Math. Stat. Learn. 8 (2025), 71–104
DOI 10.4171/MSL/49

Convergence in total variation for the kinetic Langevin algorithm

Joseph Lehec

Abstract. We prove non-asymptotic total variation estimates for the kinetic Langevin algorithm in high dimension when the target measure satisfies a Poincaré inequality and has gradient Lipschitz potential. The main point is that the estimate improves significantly upon the corresponding bound for the non-kinetic version of the algorithm, due to Dalalyan. In particular, the dimension dependence drops from $O(n)$ to $O(\sqrt{n})$.

1. Introduction

1.1. Context

Suppose we want to sample from a probability measure μ on $\mathbb{R}^n$ of the form

$$\mu(dx) = \mathrm{e}^{-V(x)}\, dx,$$

where V is some smooth function from $\mathbb{R}^n$ to $\mathbb{R}$ (never mind the precise hypothesis for now) which we call the potential of μ. This is a very common problem in applied mathematics, it shows up in many different contexts, from Bayesian statistics, to optimization, machine learning and many more. We will not discuss applications here at all. Instead we focus on the sampling problem from a theoretical point of view. We shall investigate a particular algorithm called the *kinetic Langevin algorithm*. This is an order 1 algorithm, in the sense that it relies on the knowledge of the gradient of V. Therefore, it is assumed throughout this article that there is some oracle that given a point x in $\mathbb{R}^n$ returns $\nabla V(x)$. In this context the complexity of the sampling algorithm is the number of oracle queries.

1.2. The Langevin algorithm and its kinetic version

Consider the following stochastic differential equation

$$dX_t = \sqrt{2}\, dW_t - \nabla V(X_t)\, dt,$$

Mathematics Subject Classification 2020: 65C05 (primary); 35H10, 68W20 (secondary).
Keywords: hypoelliptic equations, Monte Carlo methods, randomized algorithms.

where (W_t) is a standard Brownian motion on $\mathbb{R}^n$. Under mild hypotheses the equation admits a unique strong solution, which is a Markov process, for which μ is the unique stationary measure. Moreover, the process is ergodic, in the sense that we have convergence of X_t to μ in law as t tends to $+\infty$. The Langevin algorithm is the Markov chain induced by the Euler scheme associated to this diffusion. This means that given a time step parameter η, the algorithm is given by

$$x_{k+1} = x_k + \sqrt{2\eta}\,\xi_{k+1} - \eta \nabla V(x_k),$$

where (ξ_k) is an i.i.d. sequence of standard Gaussian vectors on $\mathbb{R}^n$. There is quite a lot of literature studying the performance of this algorithm, either empirically or from a more theoretical point of view. Explicit non-asymptotic bounds seem to have started with the seminal work of Dalalyan [10] to which we will come back later on.

Let us move on to the kinetic version of the algorithm. The main idea is to add another variable which is interpreted as a speed variable. We thus consider the stochastic differential equation on $\mathbb{R}^n \times \mathbb{R}^n$ given by

$$\begin{cases} dX_t = Y_t\,dt, \\ dY_t = \sqrt{2\beta}\,dW_t - \beta Y_t\,dt - \nabla V(X_t)\,dt, \end{cases} \tag{1}$$

where (W_t) is a standard Brownian motion on $\mathbb{R}^n$ and β is a positive parameter, called the *friction* parameter hereafter. Note that this equation is degenerate, in the sense that we only have a diffusion on the speed variable and not on the space variable. The kinetic version of the Langevin diffusion is sometimes called *underdamped* Langevin diffusion, whereas the usual diffusion is called *overdamped*. We will use both terminologies here, so underdamped and overdamped are synonyms for kinetic and non-kinetic, respectively. The underdamped diffusion admits a unique stationary distribution, namely the measure $\pi := \mu \otimes \gamma$, where γ is the standard Gaussian measure on $\mathbb{R}^n$. Under mild assumptions, the diffusion is also ergodic, and (X_t, Y_t) converges in distribution to π. As far as sampling is concerned what matters is that the first factor is our target measure μ, so that the position X_t converges to μ. The kinetic Langevin algorithm is the algorithm obtained by discretizing the diffusion (1). Namely, we fix a time step parameter η and we consider the following system of equations

$$\begin{cases} dX_t^\eta = Y_t^\eta\,dt, \\ dY_t^\eta = \sqrt{2\beta}\,dW_t - \beta Y_t^\eta\,dt - \nabla V(X^\eta_{\lfloor t/\eta \rfloor \eta})\,dt, \end{cases} \tag{2}$$

where η is the time step and $\lfloor \cdot \rfloor$ denotes the integer part. Thus the only difference with (1) is that in the equation for the speed, the gradient is not queried at the current position but at some past position, corresponding to latest integer multiple of η. We initiate this at some (possibly random) point (x_0, y_0) and we set $(x_k, y_k) = (X^\eta_{k\eta}, Y^\eta_{k\eta})$

for every integer k. The process (x_k, y_k) is Markov chain whose transition kernel is explicit. Namely, the transition measure at point (x, y) is the Gaussian measure on $\mathbb{R}^n \times \mathbb{R}^n$, centered at point (x', y') and with covariance matrix $\left(\begin{smallmatrix} a & c \\ c & b \end{smallmatrix}\right) \otimes I_n$, where

$$\begin{aligned} x' &= x + \frac{1 - \mathrm{e}^{-\beta\eta}}{\beta} y - \frac{\mathrm{e}^{-\beta\eta} - 1 + \beta\eta}{\beta^2} \nabla V(x), \\ y' &= \mathrm{e}^{-\beta\eta} y - \frac{1 - \mathrm{e}^{-\beta\eta}}{\beta} \nabla V(x), \end{aligned}$$

$$a = \frac{1}{\beta^2}(4\mathrm{e}^{-\beta\eta} - \mathrm{e}^{-2\beta\eta} + 2\beta\eta - 3), \quad b = 1 - \mathrm{e}^{-2\beta\eta}, \quad c = \frac{1}{\beta}(1 - \mathrm{e}^{-\beta\eta})^2.$$

Indeed the solution of the equation (2) on $[0, \eta)$ is completely explicit. Namely, if we start from (x, y), then we have

$$Y_t^\eta = \mathrm{e}^{-\beta t} y - \frac{1 - \mathrm{e}^{-\beta t}}{\beta} \nabla V(x) + \sqrt{2\beta} \int_0^t \mathrm{e}^{\beta(s-t)} \, dW_s, \quad \forall t \leq \eta.$$

From this we also get an explicit formula for X_t^η. This shows that conditioned on the initial point (x, y) the random vector $(X_\eta^\eta, Y_\eta^\eta)$ is a Gaussian vector on $\mathbb{R}^n \times \mathbb{R}^n$. Finding the different parameters is only a matter of computation which is omitted here. See e.g. [7,28] for more details.

Thus the transition kernel is just a Gaussian kernel with a somewhat intricate but completely explicit covariance matrix, which depends only on the friction parameter β and the time step η, and not on the potential V. The potential V only appears via its gradient in the center of mass of the Gaussian kernel. Each step of this Markov chain is thus easy to sample, under the assumption that we have an oracle for ∇V.

This Markov chain is what we call the kinetic Langevin algorithm associated to μ. It depends on two parameters, the friction parameter β and the discretization parameter η. We will show that if those parameters are chosen appropriately then after a polynomial number of steps the distribution of (x_k, y_k) is very close to π.

1.3. Main result

We will quantify the sampling error in terms of total variation distance: if μ and ν are two probability measures defined on the same space E equipped with some σ-field $\mathcal{A}$ then

$$TV(\mu, \nu) = \sup_{A \in \mathcal{A}} \{|\mu(A) - \nu(A)|\}.$$

A related notion is the chi-square divergence, defined by

$$\chi_2(\mu \mid \nu) = \int_E \frac{d\mu}{d\nu} \, d\mu - 1,$$

assuming that μ is absolutely continuous with respect to ν. If this is not the case the convention is that the chi-square divergence is $+\infty$. The chi-square divergence is not a distance, it is not symmetric and it does not satisfy the triangle inequality either. However it controls the total variation distance. Indeed, we have

$$TV(\mu,\nu) \le \sqrt{\log(1+\chi_2(\mu \mid \nu))} \le \sqrt{\chi_2(\mu \mid \nu)}.$$

By a slight abuse of notations, when X, Y are random vectors we shall often write $TV(X,Y)$ for $TV(\mu,\nu)$ where μ,ν are the respective laws of X,Y. We adopt a similar convention for the chi-square divergence. For instance, we shall write $\chi_2(X \mid \nu)$ for the relative chi-square divergence of the law of X with respect to the measure ν.

