various domains such as probability, statistics or partial differential equations (PDEs). For example, in PDEs they play a crucial role for showing the existence of solutions of Poisson equations or Sobolev embeddings [Gilbarg and Trudinger, 2001], and they lead in statistics to concentration of measure results [Gozlan, 2010]. In this paper, the property that we are the most interested in is the convergence rate of diffusions to their stationary measure $d\mu$. In this section, we consider a very general class of measures: $d\mu(x) = \mathrm{e}^{-V(x)}dx$ (called Gibbs measures with potential V), which allows for a clearer explanation. Note that all measures admitting a positive density can be written like this and are typical in Bayesian machine learning [Robert, 2007] or molecular dynamics [Lelièvre and Stoltz, 2016]. Yet, the formalism of this section can be extended to more general cases [Bakry et al., 2014].

Let us consider the overdamped Langevin diffusion in $\mathbb{R}^d$, that is the solution of the following stochastic differential equation (SDE):

$$\mathrm{d}X_t = -\nabla V(X_t)\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t, \quad (3)$$

where $(B_t)_{t\geqslant 0}$ is a d-dimensional Brownian motion. It is well-known [Bakry et al., 2014] that the law of $(X_t)_{t\geqslant 0}$ converges to the Gibbs measure $d\mu$ and that the Poincaré constant controls the rate of convergence to equilibrium in $L^2(\mu)$. Let us denote by $P_t(f)$ the Markovian semi-group associated with the Langevin diffusion $(X_t)_{t\geqslant 0}$. It is defined in the following way: $P_t(f)(x) = \mathbb{E}[f(X_t)|X_0 = x]$. This semi-group satisfies the dynamics

$$\frac{d}{dt}P_t(f) = \mathcal{L}P_t(f),$$

where $\mathcal{L}\phi = \Delta^L\phi - \nabla V\cdot\nabla\phi$ is a differential operator called the infinitesimal generator of the Langevin diffusion (3) (Δ^L denotes the standard Laplacian on $\mathbb{R}^d$). Note that by integration by parts, the semi-group $(P_t)_{t\geqslant 0}$ is reversible with respect to $d\mu$, that is: $-\int f(\mathcal{L}g)\,d\mu = \int \nabla f\cdot\nabla g\,d\mu = -\int(\mathcal{L}f)g\,d\mu$. Let us now state a standard convergence theorem (see e.g. [Bakry et al., 2014, Theorem 2.4.5]), which proves that $\mathcal{P}_\mu$ is the characteristic time of the exponential convergence of the diffusion to equilibrium in $L^2(\mu)$.

Theorem 1 (Poincaré and convergence to equilibrium)**.** *With the notation above, the following statements are equivalent:*

(*i*) *μ satisfies a Poincaré inequality with constant $\mathcal{P}_\mu$;*

(*ii*) *For all f smooth and compactly supported, $\mathrm{Var}_\mu(P_t(f)) \leqslant \mathrm{e}^{-2t/\mathcal{P}_\mu}\mathrm{Var}_\mu(f)$ for all $t\geqslant 0$.*

Proof. The proof is standard. Note that upon replacing f by $f - \int f d\mu$, one can assume that $\int f d\mu = 0$. Then, for all $t\geqslant 0$,

$$\begin{aligned}\frac{d}{dt}\mathrm{Var}_\mu(P_t(f)) &= \frac{d}{dt}\int (P_t(f))^2 d\mu \\ &= 2\int P_t(f)(\mathcal{L}P_t(f))d\mu \\ &= -2\int \|\nabla P_t(f)\|^2 d\mu \qquad (*)\end{aligned}$$

Let us assume (*i*). With equation (*), we have

$$\begin{aligned}\frac{d}{dt}\mathrm{Var}_\mu(P_t(f)) &= -2\int \|\nabla P_t(f)\|^2 d\mu \\ &\leqslant -2\,\mathcal{P}_\mu^{-1}\int (P_t(f))^2 d\mu \\ &= -2\,\mathcal{P}_\mu^{-1}\mathrm{Var}_\mu(P_t(f)).\end{aligned}$$

The proof is then completed by using Grönwall's inequality.

Let us assume (*ii*). We write, for $t > 0$,

$$\frac{(\mathrm{Var}_\mu(P_t(f)) - \mathrm{Var}_\mu(f))}{t} \leqslant \frac{(\mathrm{e}^{-2t/\mathcal{P}_\mu} - 1)\mathrm{Var}_\mu(f)}{t}.$$

By letting t go to 0 and using equation (*),

$$2\mathcal{P}_\mu^{-1}\mathrm{Var}_\mu(f) \leqslant \frac{d}{dt}\mathrm{Var}_\mu(P_t(f))_{t=0} = 2\int \|\nabla f\|^2 d\mu,$$

which shows the converse implication. □

Remark 2. *Let f be a centered eigenvector of $-\mathcal{L}$ with eigenvalue $\lambda \neq 0$. By the Poincaré inequality,*

$$\begin{aligned}\int f^2 d\mu \leqslant \mathcal{P}_\mu \int \|\nabla f\|^2 d\mu &= \mathcal{P}_\mu \int f(-\mathcal{L}f)d\mu \\ &= \lambda\mathcal{P}_\mu \int f^2 d\mu,\end{aligned}$$

from which we deduce that every non-zero eigenvalue of $-\mathcal{L}$ is larger that $1/\mathcal{P}_\mu$. The best Poincaré constant is thus the inverse of the smallest non zero eigenvalue of $-\mathcal{L}$. The finiteness of the Poincaré constant is therefore equivalent to a spectral gap *property of $-\mathcal{L}$. Similarly, a discrete space Markov chain with transition matrix P converges at a rate determined by the spectral gap of $I - P$.*

There have been efforts in the past to estimate spectral gaps of Markov chains [Hsu et al., 2015, Levin and Peres, 2016, Qin et al., 2019, Wolfer and Kontorovich, 2019, Combes and Touati, 2019] but these have been done with samples from trajectories of the dynamics. The main difference here is that the estimation will only rely on samples from the stationary measure.