There will be two hypotheses for the target measure. First of all we assume that μ satisfies the Poincaré inequality. Namely, we assume that there is a constant C_P such that for any locally Lipschitz function f, we have

$$\operatorname{var}_\mu(f) \le C_P \int_{\mathbb{R}^n} |\nabla f|^2 \, d\mu.$$

The second hypothesis is a regularity hypothesis: We assume that the potential V of μ is $\mathcal{C}^2$-smooth with Lipschitz gradient. The Lipschitz constant of ∇V will be denoted by L throughout. We shall give two estimates. One that is valid under the above two assumptions only, and one slightly better, under the additional assumption that μ is log-concave, in the sense that the potential V is a convex function.

Theorem 1. *Suppose that μ satisfies Poincaré with constant C_P and has $\mathcal{C}^2$-smooth and gradient Lipschitz potential, with constant L. Assume that the kinetic Langevin algorithm is initiated at (x_0, y_0) with x_0 independent of y_0, and y_0 distributed according to the standard Gaussian measure on $\mathbb{R}^n$. Fix the friction parameter $\beta = \sqrt{L}$, and set the time step parameter η and the number of steps k by*

$$\eta = c \cdot \varepsilon L^{-1} C_P^{-1/2} \cdot \left(\sqrt{n} + \sqrt{\log(1+\chi_2(x_0 \mid \mu))}\right)^{-1} \cdot \log^{-1/2}\left(\frac{\chi_2(x_0 \mid \mu)}{\varepsilon}\right),$$
$$k = C \cdot \varepsilon^{-1} (L C_P)^{3/2} \cdot \left(\sqrt{n} + \sqrt{\log(1+\chi_2(x_0 \mid \mu))}\right) \cdot \log^{3/2}\left(\frac{\chi_2(x_0 \mid \mu)}{\varepsilon}\right),$$

where c, C are universal constants. Then we have

$$TV(x_k, \mu) \le \varepsilon.$$

If in addition μ is log-concave then we have the same result with friction $\beta = C_P^{-1/2}$, time step η as above, and number of steps k given by

$$k = C \cdot \varepsilon^{-1} L C_P \cdot \left(\sqrt{n} + \sqrt{\log(1+\chi_2(x_0 \mid \mu))}\right) \cdot \log^{3/2}\left(\frac{\chi_2(x_0 \mid \mu)}{\varepsilon}\right).$$

Some comments are in order. First of all the constants c, C could be explicitly tracked down from the proof, but we have chosen not to do so in order to lighten the exposition. In particular, the proof has not been optimized so as minimize the value of C. That being said, our proof does not yield horribly large constants either, see the discussion right after Theorem 2 below.

Secondly, the result depends on some *warm start* hypothesis. Namely the algorithm must be initiated at some point x_0 for which we have some control on the chi-square divergence to the target measure. Assuming that this chi-square divergence is polynomial in the dimension, we then get

$$O^*\big(\varepsilon^{-1}(LC_P)^{3/2} \cdot n^{1/2}\big)$$

complexity for the kinetic Langevin algorithm in general and $O^*(\varepsilon^{-1}LC_P \cdot n^{1/2})$ complexity in the log-concave case. The notation O^* means up to a multiplicative constant and possibly poly-logarithmic factors. This is a common notation when studying the complexity of algorithms. Note that even in the more pessimistic scenario were the chi-square divergence of the initial distribution is exponentially large in the dimension, Theorem 1 still provides polynomial bounds, though with a worst dependence on the dimension.

Coming back to the situation where a polynomial warm-start is available, our result should be compared with the complexity one gets for the overdamped version of the algorithm under the same set of hypothesis. The state of the art is Dalalyan [10], which gives convergence after $O^*(\varepsilon^{-2}(LC_P)^2 n)$ steps of the algorithm. The result in [10] is not quite written this way but the above complexity does follow from the proof, see the discussion at the end of the introduction of [19]. Well, in this discussion the assumption is that μ satisfies the log-Sobolev inequality, and that we have a warm start in the relative entropy sense, but the argument is exactly the same assuming Poincaré and a chi-square divergence warm start assumption.

The most important improvement given by the kinetic version is probably the dependence in the dimension, which drops from $O(n)$ to $O(\sqrt{n})$, but it should be noted that also the dependence on the precision ε, on the Lipschitz constant L of the potential and on the Poincaré constant C_P of the measure are definitely improved by the kinetic version of the algorithm, all the more so in the log-concave case.

1.4. The hypocoercive estimate

The proof of the main theorem splits into two parts: firstly we need to estimate the speed of convergence of the true diffusion towards its equilibrium measure and secondly to control the discretization error. For the first part the difficulty lies in the fact that the diffusion is degenerate, in the sense that there is only a diffusion term in the speed variable and not in the space variable. This implies that we cannot expect an exponential decay of the chi-square divergence along the diffusion of the form:

$$\chi_2(\nu P_t \mid \pi) \leq \mathrm{e}^{-ct} \chi_2(\nu \mid \pi),$$

where c is the positive constant. Here ν is some probability measure on the product space $\mathbb{R}^n \times \mathbb{R}^n$ and (P_t) denotes the semigroup associated to the kinetic Langevin equation (1). In other words νP_t denotes the law of (X_t, Y_t) when (X_0, Y_0) is distributed according to ν. The terminology is that the diffusion fails to be coercive. However, there is no obvious obstruction to having this inequality with a prefactor. Such estimates are called hypocoercive estimates, and there are a number of them available in the literature.

The result that we shall use is essentially from Cao, Lu and Wang [5], but for reasons to be explained later on we will reprove the result from scratch rather than taking it for granted. In terms of hypothesis, the hypocoercive estimate still requires Poincaré, but a weaker condition than gradient Lipschitz potential is enough. Indeed, it only requires a semi-convexity property. Namely, the assumption is that there exists a constant $\kappa \geq 0$ such that the potential V of μ has the property that $V(x) + \frac{\kappa}{2}|x|^2$ is a convex function of x. Note that if ∇V is Lipschitz with constant L then this holds true with $\kappa = L$ but the converse is not true.

Theorem 2. *If μ has $\mathscr{C}^2$-smooth and semi-convex potential, with constant $\kappa \geq 0$, and satisfies Poincaré with constant C_P, then for every probability measure ν on $\mathbb{R}^n \times \mathbb{R}^n$, we have*

$$\chi_2(\nu P_t \mid \pi) \leq 2 \cdot \exp\left(-c \cdot \frac{\beta t}{1 + (\beta^2 + \kappa) C_P} \right) \cdot \chi_2(\nu \mid \pi),$$

for all $t > 0$, where β is the friction parameter of the kinetic Langevin diffusion and c is a positive universal constant.

It should be noted that most hypocoercivity results are established with a prefactor that depends on the initial measure ν. A nice feature of this one is that the prefactor is just a universal constant.

Again, the focus of this work is on the theoretical aspects of the kinetic Langevin algorithm rather than its practical implementation, but for this result we do provide a version of the inequality that is completely free of hidden constants, namely

$$\chi_2(\nu P_t \mid \pi) \leq \exp\left(-\frac{\beta t}{10 \cdot (3 + \beta\sqrt{C_P} + 2\sqrt{1 + \kappa C_P})^2} + \frac{1}{60} \right) \cdot \chi_2(\nu \mid \pi)$$

for all $t > 0$. This is obtained by simply keeping track of the various constants involved in the proof of the theorem, see Section 2 below.

1.5. Related works

The fact that the kinetic version of the algorithm has better performance was already observed in a number of works. Maybe the first theoretical results in this direction are [16, 17] in which the case where the target measure is Gaussian is studied in details. Of course a Gaussian target measure is of little interest for sampling but the point there was to show that when the friction parameter is set appropriately then the kinetic diffusion is faster than the overdamped one. As far as sampling is concerned, relevant references include [7, 11] in which non-asymptotic bounds for the kinetic Langevin algorithm are established. The two results are very similar and differ in two ways from the present article.

Firstly, the performance is measured in terms of the Wasserstein 2 distance (i.e. the optimal transportation associated to the cost function $|x-y|^2$) rather than total variation. Also the hypothesis is more stringent, the target measure is assumed to be uniformly log-concave and gradient Lipschitz. In other words the potential V is assumed to satisfy

$$mI_n \leq \nabla^2 V(x) \leq MI_n \tag{3}$$

for all x, and where m, M are positive constants. Notice that such a potential is gradient Lipschitz with constant $L \leq M$ and satisfies Poincaré with constant $C_P \leq m^{-1}$. Nevertheless, [7, 11] establish that all things equal (namely under the assumption (3) and as far as the Wasserstein distance is concerned) the underdamped version outperforms the best bounds available for the overdamped algorithm, which were established previously in [9, 13]. In particular, it is shown that the dimension dependence drops from $O(n)$ to $O(\sqrt{n})$, as in the present work.

In [22] a relative entropy estimate for the kinetic Langevin algorithm is established. Although relative entropy controls total variation (see Section 3 below) this does not recover our main result. First of all, the result from [22] is proven under the assumption that the target measure satisfies a log-Sobolev inequality, which is a stronger hypothesis than Poincaré, and more importantly it is only partly quantitative, in the sense that the dependence on certain parameters of the problem is not made explicit. The reference that comes closer to our work is [28] (which we were not aware of until the first version of this paper was released). At a high level, the results and methods of proof there are very similar to what is done in the current work. However, our analysis of the discretization error is much simpler and also allows to capture more accurately the dependence on the initial condition. In [28] the convergence is established for a specific warm start condition for which both the log chi-square divergence and the Fisher information are order n (essentially). In such a situation the authors of [28] obtain convergence in total variation after $O^*(\varepsilon^{-1}(LC_P)^{3/2}n^2)$ steps of the algorithm in general and $O^*(\varepsilon^{-1}LC_P n^2)$ steps in the log-concave case. While it is true that when $\log \chi_2(x_0 \mid \mu) = O^*(n)$ we get exactly the same complexity, our result has the advantage of not requiring bounded Fisher information. Also we get a better dimension dependence if one happens to have a better warm start hypothesis. In particular, as we mentioned already the dimension dependence becomes as low as $O^*(\sqrt{n})$ when the initial chi-square divergence is polynomial in the dimension. This does not seem to follow from the analysis of [28].

Let us discuss also the literature on the Hamiltonian Monte Carlo algorithm, which is a sampling algorithm very much related to the kinetic Langevin algorithm. It is based on the observation that the Hamiltonian dynamic

$$\begin{cases} y'(t) = x(t), \\ x'(t) = -\nabla V(y(t)) \end{cases} \tag{4}$$

preserves the Lebesgue measure as well as the potential $\mathscr{H}(x, y) := V(x) + |y|^2/2$. As a result it also preserves the probability measure π. The HMC process is the piecewise deterministic process obtained by choosing a time step δ, resampling the speed y at every integer multiple of δ and following the system of equations (4) in between. Of course this does not admit an explicit solution and in order to turn this ideal HMC dynamic into a proper algorithm one needs to replace the Hamiltonian dynamic phase by an Euler-type discretization. The resulting algorithm looks a lot like the kinetic Langevin algorithm, and in particular the two algorithms should have pretty much the same performance. Non-asymptotic theoretical bounds for the HMC algorithm are established in [6, 23, 24]. As for the works on the kinetic Langevin algorithm mentioned above, they prove convergence estimates in the Wasserstein sense and under the assumption that the Hessian of the potential of the target measure is bounded from above and below (by a positive constant). The result from [6] gives essentially the same estimate as what [7, 11] get for the kinetic Langevin algorithm.

We have only presented a very short selection of the literature on the Langevin and the HMC algorithms. For instance there are also many references where the focus is on the discretization scheme. Indeed, we used the most natural one in this paper but there exist variants in the literature. These typically yield better dependence on the error ε at the cost of higher order regularity estimates on the potential. See [4] and the references therein.

1.6. Perspectives

First, we conjecture that Theorem 1 should hold true for the HMC algorithm as well. Maybe more interestingly, one drawback with our result is that we lose regularity from the hypothesis to the conclusion. We prove a total variation estimate under a warm start hypothesis in the chi-square sense. It would be more satisfactory to get a chi-square estimate in the conclusion as well. Note that the chi-square divergence controls the Wasserstein distance under Poincaré (see [21]). Therefore, such a result would imply also a convergence estimate for the Wasserstein distance, similar to that from the aforementioned works [7, 11], but under significantly weaker hypothesis on the target measure. In the same way, it would be interesting to have an analogue result for the relative entropy (both in the hypothesis and in the conclusion) under a log-Sobolev hypothesis for the target measure. In the overdamped version of the algorithm this task was completed by Vempala and Wibisono [26] but it is not clear at all whether their approach can be adapted to the kinetic Langevin algorithm.

2. The hypocoercive estimate

Itô's formula shows that the generator of the kinetic diffusion (1) is the operator

$$\mathcal{L} f(x,y) = \beta \Delta_y f(x,y) + \nabla_x f(x,y) \cdot y - \beta \nabla_y f(x,y) \cdot y - \nabla_y f(x,y) \cdot \nabla V(x).$$

Integrating by parts we see that the probability measure π on $\mathbb{R}^n \times \mathbb{R}^n$ whose density is proportional to

$$\mathrm{e}^{-V(x)} \cdot \mathrm{e}^{-|y|^2/2}$$

is stationary. One important fact is that the diffusion (X_t, Y_t) is not reversible. In terms of the operator $\mathcal{L}$, this means that the operator is not symmetric in $L^2(\pi)$. More precisely, direct calculations show that the part

$$\Delta_y f(x,y) - \nabla_y f(x,y) \cdot y =: \mathcal{L}_{\mathrm{OU}} f$$

is symmetric and that the part

$$\nabla_x f(x,y) \cdot y - \nabla_y f(x,y) \cdot \nabla V(x) =: \mathcal{L}_{\mathrm{Ham}} f$$

is antisymmetric. The indices "OU" and "Ham" stand for Ornstein–Ulhenbeck and Hamiltonian, respectively. As a result the adjoint operator is $\mathcal{L}^* = \beta \mathcal{L}_{\mathrm{OU}} - \mathcal{L}_{\mathrm{Ham}}$.

We let (P_t) be the semigroup with generator $\mathcal{L}$, and (P_t^*) be the adjoint semigroup. In other words, if ν is a probability measure on $\mathbb{R}^n \times \mathbb{R}^n$ which is absolutely continuous with respect to π, with density f, then νP_t has density $P_t^* f$ with respect to π. Moreover, the chi-square divergence between ν and π is nothing but the variance of the relative density, so that $\chi_2(\nu P_t \mid \pi) = \mathrm{var}_\pi(P_t^* f)$. Therefore, Theorem 2 can be reformulated as follows.

Theorem 3. *If the potential of μ is $\mathcal{C}^2$-smooth and semi-convex, with constant $\kappa \geq 0$, and if μ satisfies Poincaré with constant C_P, then for every function $f \in L^2(\pi)$, we have*

$$\mathrm{var}_\pi(P_t^* f) \leq 2 \exp\left(-c \cdot \frac{\beta t}{1 + (\beta^2 + \kappa) C_P} \right) \mathrm{var}_\pi(f).$$

As we mentioned in the introduction, the difficulty arises from the degeneracy of the diffusion. Let us say a few more words about this. By definition

$$\partial_t P_t^* f = \mathcal{L}^* P_t^* f.$$

Integrating by parts, we see that

$$\frac{d}{dt} \operatorname{var}_\pi(P_t^* f) = 2 \int_{\mathbb{R}^{2n}} (\mathcal{L}^* P_t^* f) P_t^* f \, d\pi = -2\beta \int_{\mathbb{R}^{2n}} |\nabla_y P_t^* f|^2 \, d\pi.$$

This shows that the variance of $P_t^* f$ is non-increasing in time. If we want a more quantitative statement, the issue is that only the gradient in y appears in the dissipation of the variance, and not the full gradient. This comes from the fact that the associated diffusion is degenerate and that the Brownian term only appears in the y variable. Thus we cannot hope to lower bound the dissipation of variance by the variance itself.