Poincaré constant and sampling. In high dimensional settings (in Bayesian machine learning [Robert, 2007]) or molecular dynamics [Lelièvre and Stoltz, 2016] where d can be large – from 100 to 10^7), one of the standard techniques to sample $d\mu(x) = \mathrm{e}^{-V(x)}dx$ is to build a Markov chain by discretizing in time the overdamped Langevin diffusion (3) whose law converges to $d\mu$. According to Theorem 1, the typical time to wait to reach equilibrium is $\mathcal{P}_\mu$. Hence, the larger the Poincaré constant of a probability measure $d\mu$ is, the more difficult the sampling of $d\mu$ is. Note also that V need not be convex for the Markov chain to converge.

2.3 Examples

Gaussian distribution. For the Gaussian measure on $\mathbb{R}^d$ of mean 0 and variance 1: $d\mu(x) = \frac{1}{(2\pi)^{d/2}}\mathrm{e}^{-\|x\|^2/2}dx$, it holds for all f smooth and compactly supported,

$$\mathrm{Var}_\mu(f) \leqslant \int_{\mathbb{R}^d} \|\nabla f\|^2 d\mu,$$

and one can show that $\mathcal{P}_\mu = 1$ is the optimal Poincaré constant (see [Chernoff, 1981]). More generally, for

a Gaussian measure with covariance matrix Σ, the Poincaré constant is the spectral radius of Σ.

Other examples of analytically known Poincaré constant are $1/d$ for the uniform measure on the unit sphere in dimension d [Ledoux, 2014] and 4 for the exponential measure on the real line [Bakry et al., 2014]. There also exist various criteria to ensure the existence of **(PI)**. We will not give an exhaustive list as our aim is rather to emphasize the link between sampling and optimization. Let us however finish this part with particularly important results.

A measure of non-convexity. Let $d\mu(x) = \mathrm{e}^{-V(x)}dx$. It has been shown in the past decades that the "more convex" V is, the smaller the Poincaré constant is. Indeed, if V is ρ-strongly convex, then the Bakry-Emery criterion [Bakry et al., 2014] tells us that $\mathcal{P}_\mu \leqslant 1/\rho$. If V is only convex, it has been shown that $d\mu$ satisfies also a **(PI)** (with a possibly very large Poincaré constant) [Ravindran et al., 1995, Bobkov, 1999]. Finally, the case where V is non-convex is explored in detail in a one-dimensional setting and it is shown that for potentials V with an energy barrier of height h between two wells, the Poincaré constant explodes exponentially with respect the height h [Menz and Schlichting, 2014]. In that spirit, the Poincaré constant of $d\mu(x) = \mathrm{e}^{-V(x)}dx$ can be a quantitative way to quantify how multimodal the distribution $d\mu$ is and hence how non-convex the potential V is [Jain and Kar, 2017, Raginsky et al., 2017].

3 Statistical Estimation of the Poincaré Constant

The aim of this section is to provide an estimator of the Poincaré constant of a measure μ when we only have access to n samples of it, and to study its convergence properties. More precisely, given n independent and identically distributed (i.i.d.) samples $(x_1, \dots, x_n)$ of the probability measure $d\mu$, our goal is to estimate $\mathcal{P}_\mu$. We will denote this estimator (function of $(x_1, \dots, x_n)$) by the standard notation $\widehat{\mathcal{P}}_\mu$.

3.1 Reformulation of the problem in a reproducing kernel Hilbert Space

Definition and first properties. Let us suppose here that the space of test functions of the **(PI)**, $\mathcal{H}$, is a reproducing kernel Hilbert space (RKHS) associated with a kernel K on $\mathbb{R}^d$ [Schölkopf and Smola, 2002, Shawe-Taylor and Cristianini, 2004]. This has two important consequences:

1. $\mathcal{H}$ is the linear function space $\mathcal{H} = \mathrm{span}\{K(\cdot, x),\ x \in \mathbb{R}^d\}$, and in particular, for all $x \in \mathbb{R}^d$, the function $y \mapsto K(y, x)$ is an element of $\mathcal{H}$ that we will denote by K_x.
2. The reproducing property: $\forall f \in \mathcal{H}$ and $\forall x \in \mathbb{R}^d$, $f(x) = \langle f, K(\cdot, x)\rangle_{\mathcal{H}}$. In other words, function evaluations are equal to dot products with canonical elements of the RKHS.

We make the following mild assumptions on the RKHS:

Ass. 1. *The RKHS $\mathcal{H}$ is dense in $H^1(\mu)$.*

Note that this is the case for most of the usual kernels: Gaussian, exponential [Micchelli et al., 2006]. As **(PI)** involves derivatives of test functions, we will also need some regularity properties of the RKHS. Indeed, to represent ∇f in our RKHS we need a partial derivative reproducing property of the kernel space.

Ass. 2. *K is a Mercer kernel and $K \in C^2(\mathbb{R}^d \times \mathbb{R}^d)$.*

Let us denote by $\partial_i = \partial_{x^i}$ the partial derivative operator with respect to the i-th component of x. It has been shown [Zhou, 2008] that under assumption **(Ass. 2)**, $\forall i \in [\![1, d]\!]$, $\partial_i K_x \in \mathcal{H}$ and that a partial derivative reproducing property holds true: $\forall f \in \mathcal{H}$ and $\forall x \in \mathbb{R}^d$, $\partial_i f(x) = \langle \partial_i K_x, f\rangle_{\mathcal{H}}$. Hence, thanks to assumption **(Ass. 2)**, ∇f is easily represented in the RKHS. We also need some boundedness properties of the kernel.