There are a number of ways around this issue, and this line of research usually goes by the name of *hypocoercivity*. This was pioneered by Villani [27], other classical references include [2, 12] to name only a very few of them. In Villani's work the main idea is to consider the dissipation of some perturbed energy of the form

$$\mathcal{E}(f_t) := \operatorname{var}_\pi(f_t) + \int_{\mathbb{R}^n \times \mathbb{R}^n} \langle A \nabla f_t, \nabla f_t \rangle \, d\pi,$$

where A is a suitably chosen positive semi-definite matrix. This is also the approach taken by many of the subsequent works, including the works having application to sampling [7, 11] which we already mentioned. To arrive at Theorem 3, we take a slightly different route, inspired by the work of Albritton, Armstrong, Mourrat and Novack [1]. The main idea is that while an inequality of the form

$$\operatorname{var}_\pi(f) \leq C \int |\nabla_y f|^2 \, d\pi$$

is impossible, if we integrate this on some time interval along the kinetic Langevin diffusion, then the inequality becomes plausible. This is called *space-time* Poincaré inequality. This approach is quite general and not restricted to the case of the kinetic Langevin diffusion. However the results from [1] are mostly qualitative, whereas here we need quantitative estimates with explicit dependence on all parameters of the problem. A quantitative version of [1] was developed by Cao, Lu and Wang in [5] and the proof spelled out below is very much inspired by their argument.

Before embarking for the proof of Theorem 3 let us remark that it is invariant by scaling. Indeed, notice that if (X_t, Y_t) is a kinetic Langevin diffusion associated to μ and with friction parameter β, then $(\frac{1}{\lambda} X_{\lambda t}, Y_{\lambda t})$ is a kinetic Langevin diffusion associated to the measure μ scaled by $1/\lambda$ (i.e. the law of X/λ if X is a vector with law μ) and friction parameter $\lambda\beta$. This implies easily that if we define $\alpha = \alpha(t, \beta, \mu)$ to be the best estimate one could get for μ in this theorem, namely

$$\alpha(t, \beta, \mu) = \sup\left\{ \frac{\operatorname{var}_\pi(P_t^* f)}{\operatorname{var}_\pi(f)} \right\},$$

where the supremum is taken over every function $f \in L^2(\pi)$, then α has the property that

$$\alpha(t, \beta, \mu) = \alpha\Big(\frac{t}{\lambda}, \beta\lambda, \mu_{1/\lambda}\Big),$$

where $\mu_{1/\lambda}$ denotes the measure μ scaled by $1/\lambda$. The upper bound for α provided by Theorem 3 is also invariant by this change of variable. As a result it is enough to prove the result when one of the parameters has a prescribed value, we will exploit this observation later on.

2.1. The L^2 method for Poincaré

In this section we gather some well-known facts about the Laplace operator associated to some measure that will be needed later on. Let μ be a probability measure on $\mathbb{R}^n$, of the form

$$\mu(dx) = \mathrm{e}^{-V(x)}\, dx$$

for some $\mathcal{C}^2$-smooth function $V \colon \mathbb{R}^n \to \mathbb{R}$. The Laplace operator associated to μ is the differential operator L_μ defined by

$$L_\mu f = \Delta f - \langle \nabla f, \nabla V \rangle.$$

Originally, L_μ is defined on the space of $\mathcal{C}^\infty$-smooth and compactly supported functions f, in which case an integration by parts gives

$$\int_{\mathbb{R}^n} (L_\mu f) g \, d\mu = -\int_{\mathbb{R}^n} \langle \nabla f, \nabla g \rangle \, d\mu.$$

This shows in particular

$$\langle L_\mu f, g \rangle_{L^2(\mu)} = \langle f, L_\mu g \rangle_{L^2(\mu)},$$

and that $\langle L_\mu f, f \rangle_{L^2(\mu)} \leq 0$ for any functions f, g in the domain of L_μ. In the language of operator theory, the operator $-L_\mu$ is said to be symmetric and monotone. It turns out that L_μ admits a unique self-adjoint extension, the domain of which contains $H^1(\mu)$, and that the above integration by parts is true for every f, g in $H^1(\mu)$. Recall that $H^1(\mu)$ is the space of functions $f \in L^2(\mu)$ whose weak gradient also belongs to $L^2(\mu)$. The operator L_μ admits 0 as a simple a eigenvalue, the corresponding eigenspace consists of constant functions. We say that L_μ, or rather $-L_\mu$, has a spectral gap if there exists $\lambda_0 > 0$ such that the rest of the spectrum of $-L_\mu$ is included in some interval $[\lambda_0; +\infty)$. This property turns out to be equivalent to the Poincaré inequality, as we shall see now.

Lemma 4. *The measure μ satisfies Poincaré with constant C_P if and only if the spectral gap of the operator L_μ is at least C_P^{-1}. Moreover, this is also equivalent to the inequality*

$$\int_{\mathbb{R}^n} |\nabla f|^2 \, d\mu \le C_P \cdot \int_{\mathbb{R}^n} (L_\mu f)^2 \, d\mu$$

for every f in the domain of L_μ.

Lemma 5 (Bochner formula). *For every smooth and compactly supported function f, we have*

$$\int_{\mathbb{R}^n} (L_\mu f)^2 \, d\mu = \int_{\mathbb{R}^n} \|\nabla^2 f\|_F^2 + \nabla^2 V(\nabla f, \nabla f) \, d\mu.$$

Here and in the sequel

$$\|A\|_F = \Big(\sum_{i,j} a_{ij}^2 \Big)^{1/2}$$

denotes the Frobenius norm of a symmetric matrix $A = (a_{ij})$. If we combine together these two lemmas we easily get the so-called Lichnerowicz estimates: If the potential V of μ is uniformly convex, in the sense that there exists a constant $\alpha > 0$ such that $\nabla^2 V \ge \alpha \cdot \mathrm{Id}$ pointwise and for the order given by the cone of positive semi-definite matrices, then the measure μ satisfies Poincaré with constant $1/\alpha$. Indeed Bochner's formula then implies

$$\int_{\mathbb{R}^n} (L_\mu f)^2 \, d\mu \ge \alpha \int_{\mathbb{R}^n} |\nabla f| \, d\mu,$$

which yields the claimed bound on the Poincaré constant thanks to Lemma 4.

This approach for the Poincaré inequality dates back to the work of Lichnerowicz [20]. More recent references where this method plays a role include [8, 18] among others. We refer to these for the proofs of Lemmas 4 and 5. For our purposes the following immediate consequence of the two lemmas will be important.

Corollary 6. *If μ satisfies Poincaré with constant C_P and if the potential V of μ is $\mathcal{C}^2$-smooth and semi-convex with constant κ, then for every f in the domain of L_μ, we have*

$$\int_{\mathbb{R}^n} \|\nabla^2 f\|_F^2 \, d\mu \le (1 + \kappa \cdot C_P) \int_{\mathbb{R}^n} (L_\mu f)^2 \, d\mu.$$

Proof. Bochner's formula and the semi-convexity hypothesis yield

$$\int_{\mathbb{R}^n} (L_\mu f)^2 \, d\mu \ge \int_{\mathbb{R}^n} \|\nabla^2 f\|_F^2 \, d\mu - \kappa \int_{\mathbb{R}^n} |\nabla f|^2 \, d\mu.$$

By Lemma 4,

$$\int_{\mathbb{R}^n} |\nabla f|^2 \, d\mu \le C_P \int_{\mathbb{R}^n} (L_\mu f)^2 \, d\mu,$$

and the result follows. ■

2.2. A divergence equation

The next proposition, taken from the Cao, Lu, Wang paper, is the main ingredient in the proof of the hypocoercive estimate. This proposition is revisited and extended to a somewhat more general context in the preprint [14]. However, instead of just taking the result for granted we shall reprove it. One reason is that in the aforementioned works the proposition is proven under annoying additional technical assumptions which are actually not needed. Maybe more importantly the existing proofs are mostly computational and we find them hard to follow. Our proof relies on spectral theory and is arguably more conceptual. This hopefully sheds new light on this somewhat delicate proposition.