Ass. 3. *K is a kernel such that $\forall x \in \mathbb{R}^d$, $K(x, x) \leqslant \mathcal{K}$ and[1] $\|\nabla K_x\|^2 \leqslant \mathcal{K}_d$, where $\|\nabla K_x\|^2 := \sum_{i=1}^d \langle \partial_i K_x, \partial_i K_x\rangle = \sum_{i=1}^d \frac{\partial^2 K}{\partial x^i \partial y^i}(x, x)$ (see calculations below), x and y standing respectively for the first and the second variables of $(x, y) \mapsto K(x, y)$.*

The equality mentioned in the expression of $\|\nabla K_x\|^2$ arises from the following computation: $\partial_i K_y(x) = \langle \partial_i K_y, K_x\rangle = \partial_{y^i} K(x, y)$ and we can write that for all $x, y \in \mathbb{R}^d$, $\langle \partial_i K_x, \partial_i K_y\rangle = \partial_{x^i}\left(\partial_i K_y(x)\right) = \partial_{x^i}\partial_{y^i} K(x, y)$. Note that, for example, the Gaussian kernel satisfies **(Ass. 1)**, **(Ass. 2)**, **(Ass. 3)**.

A spectral point of view. Let us define the following operators from $\mathcal{H}$ to $\mathcal{H}$:

$$\Sigma = \mathbb{E}\left[K_x \otimes K_x\right], \qquad \Delta = \mathbb{E}\left[\nabla K_x \otimes_d \nabla K_x\right],$$

and their empirical counterparts,

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n K_{x_i} \otimes K_{x_i}, \qquad \widehat{\Delta} = \frac{1}{n}\sum_{i=1}^n \nabla K_{x_i} \otimes_d \nabla K_{x_i},$$

where $\otimes$ is the standard tensor product: $\forall f, g, h \in \mathcal{H}$, $(f \otimes g)(h) = \langle g, h\rangle_{\mathcal{H}} f$ and $\otimes_d$ is defined as follows: $\forall f, g \in \mathcal{H}^d$ and $h \in \mathcal{H}$, $(f \otimes_d g)(h) = \sum_{i=1}^d \langle g_i, h\rangle_{\mathcal{H}} f_i$.

[1] The subscript d in $\mathcal{K}_d$ accounts for the fact that this quantity is expected to scale linearly with d (as is the case for the Gaussian kernel).

Proposition 1 (Spectral characterization of the Poincaré constant)**.** *Suppose that assumptions **(Ass. 1)**, **(Ass. 2)**, **(Ass. 3)** hold true. Then the Poincaré constant $\mathcal{P}_\mu$ is the maximum of the following Rayleigh ratio:*

$$\mathcal{P}_\mu = \sup_{f \in \mathcal{H} \backslash \mathrm{Ker}(\Delta)} \frac{\langle f, Cf \rangle_{\mathcal{H}}}{\langle f, \Delta f \rangle_{\mathcal{H}}} = \left\| \Delta^{-1/2} C \Delta^{-1/2} \right\|, \quad (4)$$

with $\|\cdot\|$ the operator norm on $\mathcal{H}$ and $C = \Sigma - m \otimes m$ where $m = \int_{\mathbb{R}^d} K_x d\mu(x) \in \mathcal{H}$ is the covariance operator, considering Δ^{-1} as the inverse of Δ restricted to $(\mathrm{Ker}(\Delta))^\perp$.

Note that C and Δ are symmetric positive semi-definite trace-class operators (see Appendix C.2). Note also that $\mathrm{Ker}(\Delta)$ is the set of constant functions, which suggests introducing $\mathcal{H}_0 := (\mathrm{Ker}(\Delta))^\perp = \mathcal{H} \cap L^2_0(\mu)$, where $L^2_0(\mu)$ is the space of $L^2(\mu)$ functions with mean zero with respect to μ. Finally note that $\mathrm{Ker}(\Delta) \subset \mathrm{Ker}(C)$ (see Section A of the Appendix). With the characterization provided by Proposition 1, we can easily define an estimator of the Poincaré constant $\widehat{\mathcal{P}}_\mu$, following standard regularization techniques from kernel methods [Schölkopf and Smola, 2002, Shawe-Taylor and Cristianini, 2004, Fukumizu et al., 2007].

Definition 2. *The estimator $\widehat{\mathcal{P}}^{n,\lambda}_\mu$ of the Poincaré constant is the following:*

$$\widehat{\mathcal{P}}^{n,\lambda}_\mu := \sup_{f \in \mathcal{H} \backslash \mathrm{Ker}(\Delta)} \frac{\langle f, \widehat{C} f \rangle_{\mathcal{H}}}{\langle f, (\widehat{\Delta} + \lambda I) f \rangle_{\mathcal{H}}} = \left\| \widehat{\Delta}_\lambda^{-1/2} \widehat{C} \widehat{\Delta}_\lambda^{-1/2} \right\|, \quad (5)$$

with $\widehat{C} = \widehat{\Sigma} - \widehat{m} \otimes \widehat{m}$ and where $\widehat{m} = \frac{1}{n} \sum_{i=1}^n K_{x_i}$. $\widehat{C}$ is the empirical covariance operator and $\widehat{\Delta}_\lambda = \widehat{\Delta} + \lambda I$ is a regularized empirical version of the operator Δ restricted to $(\mathrm{Ker}(\Delta))^\perp$ as in Proposition 1.

Note that regularization is necessary as the nullspace of $\widehat{\Delta}$ is no longer included in the nullspace of $\widehat{C}$ so that the Poincaré constant estimates blows up when $\lambda \to 0$. The problem in Equation (5) has a natural interpretation in terms of Poincaré inequality as it corresponds to a regularized **(PI)** for the empirical measure $\widehat{\mu}_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ associated with the i.i.d. samples $x_1, \dots, x_n$ from $d\mu$. To alleviate the notation, we will simply denote the estimator by $\widehat{\mathcal{P}}_\mu$ until the end of the paper.