Proposition 7. *Suppose that the potential of μ is $\mathcal{C}^2$-smooth and semi-convex, with constant κ, and that μ satisfies Poincaré with constant* 1. *Fix a time horizon T and let*

$$f \in L^2\big([0,T] \times \mathbb{R}^n, \lambda \otimes \mu\big),$$

where λ is the Lebesgue measure, and assume that f is orthogonal to constants. Then there exist u and v in $L^2(\lambda \otimes \mu)$ satisfying the Dirichlet boundary conditions, namely $u(0,\cdot) = u(T,\cdot) = 0$, and similarly for v, such that

$$\partial_t u + L_\mu v = f, \tag{5}$$

and such that the following estimates hold true:

(i) $\|\nabla u\|_{L^2(\lambda\otimes\mu)} \lesssim T \|f\|_{L^2(\lambda\otimes\mu)}$,

(ii) $\|\nabla \partial_t v\|_{L^2(\lambda\otimes\mu)} \lesssim T^{-1} \|f\|_{L^2(\lambda\otimes\mu)}$,

(iii) $\|\nabla v\|_{L^2(\lambda\otimes\mu)} \lesssim \|f\|_{L^2(\lambda\otimes\mu)}$,

(iv) $\|\nabla^2 v\|_{L^2(\lambda\otimes\mu)} \lesssim (1+\kappa)^{1/2} \|f\|_{L^2(\lambda\otimes\mu)}$.

Here, the notation $\lesssim$ means up to a multiplicative universal constant. Also, when applied to tensors, norms should be interpreted coordinate-wise. For instance,

$$\|\nabla^2 v\|^2_{L^2(\lambda\otimes\mu)} = \sum_{ij} \|\partial_{ij} v\|^2_{L^2(\lambda\otimes\mu)}.$$

The results from the previous subsection section will play a role at some point but there is still some work to be done here. The main difficulty is that there is a mismatch between the boundary condition for u and that for v.

As a preliminary step to the proof of Proposition 7, we notice that it is enough to consider the case where $T = 1$. Indeed given three functions u, v, f in

$$L^2\big([0,T] \times \mathbb{R}^n, \lambda \otimes \mu\big),$$

we let $\tilde{u}$, $\tilde{v}$, $\tilde{f}$ be the functions in $L^2([0,1] \times \mathbb{R}^n, \lambda \otimes \mu)$ given by

$$\begin{aligned} \tilde{f}(t,x) &= f(tT,x), \\ \tilde{u}(t,x) &= T^{-1} \cdot u(tT,x), \\ \tilde{v}(t,x) &= v(tT,x). \end{aligned}$$

Obviously, f is orthogonal to constants if and only if $\tilde{f}$ is, and u, v satisfy the Dirichlet boundary conditions if and only if $\tilde{u}$, $\tilde{v}$ do. It is also clear that the divergence equation (5) for (u, v, f) is equivalent to that for $(\tilde{u}, \tilde{v}, \tilde{f})$. Lastly, elementary computations show that the Sobolev-type estimates (i), (ii), (iii), (iv) for u, v, f, T amount to the same estimates for $\tilde{u}$, $\tilde{v}$, $\tilde{f}$, 1. Therefore, we assume that $T = 1$ from now on.

Remark. In the same way, while we stated this proposition under the assumption that $C_P = 1$ (which will be sufficient for our needs later on), one can show a more general estimate that depends on the Poincaré constant of the measure μ by rescaling the space variable.

It will be convenient to introduce the following notations. We let $A = -\partial_t \partial_t$ with Dirichlet boundary conditions. By a slight abuse of notation we will view it either as an operator on $L^2(\lambda)$ or as an operator on $L^2(\lambda \otimes \mu)$ that acts on the time variable only. In both cases it is a self-adjoint positive semi-definite unbounded operator. We also let $B = -L_\mu$, and again we view this either as a positive semi-definite operator on $L^2(\mu)$ or on $L^2(\lambda \otimes \mu)$. We also denote by L_0^2 the subspace of L^2 consisting of functions which are orthogonal to constants. Lastly for $h \in L^2(\lambda \otimes \mu)$, we define

$$\Pi h(x) = \int_0^1 h(t,x)\, dt.$$

In other words, Πh is the second marginal of h. Note that $\Pi \colon L^2(\lambda \otimes \mu) \to L^2(\mu)$ is a bounded operator, and that its adjoint Π^* is the canonical injection of $L^2(\mu)$ into $L^2(\lambda \otimes \mu)$. Observe that both Π and Π^* preserve L_0^2. The next lemma is the key to the proof of Proposition 7.

Lemma 8. *Recall that we assume that $T = 1$. Consider the following operators:*

- $Q_1 := \Pi AB(A^2 + B^2)^{-1}\Pi^*$;
- $Q_2 := \Pi A^4(A^2 + B^2)^{-2}\Pi^*$;
- $Q_3 := \Pi A^2(A^2 + B^2)^{-1}$.

The following properties hold true:

(a) *Both Q_1 and Q_2 are bounded and positive semi-definite on $L^2(\mu)$ and leave the subspace $L^2_0(\mu)$ invariant. The operator Q_3 is bounded from $L^2(\lambda \otimes \mu)$ to $L^2(\mu)$ and maps $L^2_0(\lambda \otimes \mu)$ to $L^2_0(\mu)$.*

(b) *As operators on $L^2_0(\mu)$ the operators Q_1, Q_2 satisfy $Q_2 \le C \cdot Q_1$ for the order given by the cone of positive semi-definite operators, and where C is a universal constant.*

(c) *As operators on $L^2_0(\mu)$ both Q_1 and Q_2 are actually positive definite.*

(d) *On $L^2_0(\mu)$, we also have* $0 \le Q_3^* Q_1^{-1} Q_3 \le C \cdot \mathrm{Id}$.

Proof. We start with claim (a). Since A and B commute and are positive semi-definite, we have

$$0 \le AB(A^2 + B^2)^{-1} \le \frac{1}{2}\mathrm{Id}.$$

In particular, we note that $AB(B^2 + A^2)^{-1}$ is a bounded positive semi-definite operator on $L^2(\lambda \otimes \mu)$. Observe that $\mathbb{1}$ is an eigenfunction

$$AB(A^2 + B^2)^{-1}\mathbb{1} = A(B^2 + A^2)^{-1}B\mathbb{1} = 0.$$

Since $AB(A^2 + B^2)^{-1}$ is self adjoint, it also leaves the space $\mathrm{span}\{\mathbb{1}\}^\perp = L^2_0(\lambda \otimes \mu)$ invariant. Composing by Π and Π^* we thus see that Q is a positive semi-definite bounded operator on $L^2(\mu)$ leaving $L^2_0(\mu)$ invariant. In a similar way, we have

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

Thus $A^2(B^2 + A^2)^{-1}$ leaves $L^2_0(\lambda \otimes \mu)$ invariant. It is also clearly positive semi-definite and bounded. This implies easily that Q_2 is also a positive semi-definite operator on $L^2(\mu)$ that preserves $L^2_0(\mu)$, and that Q_3: $L^2(\lambda \otimes \mu) \to L^2(\mu)$ is bounded and maps $L^2_0(\lambda \otimes \mu)$ to $L^2_0(\mu)$.

We move one to the proof of (b). We introduce the sine basis of $L^2([0, 1], \lambda)$, namely

$$e_n(t) = \sqrt{2} \cdot \sin(\pi n t), \quad n \ge 1.$$

This is an orthonormal basis of $L^2([0, 1], \lambda)$ for which A is diagonal. More precisely,

$$Ae_n = n^2 e_n, \quad \forall n \ge 1.$$

An element $f \in L^2(\lambda \otimes \mu)$ can be written uniquely

$$f = \sum_{n \ge 1} e_n \otimes f_n,$$

where (f_n) is a sequence of elements of $L^2(\mu)$ such that the series

$$\sum_{n\geq 1} \| f_n \|^2_{L^2(\mu)}$$

is converging. Here for $h \in L^2(\lambda)$ and $k \in L^2(\mu)$, we denote by $h \otimes k$ the function that maps (t, x) to $h(t)k(x)$. Then the action of A is simply given by the equation

$$Af = \sum_{n\geq 1} n^2 e_n \otimes f_n. \tag{6}$$

A bit more precisely, f belongs to the domain of A if and only if the series

$$\sum_{n\geq 1} n^4 \| f_n \|^2_{L^2(\mu)}$$

is converging, in which case (6) holds true. In the same way, f belongs to the domain of B if every function f_n does and if the series

$$\sum_{n\geq 1} \| B f_n \|^2_{L^2(\mu)}$$

is converging. When this is the case we have

$$Bf = \sum_{n\geq 1} e_n \otimes (B f_n).$$

Therefore,

$$(A^2 + B^2) f = \sum_{n\geq 1} e_n \otimes \big((n^4 \mathrm{Id} + B^2) f_n \big),$$

and also

$$AB(A^2 + B^2)^{-1} f = \sum_{n\geq 1} e_n \otimes \big(n^2 B (n^4 \mathrm{Id} + B^2)^{-1} f_n \big). \tag{7}$$

The constant function admits the following decomposition in the sine basis:

$$\mathbb{1} = \frac{2\sqrt{2}}{\pi} \sum_{n \text{ odd}} \frac{e_n}{n}.$$

Therefore, for $g \in L^2_0(\mu)$, we have

$$\Pi^* g = \mathbb{1} \otimes g = \frac{2\sqrt{2}}{\pi} \sum_{n \text{ odd}} \frac{e_n \otimes g}{n},$$

and for $f = \sum_{n\geq 1} e_n \otimes f_n \in L^2_0(\lambda \otimes \mu)$, we have

$$\Pi f = \frac{2\sqrt{2}}{\pi} \sum_{n \text{ odd}} \frac{f_n}{n}.$$

Combining with (7), we thus get

$$Q_1 = \Pi AB(A^2 + B^2)^{-1}\Pi^* = \frac{8}{\pi^2} \sum_{n \text{ odd}} B(n^4 \mathrm{Id} + B^2)^{-1}.$$

In the same way, we have

$$Q_2 = \Pi A^4(A^2 + B^2)^{-2}\Pi^* = \frac{8}{\pi^2} \cdot \sum_{n \text{ odd}} n^6(n^4 \mathrm{Id} + B^2)^{-2}. \tag{8}$$

Therefore, the desired inequality (b) can be reformulated as

$$\sum_{n \text{ odd}} n^6(n^4 \mathrm{Id} + B^2)^{-2} \leq C \sum_{n \text{ odd}} B(n^4 \mathrm{Id} + B^2)^{-1}, \tag{9}$$

as self-adjoint operators on $L_0^2(\mu)$.

Recall that μ satisfies Poincaré with constant 1, which by Lemma 4 shows that the spectral gap of $B = -L_\mu$ is at least 1. This means that when we restrict to $L_0^2(\mu)$ the spectrum of B is included in the interval $[1, +\infty)$. We claim that this information alone yields (9). In other words, the inequality would be true for any Hilbert space, and any unbounded positive definite operator B whose spectrum lies in the interval $[1, \infty)$. Indeed, by the spectral theorem there is some resolution (E_λ) of the identity of $L_0^2(\mu)$ such that

$$B = \int_{\mathbb{R}} \lambda \, dE_\lambda. \tag{10}$$

Moreover, since the spectrum of B is above 1, the integral in (10) is actually supported on $[1, \infty)$ rather than on the whole line. Then for any integer $n \neq 0$ and any function $g \in L_0^2(\mu)$, we have

$$\langle B(n^4 \mathrm{Id} + B^2)^{-1} g, g \rangle_{L^2(\mu)} = \int_1^\infty \frac{\lambda}{n^4 + \lambda^2} \, \nu_g(d\lambda),$$

where ν_g is the spectral measure associated to g, namely the measure on $[1, \infty)$ whose distribution is given by

$$\nu_g\big([1, \lambda]\big) = \langle E_\lambda g, g \rangle_{L^2(\mu)}, \quad \forall \lambda \geq 1.$$

There is an analogous formula for $n^6(n^4 + B^2)^{-2}$ and (9) can thus be reformulated as

$$\int_1^\infty \sum_{n \text{ odd}} \frac{n^6}{(n^4 + \lambda^2)^2} \, \nu_g(d\lambda) \leq C \int_1^\infty \sum_{n \text{ odd}} \frac{\lambda}{n^4 + \lambda^2} \, \nu_g(d\lambda), \quad \forall g \in L_0^2(\mu).$$

But this would obviously follow from the pointwise inequality

$$\sum_{n \text{ odd}} \frac{n^6}{(n^4 + \lambda^2)^2} \leq C \sum_{n \text{ odd}} \frac{\lambda}{n^4 + \lambda^2}, \quad \forall \lambda \geq 1. \tag{11}$$

Thus all we have to do is to prove this relatively elementary inequality, which can be done as follows. Since $n^6 \leq (n^4+\lambda^2)^{3/2}$ it is enough to prove that

$$\sum_{n \text{ odd}} \frac{1}{(n^4+\lambda^2)^{1/2}} \leq C\lambda \sum_{n \text{ odd}} \frac{1}{n^4+\lambda^2}, \qquad \forall \lambda \geq 1.$$

Now observe that

$$\begin{aligned} \sum_{n \text{ odd}} \frac{1}{(n^4+\lambda^2)^{1/2}} &\leq \sum_{n \geq 1} \frac{1}{(n^4+\lambda^2)^{1/2}} \\ &\leq \int_0^\infty \frac{1}{(x^4+\lambda^2)^{1/2}}\, dx = C_1 \lambda^{-1/2}, \end{aligned} \tag{12}$$

where $C_1 = \int_0^\infty (x^4+1)^{-1/2}\, dx$. In a similar way,

$$\begin{aligned} \sum_{n \text{ odd}} \frac{1}{n^4+\lambda^2} &\geq \frac{1}{2} \sum_{n \geq 1} \frac{1}{n^4+\lambda^2} \\ &= -\frac{1}{2\lambda^2} + \frac{1}{2} \sum_{n \geq 0} \frac{1}{n^4+\lambda^2} \\ &\geq -\frac{1}{2\lambda^2} + \frac{1}{2} \int_0^\infty \frac{1}{x^4+\lambda^2}\, dx = -\frac{1}{2\lambda^2} + \frac{C_2}{\lambda^{3/2}}, \end{aligned}$$

where $C_2 = (1/2) \int_0^\infty (x^4+1)^{-1}\, dx$. This clearly implies that

$$\sum_{n \text{ odd}} \frac{1}{n^4+\lambda^2} \geq C_3 \lambda^{-3/2} \tag{13}$$

for all $\lambda \geq 1$ and some universal constant C_3. Putting (12) and (13) together yields (11) and finishes the proof of (b).

To prove (c), observe that if $g \in L_0^2(\mu)$ is different from 0 then the spectral measure ν_g is not identically zero. As a result

$$\langle n^6 (n^4+B^2)^{-4} g, g \rangle_{L^2(\mu)} = \int_1^\infty \frac{n^6}{(n^4+\lambda^2)^2} \nu_g\, (d\lambda) > 0.$$

Summing over odd integers and combining with (8) we see that Q_2 restricted to $L_0^2(\mu)$ is positive definite. By claim (b) this implies that also Q_1 is positive definite on $L_0^2(\mu)$.

Lastly, (a) and (c) imply that the operator $Q_3^* Q_1^{-1} Q_3$ is well defined and positive semi-definite on $L_0^2(\mu)$. A priori this operator could be unbounded. However, since $\Pi^* \Pi$ is the identity map, we have $Q_3 Q_3^* = Q_2$. As a result

$$(Q_3^* Q_1^{-1} Q_3)^2 = Q_3^* Q_1^{-1} Q_2 Q_1^{-1} Q_3.$$

Now (b) implies that

$$Q_1^{-1}Q_2Q_1^{-1} \leq C \cdot Q_1^{-1}Q_1Q_1^{-1} = C \cdot Q_1^{-1}.$$

Therefore, $(Q_3^*Q_1^{-1}Q_3)^2 \leq C \cdot Q_3^*Q_1^{-1}Q_3$. This implies that $Q_3^*Q_1^{-1}Q_3 \leq C \cdot \mathrm{Id}$, and finishes the proof of the lemma. ∎

Remark. As is apparent from this proof, Q_2 is also positive definite on $L^2(\mu)$, whereas constant functions belong to the kernel of Q_1. So it is important to restrict to $L_0^2(\mu)$ for the inequality $Q_2 \leq CQ_1$ to be valid.

We are now in a position to prove the key divergence estimate of Cao, Lu and Wang.