3.2 Statistical consistency of the estimator

We show that, under some assumptions and by choosing carefully λ as a function of n, the estimator $\widehat{\mathcal{P}}_\mu$ is statistically consistent, i.e., almost surely:

$$\widehat{\mathcal{P}}_\mu \xrightarrow{n \to \infty} \mathcal{P}_\mu.$$

As we regularized our problem, we prove the convergence in two steps: first, the convergence of $\widehat{\mathcal{P}}_\mu$ to the regularized problem $\mathcal{P}^\lambda_\mu = \sup_{f \in \mathcal{H} \backslash \{0\}} \frac{\langle f, Cf \rangle}{\langle f, (\Delta + \lambda I) f \rangle} = \|\Delta_\lambda^{-1/2} C \Delta_\lambda^{-1/2}\|$, which corresponds to controlling the statistical error associated with the estimator $\widehat{\mathcal{P}}_\mu$ (variance); second, the convergence of $\mathcal{P}^\lambda_\mu$ to $\mathcal{P}_\mu$ as λ goes to zero which corresponds to the bias associated with the estimator $\widehat{\mathcal{P}}_\mu$. The next result states the statistical consistency of the estimator when λ is a sequence going to zero as n goes to infinity (typically as an inverse power of n).

Theorem 2 (Statistical consistency)**.** *Assume that **(Ass. 1)**, **(Ass. 2)**, **(Ass. 3)** hold true and that the operator $\Delta^{-1/2} C \Delta^{-1/2}$ is compact on $\mathcal{H}$. Let $(\lambda_n)_{n \in \mathbb{N}}$ be a sequence of positive numbers such that $\lambda_n \to 0$ and $\lambda_n \sqrt{n} \to +\infty$. Then, almost surely,*

$$\widehat{\mathcal{P}}_\mu \xrightarrow{n \to \infty} \mathcal{P}_\mu.$$

As already mentioned, the proof is divided into two steps: the analysis of the statistical error for which we have an explicit rate of convergence in probability (see Proposition 2 below) and which requires $n^{-1/2}/\lambda_n \to 0$, and the analysis of the bias for which we need $\lambda_n \to 0$ and the compactness condition (see Proposition 3). Notice that the compactness assumption in Proposition 3 and Theorem 2 is stronger than **(PI)**. Indeed, it can be shown that satisfying **(PI)** is equivalent to having the operator $\Delta^{-1/2} C \Delta^{-1/2}$ bounded whereas to have convergence of the bias we need compactness. Note also that $\lambda_n = n^{-1/4}$ matches the two conditions stated in Theorem 2 and is the optimal balance between the rate of convergence of the statistical error (of order $\frac{1}{\lambda\sqrt{n}}$, see Proposition 2) and of the bias we obtain in some cases (of order λ, see Section B of the Appendix). Note that the rates of convergence do not depend on the dimension d of the problem which is a usual strength of kernel methods and differ from local methods like diffusion maps Coifman and Lafon [2006], Hein et al. [2007].

For the statistical error term, it is possible to quantify the rate of convergence of the estimator to the regularized Poincaré constant as shown below.

Proposition 2 (Analysis of the statistical error)**.** *Suppose that **(Ass. 1)**, **(Ass. 2)**, **(Ass. 3)** hold true. For any $\delta \in (0, 1/3)$, and $\lambda > 0$ such that $\lambda \leqslant \|\Delta\|$ and any integer $n \geqslant 15 \frac{\mathcal{K}_d}{\lambda} \log \frac{4 \mathrm{Tr} \Delta}{\lambda \delta}$, with probability at least $1 - 3\delta$,*

$$\left| \widehat{\mathcal{P}}_\mu - \mathcal{P}^\lambda_\mu \right| \leqslant \frac{8\mathcal{K}}{\lambda\sqrt{n}} \log(2/\delta) + \mathrm{o}\left(\frac{1}{\lambda\sqrt{n}} \right). \quad (6)$$

Note that in Proposition 2 we are only interested in the regime where $\lambda\sqrt{n}$ is large. Lemmas 5 and 6 of the Appendix give explicit and sharper bounds under refined

hypotheses on the spectra of C and Δ. Recall also that under assumption **(Ass. 3)**, C and Δ are trace-class operators (as proved in the Appendix, Section C.2) so that $\|\Delta\|$ and $\mathrm{Tr}(\Delta)$ are indeed finite. Finally, remark that (6) implies the almost sure convergence of the statistical error by applying the Borel-Cantelli lemma.

Proposition 3 (Analysis of the bias)**.** *Assume that **(Ass. 1)**, **(Ass. 2)**, **(Ass. 3)** hold true, and that the bounded operator $\Delta^{-1/2} C \Delta^{-1/2}$ is compact on $\mathcal{H}$. Then,*

$$\lim_{\lambda \to 0} \mathcal{P}_\mu^\lambda = \mathcal{P}.$$

As said above the compactness condition (similar to the one used for convergence proofs of kernel Canonical Correlation Analysis [Fukumizu et al., 2007]) is stronger than satisfying **(PI)**. The compactness condition adds conditions on the spectrum of $\Delta^{-1/2} C \Delta^{-1/2}$: it is discrete and accumulates at 0. We give more details on this condition in Section B of the Appendix and derive explicit rates of convergence under general conditions. We derive also a rate of convergence for more specific structures (Gaussian case or under an assumption on the support of μ) in Sections B and D of the Appendix.

4 Learning a Reaction Coordinate

If the measure μ is multimodal, the Langevin dynamics (3) is trapped for long times in certain regions (modes) preventing it from efficient space exploration. This phenomenon is called *metastability* and is responsible for the slow convergence of the diffusion to its equilibrium [Lelièvre, 2013, Lelièvre et al., 2008]. Some efforts in the past decade [Lelièvre, 2015] have focused on understanding this multimodality by capturing the behavior of the dynamics at a coarse-grained level, which often have a low-dimensional nature. The aim of this section is to take advantage of the estimation of the Poincaré constant to give a procedure to unravel these dynamically meaningful slow variables called reaction coordinate.

4.1 Good Reaction Coordinate

From a numerical viewpoint, a good reaction coordinate can be defined as a low dimensional function $\xi : \mathbb{R}^d \to \mathbb{R}^p$ ($p \ll d$) such that the family of conditional measures $(\mu(\cdot|\xi(x) = r))_{z \in \mathbb{R}^p}$ are "less multimodal" than the measure $d\mu$. This can be fully formalized in particular in the context of free energy techniques such as the adaptive biasing force method, see for example Lelièvre et al. [2008]. For more details on mathematical formalizations of metastability, we also refer to Lelièvre [2013]. The point of view we will follow in this work is to choose ξ in order to maximize the Poincaré constant of the pushforward distribution $\xi * \mu$. The idea is to capture in $\xi * \mu$ the essential multimodality of the original measure, in the spirit of the two scale decomposition of Poincaré or logarithmic Sobolev constant inequalities [Lelièvre, 2009, Menz and Schlichting, 2014, Otto and Reznikoff, 2007].

4.2 Learning a Reaction Coordinate

Optimization problem. Let us assume in this subsection that the reaction coordinate is an orthogonal projection onto a linear subspace of dimension p. Hence ξ can be represented by $\forall x \in \mathbb{R}^d$, $\xi(x) = Ax$ with $A \in \mathcal{S}^{p,d}$ where $\mathcal{S}^{p,d} = \{A \in \mathbb{R}^{p \times d} \text{ s. t. } AA^\top = I_p\}$ is the Stiefel manifold [Edelman et al., 1998]. As discussed in Section 4.1, to find a good reaction coordinate we look for ξ for which the Poincaré constant of the pushforward measure $\xi * \mu$ is the largest. Given n samples, let us define the matrix $X = (x_1, \dots, x_n)^\top \in \mathbb{R}^{n \times d}$. We denote by $\widehat{\mathcal{P}}_X$ the estimator of the Poincaré constant using the samples $(x_1, \dots, x_n)$. Hence $\widehat{\mathcal{P}}_{AX^\top}$ defines an estimator of the Poincaré constant of the pushforward measure $\xi * \mu$. Our aim is to find $\operatorname*{argmax}_{A \in \mathcal{S}^{p,d}} \widehat{\mathcal{P}}_{AX^\top}$.

Random features. One computational issue with the estimation of the Poincaré constant is that building $\widehat{C}$ and $\widehat{\Delta}$ requires respectively constructing $n \times n$ and $nd \times nd$ matrices. Random features [Rahimi and Recht, 2008] avoid this problem by building explicitly features that approximate a translation invariant kernel $K(x, x') = K(x - x')$. More precisely, let M be the number of random features, $(w_m)_{1 \leqslant m \leqslant M}$ be random variables independently and identically distributed according to $\mathbb{P}(dw) = \int_{\mathbb{R}^d} \mathrm{e}^{-\mathrm{i} w^\top \delta} K(\delta) d\delta \, dw$ and $(b_m)_{1 \leqslant m \leqslant M}$ be independently and identically distributed according to the uniform law on $[0, 2\pi]$, then the feature vector $\phi^M(x) = \sqrt{\frac{2}{M}} \left(\cos(w_1^\top x + b_1), \dots, \cos(w_M^\top x + b_M)\right)^\top \in \mathbb{R}^M$ satisfies $K(x, x') \approx \phi^M(x)^\top \phi^M(x')$. Therefore, random features allow to approximate $\widehat{C}$ and $\widehat{\Delta}$ by $M \times M$ matrices $\widehat{C}^M$ and $\widehat{\Delta}^M$ respectively. Finally, when these matrices are constructed using the projected samples, i.e. $\left(\cos(w_m^\top A x_i + b_m)\right)_{m \leq M;\ i \leq n}$, we denote them by $\widehat{C}_A^M$ and $\widehat{\Delta}_A^M$ respectively. Hence, the problem reads

$$\operatorname*{argmax}_{A \in \mathcal{S}^{p,d}} \widehat{\mathcal{P}}_{AX^\top} = \operatorname*{argmax}_{A \in \mathcal{S}^{p,d}} \max_{v \in \mathbb{R}^M \setminus \{0\}} F(A, v), \quad (7)$$

where $F(A, v) := \frac{v^\top \widehat{C}_A^M v}{v^\top (\widehat{\Delta}_A^M + \lambda I) v}$.

Algorithm. To solve the non-concave optimization problem (7), our procedure is to do one step of non-Euclidean gradient descent to update A (gradient descent in the Stiefel manifold) and one step by solving

the generalized eigenvalue problem to update v. More precisely, the algorithm reads:

Result. Best linear Reaction Coordinate: $A_* \in \mathcal{S}^{d,p}$
A_0 random matrix in $\mathcal{S}^{d,p}$, $\eta_t > 0$ step-size
for $t = 0, \dots, T-1$ **do**
(i) Solve generalized largest eigenvalue problem with matrices $\widehat{C}_{A_t}^M$ and $\widehat{\Delta}_{A_t}^M$ to get $v^*(A_t)$:

$$v^*(A_t) = \underset{v \in \mathbb{R}^M}{\operatorname{argmax}} \frac{v^\top \widehat{C}_A^M v}{v^\top (\widehat{\Delta}_A^M + \lambda I) v}.$$

(ii) Do one gradient ascent step:

$$A_{t+1} = A_t + \eta_t \operatorname{grad}_A F(A, v^*(A_t)).$$

end

5 Numerical experiments

We divide our experiments into two parts: the first one illustrates the convergence of the estimated Poincaré constant as given by Theorem 2 (see Section 5.1), and the second one demonstrates the interest of the reaction coordinates learning procedure described in Section 4.2 (see Section 5.2).

5.1 Estimation of the Poincaré constant

In our experiments we choose the Gaussian Kernel $K(x, x') = \exp(-\|x - x'\|^2)$? This induces a RKHS satisfying **(Ass. 1)**, **(Ass. 2)**, **(Ass. 3)**. Estimating $\widehat{\mathcal{P}}_\mu$ from n samples $(x_i)_{i \leqslant n}$ is equivalent to finding the largest eigenvalue for an operator from $\mathcal{H}$ to $\mathcal{H}$. Indeed, we have

$$\widehat{\mathcal{P}}_\mu = \|(\widehat{Z}_n^* \widehat{Z}_n + \lambda I)^{-\frac{1}{2}} \widehat{S}_n^* (I - \frac{1}{n} \mathbb{1}\mathbb{1}^\top) \widehat{S}_n (\widehat{Z}_n^* \widehat{Z}_n + \lambda I)^{-\frac{1}{2}} \|_{\mathcal{H}},$$