Proof of Proposition 7. We focus on the Sobolev estimates (i) and (ii) for now. For these two estimates the semi-convexity hypothesis is not needed, only the hypothesis on the Poincaré constant matters. We need to find u, v satisfying the Dirichlet boundary conditions, the equation $\partial_t u + L_\mu v = f$, and such that

$$\|\nabla u\|^2_{L^2(\lambda\otimes\mu)} + \|\partial_t v\|^2_{L^2(\lambda\otimes\mu)} \lesssim \|f\|^2_{L^2(\lambda\otimes\mu)}.$$

Note that the choice of the function v determines u. Namely, u is the anti-derivative of $f - L_\mu v = f + Bv$:

$$u(t,x) = \int_0^t f(s,x) + Bv(s,x)\,ds. \tag{14}$$

Then u satisfies the $t = 0$ boundary condition by definition, whereas the $t = 1$ boundary condition becomes

$$\int_0^1 f(t,x) + Bv(t,x)\,dt = 0.$$

In other words, $\Pi(f + Bv) = 0$. Since B commutes with Π this amounts to

$$\Pi v = -B^{-1}\Pi f. \tag{15}$$

Recall that f is assumed to be centered, so that $\Pi f \in L_0^2(\mu)$. Since B^{-1} is bounded operator on $L_0^2(\mu)$, the function $B^{-1}\Pi f$ is a well-defined element of $L_0^2(\mu)$. Integrating by parts in time and space, we see that

$$\|\nabla\partial_t v\|^2_{L^2(\lambda\otimes\mu)} = \langle ABv, v\rangle_{L^2(\lambda\otimes\mu)}.$$

In a similar way, if u satisfies (14) and the Dirichlet boundary condition, then

$$\|\nabla u\|^2_{L^2(\lambda\otimes\mu)} = \langle A^{-1}B(Bv + f), Bv + f\rangle_{L^2(\lambda\otimes\mu)}.$$

We thus have to consider the following optimization problem:

$$\text{minimize}\langle A^{-1}B(Bv+f), Bv+f\rangle_{L^2(\lambda\otimes\mu)} + \langle ABv, v\rangle_{L^2(\lambda\otimes\mu)},$$

among functions v satisfying the Dirichlet boundary condition as well as the constraint (15). We need to show that the value of this optimization problem is at most $\|f\|^2_{L^2(\lambda\otimes\mu)}$, up to a multiplicative universal constant. Formally, we can rewrite the optimization problem as

$$\begin{cases} \text{minimize} & \langle \mathcal{A}v, v\rangle_{L^2(\lambda\otimes\mu)} + 2\langle b, v\rangle_{L^2(\lambda\otimes\mu)} + c, \\ \text{under} & \Pi v = d, \end{cases} \tag{16}$$

where

$$\begin{gathered} \mathcal{A} = A^{-1}B(A^2+B^2), \\ b = A^{-1}B^2 f, \quad c = \langle A^{-1}Bf, f\rangle_{L^2(\mu)}, \quad d = B^{-1}\Pi f. \end{gathered} \tag{17}$$

This formulation is not quite legitimate. Indeed, f need not belong to the domain of B, so b and c are not really well-defined. We can nevertheless use this formulation to guess that the solution of the quadratic optimization problem should be

$$v^{\text{opt}} = \mathcal{A}^{-1}(-b+g), \tag{18}$$

where g is the Lagrange multiplier associated to the constraint. Thus g must belong to the orthogonal complement of the kernel of Π, which is also the range of Π^*. So there is $h \in L^2_0(\mu)$ such that $g = \Pi^* h$. Note that

$$\mathcal{A}^{-1}b = A(A^2+B^2)^{-1}f.$$

Thus the constraint $\Pi v = -B^{-1}\Pi f$ becomes

$$\begin{aligned} \Pi AB^{-1}(A^2+B^2)^{-1}\Pi^* h &= \Pi B(A^2+B^2)^{-1}f - B^{-1}\Pi f \\ &= -B^{-1}\Pi A^2(A^2+B^2)^{-1}f \end{aligned}$$

(recall that B commutes with Π). So formally h is given by

$$h = -B\big(\Pi AB(A^2+B^2)^{-1}\Pi^*\big)^{-1}\Pi A^2(A^2+B^2)^{-1}f.$$

This looks unwieldy but if we use the notations from Lemma 8 we can rewrite this as

$$B^{-1}h = -Q_1^{-1}Q_3 f. \tag{19}$$

Plugging back in (18), we get the following expression for the solution of the optimization problem:

$$\begin{aligned} v^{\text{opt}} &= -B(A^2+B^2)^{-1}f - A(A^2+B^2)^{-1}\Pi^* Q_1^{-1}Q_3 f \\ &= -B(A^2+B^2)^{-1}f - A^{-1}Q_3^* Q_1^{-1}Q_3 f. \end{aligned} \tag{20}$$

Although this computation was somewhat formal, this latest expression defines a genuine element of $L^2(\lambda \otimes \mu)$. Indeed, as we have seen before $AB(A^2 + B^2)^{-1}$ is a bounded operator. Also, when we restrict to centered functions, the operator $Q_3^* Q_1^{-1} Q_3$ is bounded, thanks to the last part of Lemma 8. Therefore,

$$AB(A^2 + B^2)^{-1} f + Q_3^* Q_1^{-1} Q_3 f$$

is a well-defined element of $L^2(\mu)$. Now (20) shows that v^{opt} is well defined and belongs to the range of A^{-1}, which equals the domain of A. In particular, v^{opt} satisfies the Dirichlet boundary condition.

Now we focus on the value of the optimization problem (16). From (18), we get

$$\text{value} = -\langle \mathcal{A}^{-1} b, b\rangle_{L^2(\lambda\otimes\mu)} + c + \langle \mathcal{A}^{-1} g, g\rangle_{L^2(\lambda\otimes\mu)}. \tag{21}$$

By (17), we have

$$\begin{aligned} -\langle \mathcal{A}^{-1} b, b\rangle_{L^2(\lambda\otimes\mu)} + c &= -\langle A^{-1} B^3 (A^2 + B^2)^{-1} f, f\rangle_{L^2(\lambda\otimes\mu)} \\ &\quad + \langle A^{-1} B f, f\rangle_{L^2(\lambda\otimes\mu)} \\ &= \langle AB(A^2 + B^2)^{-1} f, f\rangle_{L^2(\lambda\otimes\mu)} \\ &\leq \frac{1}{2}\|f\|^2_{L^2(\lambda\otimes\mu)}. \end{aligned} \tag{22}$$

On the other hand,

$$\begin{aligned} \langle \mathcal{A}^{-1} g, g\rangle_{L^2(\lambda\otimes\mu)} &= \langle AB^{-1}(A^2 + B^2)^{-1} \Pi^* h, \Pi^* h\rangle_{L^2(\lambda\otimes\mu)} \\ &= \langle \Pi AB(A^2 + B^2)^{-1} \Pi^* B^{-1} h, B^{-1} h\rangle_{L^2(\mu)} \\ &= \langle Q_1 B^{-1} h, B^{-1} h\rangle_{L^2(\mu)}. \end{aligned}$$

Plugging in (19) and using Lemma 8 again, we thus get

$$\begin{aligned} \langle \mathcal{A}^{-1} g, g\rangle_{L^2(\lambda\otimes\mu)} &= \langle Q_1 Q_1^{-1} Q_3 f, Q_1^{-1} Q_3 f\rangle_{L^2(\mu)} \\ &= \langle Q_3^* Q_1^{-1} Q_3 f, f\rangle_{L^2(\lambda\otimes\mu)} \\ &\leq C \cdot \|f\|^2_{L^2(\lambda\times\mu)}. \end{aligned} \tag{23}$$

Equations (21), (22) and (23) together show that the value of the optimization problem (16) is at most $((1/2) + C)\|f\|^2_{L^2(\lambda\times\mu)}$. This proves that there exist u, v satisfying the Dirichlet boundary condition, the divergence equation (5) and the Sobolev-type estimate

$$\|\nabla u\|^2_{L^2(\lambda\times\mu)} + \|\nabla \partial_t v\|^2_{L^2(\lambda\times\mu)} \leq \left(\frac{1}{2} + C\right) \cdot \|f\|^2_{L^2(\lambda\times\mu)}. \tag{24}$$