where $\widehat{Z}_n = \sum_{i=1}^d \widehat{Z}_n^i$ and $\widehat{Z}_n^i$ is the operator from $\mathcal{H}$ to $\mathbb{R}^n$: $\forall g \in \mathcal{H}$, $\widehat{Z}_n^i(g) = \frac{1}{\sqrt{n}} \left(\langle g, \partial_i K_{x_j} \rangle \right)_{1 \leqslant j \leqslant n}$ and $\widehat{S}_n$ is the operator from $\mathcal{H}$ to $\mathbb{R}^n$: $\forall g \in \mathcal{H}$, $\widehat{S}_n(g) = \frac{1}{\sqrt{n}} \left(\langle g, K_{x_j} \rangle \right)_{1 \leqslant j \leqslant n}$. By the Woodbury operator identity, $(\lambda I + \widehat{Z}_n^* \widehat{Z}_n)^{-1} = \frac{1}{\lambda} \left(I - \widehat{Z}_n^* (\lambda I + \widehat{Z}_n \widehat{Z}_n^*)^{-1} \widehat{Z}_n \right)$, and the fact that for any operator $\|T^* T\| = \|T T^*\|$, we can show that

$$\widehat{\mathcal{P}}_\mu = \frac{1}{\lambda} \|(I - \frac{1}{n} \mathbb{1}\mathbb{1}^\top)(\widehat{S}_n \widehat{S}_n^* - \widehat{S}_n \widehat{Z}_n^* (\widehat{Z}_n \widehat{Z}_n^* + \lambda I)^{-1} \widehat{Z}_n \widehat{S}_n^*)(I - \frac{1}{n} \mathbb{1}\mathbb{1}^\top) \|_2,$$

which is now the largest eigenvalue of a $n \times n$ matrix built as the product of matrices involving the kernel K and its derivatives. Note for the above calculation that we used that $\left(I - \frac{1}{n} \mathbb{1}\mathbb{1}^\top\right)^2 = \left(I - \frac{1}{n} \mathbb{1}\mathbb{1}^\top\right)$.

We illustrate in Figure 1 the rate of convergence of the estimated Poincaré constant to 1 for the Gaussian $\mathcal{N}(0, 1)$ as the number of samples n grows. Recall that in this case the Poincaré constant is equal to 1 (see Subsection 2.3). We compare our prediction to the one given by diffusion maps techniques [Coifman and Lafon, 2006]. For our method, in all the experiments we set $\lambda_n = \frac{C_\lambda}{n}$, which is smaller than what is given by Theorem 2, and optimize the constant C_λ with a grid search. Following [Hein et al., 2007], to find the correct bandwidth ε_n of the kernel involved in diffusion maps, we performed a similar grid search on the constant C_ε for the Diffusion maps with the scaling $\varepsilon_n = \frac{C_\varepsilon}{n^{1/4}}$. Additionally to a faster convergence when n become large, the kernel-based method is more robust with respect to the choice of itss hyperparameter, which is of crucial importance for the quality of diffusion maps. Note also that we derive an explicit convergence rate for the bias in the Gaussian case in Section D of the Appendix. In Figure 1, we also show the growth of the Poincaré constant for a mixture of Gaussians of variances 1 as a function of the distance between the two means of the Gaussians. This is a situation for which the estimation provides an estimate when, up to our knowledge, no precise Poincaré constant is known (even if lower and upper bounds are known [Chafaï and Malrieu, 2010]).

5.2 Learning a reaction coordinate

We next illustrate the algorithm described in Section 4 to learn a reaction coordinate which, we recall, encodes directions which are difficult to sample. To perform the gradient step over the Stiefel manifold we used Pymanopt [Townsend et al., 2016], a Python library for manifold optimization derived from Manopt [Boumal et al., 2014] (Matlab). We show here a synthetic two-dimensional example example. We first preprocessed the samples with "whitening", i.e., making it of variance 1 in all directions to avoid scaling artifacts. In both examples, we took $M = 200$ for the number of random features and $n = 200$ for the number of samples.

We show (Figure 2) one synthetic example for which our algorithm found a good reaction coordinate. The samples are taken from a mixture of three Gaussians of means $(0, 0), (1, 1)$ and $(2, 2)$ and covariance $\Sigma = \sigma^2 I$ where $\sigma = 0.1$. The three means are aligned along a line which makes an angle $\theta = \pi/4$ with respect to the x-axis: one expects the algorithm to identify this direction as the most difficult one to sample (see left and center plots of Figure 2). With a few restarts, our algorithm indeed finds the largest Poincaré constant for a projection onto the line parametrized by $\theta = \pi/4$.

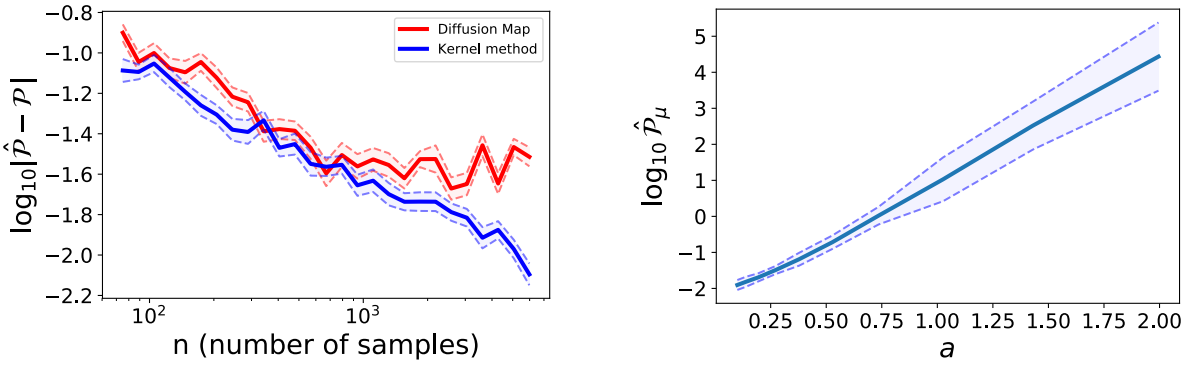

Figure 1: **(Left)** Comparison of the convergences of the kernel-based method described in this paper and diffusion maps in the case of a Gaussian of variance 1 (for each n we took the mean over 50 runs). The dotted lines correspond to standard deviations of the estimator. **(Right)** Exponential growth of the Poincaré constant for a mixture of two Gaussians $\mathcal{N}(\pm\frac{a}{2}, \sigma^2)$ as a function of the distance a between the two Gaussians ($\sigma = 0.1$ and $n = 500$).

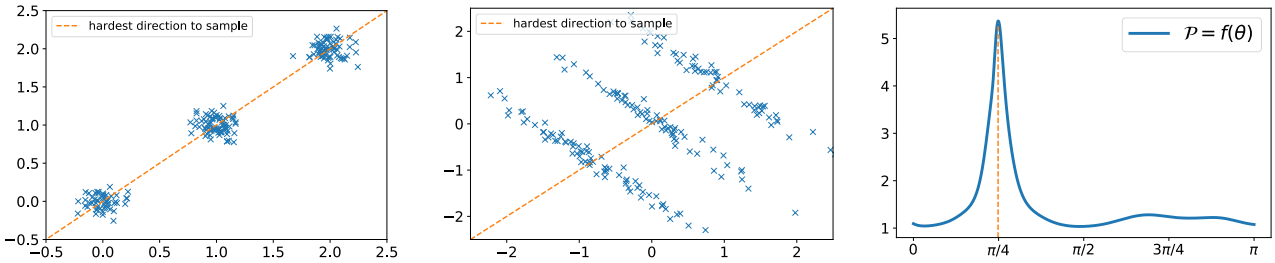

Figure 2: **(Top Left)** Samples of mixture of three Gaussians. **(Top right)** Whiten samples of Gaussian mixture on the left. **(Bottom)** Plot of the Poincaré constant of the projected samples on a line of angle θ.

6 Conclusion and Perspectives

In this paper, we have presented an efficient method to estimate the Poincaré constant of a distribution from independent samples, paving the way to learn low-dimensional marginals that are hard to sample (corresponding to the image measure of so-called reaction coordinates). While we have focused on linear projections, learning non-linear projections is important in molecular dynamics and it can readily be done with a well-defined parametrization of the non-linear function and then applied to real data sets, where this would lead to accelerated sampling [Lelièvre, 2015]. Finally, it would be interesting to apply our framework to Bayesian inference [Chopin et al., 2012] and leverage the knowledge of reaction coordinates to accelerate sampling methods.

References

Christian Robert. *The Bayesian Choice: from Decision-theoretic Foundations to Computational Implementation*. Springer Science & Business Media, 2007.

Kevin P. Murphy. *Machine Learning: a Probabilistic Perspective*. MIT Press, 2012.

Dani Gamerman and Hedibert Lopes. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman and Hall/CRC, 2006.

Walter Gilks, Sylvia Richardson, and David Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman and Hall/CRC, 1995.

Alain Durmus and Eric Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Annals of Applied Probability*, 27(3):1551–1587, 2017.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1674–1703, Amsterdam, Netherlands, 07–10 Jul 2017. PMLR.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

Stephan Mandt, Matthew D. Hoffman, and David M. Blei. Stochastic gradient descent as approximate

Bayesian inference. *J. Mach. Learn. Res.*, 18(1):4873–4907, January 2017. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=3122009.3208015.

Tony Lelièvre and Gabriel Stoltz. Partial differential equations and stochastic methods in molecular dynamics. *Acta Numerica*, 25:681–880, 2016. doi: 10.1017/S0962492916000039.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*. Springer, 2014.

Sean P. Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Springer Science & Business Media, 2012.

Yi-An Ma, Yuansi Chen, Chi Jin, Nicolas Flammarion, and Michael I. Jordan. Sampling can be faster than optimization. *arXiv preprint arXiv:1612.05330*, 2018.

Alain Durmus, Gareth O. Roberts, Gilles Vilmart, and Konstantinos C. Zygalakis. Fast Langevin based algorithm for mcmc in high dimensions. *Ann. Appl. Probab.*, 27(4):2195–2237, 08 2017. doi: 10.1214/16-AAP1257. URL https://doi.org/10.1214/16-AAP1257.

Tony Lelièvre, Mathias Rousset, and Gabriel Stoltz. Long-time convergence of an adaptive biasing force method. *Nonlinearity*, 21(6):1155–1181, 2008. doi: 10.1088/0951-7715/21/6/001. URL https://doi.org/10.1088%2F0951-7715%2F21%2F6%2F001.

Tony Lelièvre. Two mathematical tools to analyze metastable stochastic processes. In *Numerical Mathematics and Advanced Applications 2011*, pages 791–810, Berlin, Heidelberg, 2013. Springer.

Nicolas Chopin, Tony Lelièvre, and Gabriel Stoltz. Free energy methods for Bayesian inference: efficient exploration of univariate Gaussian mixture posteriors. *Statistics and Computing*, 22(4):897–916, 2012. doi: 10.1007/s11222-011-9257-9.

Paraskevi Gkeka. Machine learning force field and coarse-grained variables in molecular dynamics: application to materials and biological systems. *Preprint*, 2019.

Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1), 2006. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2006.04.006. URL http://www.sciencedirect.com/science/article/pii/S1063520306000546.

Ronald Coifman, Nadler Boaz, Stéphane Lafon, and Ioannis Kevrekidis. Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. *Applied and Computational Harmonic Analysis*, 21 (12):113–127, 2006.

Mary A. Rohrdanz, Wenwei Zheng, Mauro Maggioni, and Cecilia Clementi. Determination of reaction coordinates via locally scaled diffusion map. *The Journal of Chemical Physics*, 134(12):124116, 2011. doi: 10.1063/1.3569857.

David Gilbarg and Neil S. Trudinger. *Elliptic Partial Differential Equations of Second Order*. Springer Berlin Heidelberg, 2001.

Nathael Gozlan. Poincaré inequalities and dimension free concentration of measure. *Ann. Inst. H. Poincaré Probab. Statist.*, 46(3):708–739, 2010. doi: 10.1214/09-AIHP209. URL https://doi.org/10.1214/09-AIHP209.

Daniel Hsu, Aryeh Kontorovich, and Csaba Szepesvári. Mixing time estimation in reversible markov chains from a single sample path. In *Advances in neural information processing systems*, pages 1459–1467, 2015.

David Levin and Yuval Peres. Estimating the spectral gap of a reversible markov chain from a short trajectory. *arXiv preprint 1612.05330*, 2016.

Qian Qin, James P Hobert, Kshitij Khare, et al. Estimating the spectral gap of a trace-class markov operator. *Electronic Journal of Statistics*, 13(1): 1790–1822, 2019.

Geoffrey Wolfer and Aryeh Kontorovich. Estimating the mixing time of ergodic markov chains. *arXiv preprint arXiv:1902.01224*, 2019.

Richard Combes and Mikael Touati. Computationally efficient estimation of the spectral gap of a markov chain. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(1):7, 2019.

Herman Chernoff. A note on an inequality involving the normal distribution. *Ann. Probab.*, 9(3):533–535, 1981. doi: 10.1214/aop/1176994428. URL https://doi.org/10.1214/aop/1176994428.

Michel Ledoux. Concentration of measure and logarithmic Sobolev inequalities. *Séminaire de Probas XXXIII*, pages 120–216, 2014.

Kannan Ravindran, Lovasz Laszlo, and Simonovits Miklos. Isoperimetric problems for convex bodies and a localization lemma. *Discrete Comput. Geom.*, 13(3):541–559, 1995.

Sergey G. Bobkov. Isoperimetric and analytic inequalities for log-concave probability measures. *Ann. Probab.*, 27:1903–1921, 1999.

Georg Menz and André Schlichting. Poincaré and logarithmic Sobolev inequalities by decomposition of the energy landscape. *Ann. Probab.*, 42(5):1809–1884, 09 2014. doi: 10.1214/14-AOP908. URL https://doi.org/10.1214/14-AOP908.

Prateek Jain and Purushottam Kar. Non-convex optimization for machine learning. *Foundations and Trends in Machine Learning*, 10(3-4):142–336, 2017.

Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels*. MIT Press, 2002.

John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

Charles A Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *Journal of Machine Learning Research*, 7:2651–2667, 2006.

Ding-Xuan Zhou. Derivative reproducing properties for kernel methods in learning theory. *Journal of Computational and Applied Mathematics*, 220(1):456–463, 2008. ISSN 0377-0427. doi: https://doi.org/10.1016/j.cam.2007.08.023. URL http://www.sciencedirect.com/science/article/pii/S0377042707004657.

Kenji Fukumizu, Francis Bach, and Arthur Gretton. Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8: 361–383, 2007.

Matthias Hein, Jean-Yves Audibert, and Ulrike von Luxburg. Graph laplacians and their convergence on random neighborhood graphs. *Journal of Machine Learning Research*, pages 1325–1368, 2007.

Tony Lelièvre. Accelerated dynamics: Mathematical foundations and algorithmic improvements. *The European Physical Journal Special Topics*, 224(12): 2429–2444, 2015. ISSN 1951-6401. doi: 10.1140/epjst/e2015-02420-1. URL https://doi.org/10.1140/epjst/e2015-02420-1.

Tony Lelièvre. A general two-scale criteria for logarithmic Sobolev inequalities. *Journal of Functional Analysis*, 256(7):2211 – 2221, 2009. ISSN 0022-1236. doi: https://doi.org/10.1016/j.jfa.2008.09.019. URL http://www.sciencedirect.com/science/article/pii/S0022123608004084.

Felix Otto and Maria G. Reznikoff. A new criterion for the logarithmic Sobolev inequality and two applications. *Journal of Functional Analysis*, 243(1):121–157, 2007. ISSN 0022-1236. doi: https://doi.org/10.1016/j.jfa.2006.10.002. URL http://www.sciencedirect.com/science/article/pii/S0022123606004058.

Alan Edelman, Tomás A. Arias, and Steven T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20*, pages 1177–1184. 2008.

Djalil Chafaï and Florent Malrieu. On fine properties of mixtures with respect to concentration of measure and Sobolev type inequalities. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 46:72–96, 2010.

James Townsend, Niklas Koep, and Sebastian Weichwald. Pymanopt: A python toolbox for optimization on manifolds using automatic differentiation. *Journal of Machine Learning Research*, 17(137):1–5, 2016. URL http://jmlr.org/papers/v17/16-177.html.

Nicolas Boumal, Bamdev Mishra, P.-A. Absil, and Rodolphe Sepulchre. Manopt, a Matlab toolbox for optimization on manifolds. *Journal of Machine Learning Research*, 15:1455–1459, 2014. URL http://www.manopt.org.

Klaus Gansberger. An idea on proving weighted Sobolev embeddings. *arXiv:1007.3525.*, 2010.

Michael Reed and Barry Simon. *Methods of Modern Mathematical Physics: Functional Analysis*, volume IV. Elsevier, 2012.

Bernard Helffer and Francis Nier. Hypoelliptic estimates and spectral theory for Fokker-Planck operators and Witten Laplacians. *Lecture Notes in Mathematics*, 1862, 2005.

Hassler Whitney. Analytic extensions of differentiable functions defined in closed sets. *Transactions of the American Mathematical Society*, 36:63–89, 1934.

Vadim Vladimirovich Yurinsky. *Gaussian and Related Approximations for Distributions of Sums*, pages 163–216. Springer Berlin Heidelberg, 1995. ISBN 978-3-540-44791-7. doi: 10.1007/BFb0092604. URL https://doi.org/10.1007/BFb0092604.

Joel A. Tropp. User-friendly tools for random matrices: an introduction. NIPS Tutorials, 2012.

Minsker Stanislav. On some extensions of Bernstein's inequality for self-adjoint operators. *Statistics and Probability Letters*, 127:111–119, 2017. ISSN 0167-7152. doi: https://doi.org/10.1016/j.spl.2017.03.020.

Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, pages 3218–3228, 2017.