It remains to prove (iii) and (iv). This is where the semi-convexity hypothesis and the results of the previous subsection enter the picture. Since u, v satisfy the Dirichlet boundary condition, we have

$$\begin{aligned}-\langle \partial_t u, L_\mu v\rangle_{L^2(\lambda\otimes\mu)} &= \langle u, L_\mu \partial_t v\rangle_{L^2(\lambda\otimes\mu)} \\ &= -\langle \nabla u, \nabla\partial_t v\rangle_{L^2(\lambda\otimes\mu)} \\ &\leq \frac{1}{2}\|\nabla u\|^2_{L^2(\lambda\otimes\mu)} + \frac{1}{2}\|\nabla\partial_t v\|^2_{L^2(\lambda\otimes\mu)}.\end{aligned}$$

On the other hand, from the equation $\partial_t u + L_\mu v = f$, we get

$$\|f\|^2_{L^2(\lambda\otimes\mu)} = \|\partial_t u + L_\mu v\|^2_{L^2(\lambda\otimes\mu)} \geq \|L_\mu v\|^2_{L^2(\lambda\otimes\mu)} + 2\langle \partial_t u, L_\mu v\rangle_{L^2(\lambda\otimes\mu)}.$$

This, together with (24), yields

$$\begin{aligned}\|L_\mu v\|^2_{L^2(\lambda\otimes\mu)} &\leq \|f\|^2_{L^2(\lambda\otimes\mu)} + \|\nabla u\|^2_{L^2(\lambda\otimes\mu)} + \|\nabla\partial_t v\|^2_{L^2(\lambda\otimes\mu)} \\ &\leq \Big(\frac{3}{2} + C\Big)\|f\|^2_{L^2(\lambda\otimes\mu)}.\end{aligned} \tag{25}$$

Since the potential V of μ is κ semi-convex and since μ satisfies Poincaré with constant 1, we can apply Lemma 4 and Corollary 6 from the previous section to the function $v(t,\cdot)$. Integrating the two inequalities on $[0, 1]$ gives

$$\|\nabla v\|^2_{L^2(\lambda\otimes\mu)} \leq \|L_\mu v\|^2_{L^2(\lambda\otimes\mu)}, \quad \|\nabla^2 v\|^2_{L^2(\lambda\otimes\mu)} \leq (1+\kappa)\|L_\mu v\|^2_{L^2(\lambda\otimes\mu)}. \tag{26}$$

Equations (25) and (26) yield the estimates (iii) and (iv). ■

2.3. Space-time Poincaré inequality

We are now in a position to prove the integrated Poincaré inequality. The argument is essentially the same as that of Cao, Lu, Wang. We spell it out here for completeness. We fix a probability measure μ, and a friction parameter β, and we let (P_t^*) be the corresponding kinetic Fokker–Planck semigroup. Recall that the equilibrium measure is $\pi := \mu\otimes\gamma$, where γ is the standard Gaussian measure.

Proposition 9. *Assume that the potential of μ is $\mathscr{C}^2$-smooth and semi-convex, with constant κ, and that μ satisfies Poincaré with constant* 1. *Then for every $f \in L^2(\pi)$ and every $T > 0$, we have*

$$\int_0^T \operatorname{var}_\pi(P_t^* f)\,dt \lesssim (T^2 + T^{-2} + \beta^2 + \kappa)\cdot \int_{[0,T]\times\mathbb{R}^{2n}} |\nabla_y P_t^* f|^2\,dt\,d\pi.$$

Proof. Given a function g in $L^2(\pi)$, we denote by Mg its first marginal, namely

$$Mg(x) = \int_{\mathbb{R}^n} g(x,y)\,\gamma(dy).$$

We first note that it is enough to prove that

$$\int_0^T \operatorname{var}_\mu(MP_t^* f)\,dt \lesssim (T^2 + T^{-2} + \beta^2 + \kappa)\int_{[0,T]\times\mathbb{R}^{2n}} |\nabla_y P_t^* f|^2\,dt\,d\pi. \tag{27}$$

Indeed, we can decompose the variance and use the Gaussian Poincaré inequality to get

$$\begin{aligned}\operatorname{var}_\pi(P_t^* f) &= \int_{\mathbb{R}^n} \operatorname{var}_\gamma\big(P_t^* f(x,\cdot)\big)\,\mu(dx) + \operatorname{var}_\mu(MP_t^* f)\\ &\leq \int_{\mathbb{R}^{2n}} |\nabla_y P_t^* f|^2\,d\pi + \operatorname{var}_\mu(MP_t^* f).\end{aligned}$$

If we integrate on $[0, T]$ and combine with (27) we indeed obtain the desired inequality. It remains to prove (27). By a slight abuse of notation, we denote

$$f(t,x,y) = P_t^* f(x,y).$$

We can assume that $f(0,\cdot)$ is centered for π. This property is preserved along the diffusion, so for every fixed t, the function $f(t,\cdot)$ is centered for π, and the function $Mf(t,\cdot)$ is centered for μ. In particular,

$$\int_{[0,T]\times\mathbb{R}^n} Mf\,dt\,d\mu = 0.$$

We now invoke the previous proposition: There exist $u, v \in L^2(\lambda\otimes\mu)$ which vanish at $t = 0$ and $t = T$ such that

$$Mf = \partial_t u + L_\mu v,$$

and satisfying the following Sobolev-type estimates:

$$\begin{aligned}\|\nabla u\|_{L^2(\lambda\otimes\mu)} &\lesssim T\|Mf\|_{L^2(\lambda\otimes\mu)},\\ \|\nabla v\|_{L^2(\lambda\otimes\mu)} &\lesssim \|Mf\|_{L^2(\lambda\otimes\mu)},\\ \|\nabla\partial_t v\|_{L^2(\lambda\otimes\mu)} &\lesssim T^{-1}\|Mf\|_{L^2(\lambda\otimes\mu)},\\ \|\nabla^2 v\|_{L^2(\lambda\otimes\mu)} &\lesssim (1+\kappa)^{1/2}\|Mf\|_{L^2(\lambda\otimes\mu)}.\end{aligned} \tag{28}$$

Notice that as v is a function depending only on t and x but not on the y variable, we have

$$\mathcal{L}v(t,x) = \langle\nabla_x v(t,x), y\rangle$$

and also

$$\mathcal{L}^2 v(t,x) = \langle\nabla^2 v(t,x)y, y\rangle - \beta\langle\nabla v(t,x), y\rangle - \langle\nabla v(t,x), \nabla V(x)\rangle.$$

Integrating in the y variable and using the fact that the standard Gaussian has mean 0 and identity covariance matrix, we get

$$M\mathscr{L}^2 v = L_\mu v.$$

Therefore,

$$\begin{aligned}
\|Mf\|^2_{L^2(\lambda\otimes\mu)} &= \langle Mf, \partial_t u + L_\mu v\rangle_{L^2(\lambda\otimes\mu)} \\
&= \langle Mf, \partial_t u + M\mathscr{L}^2 v\rangle_{L^2(\lambda\otimes\mu)} \\
&= \langle f, \partial_t u\rangle_{L^2(\lambda\otimes\pi)} + \langle Mf, \mathscr{L}^2 v\rangle_{L^2(\lambda\otimes\pi)} \\
&= \langle f, \partial_t u + \mathscr{L}^2 v\rangle_{L^2(\lambda\otimes\pi)} - \langle f - Mf, \mathscr{L}^2 v\rangle_{L^2(\lambda\otimes M)}.
\end{aligned}$$

Moreover, since u vanishes on the boundary of $[0,T]\times\mathbb{R}^n$ and since f satisfies the kinetic Fokker–Planck equation, we have

$$\begin{aligned}
\langle f, \partial_t u\rangle_{L^2(\lambda\otimes\pi)} &= -\langle \partial_t f, u\rangle_{L^2(\lambda\otimes\pi)} \\
&= -\langle \mathscr{L}^* f, u\rangle_{L^2(\lambda\otimes\pi)} = -\langle f, \mathscr{L}u\rangle_{L^2(\lambda\otimes\pi)}.
\end{aligned}$$

In a similar way, since $v(t,x)$ vanishes when $t=0$ and $t=T$, the function

$$\mathscr{L}v(t,x,y) = \langle \nabla v(t,x), y\rangle$$

also has this property. As a result

$$\begin{aligned}
\langle f, \mathscr{L}^2 v\rangle_{L^2(\lambda\otimes\pi)} &= \langle \mathscr{L}^* f, \mathscr{L}v\rangle_{L^2(\lambda\otimes\pi)} \\
&= \langle \partial_t f, \mathscr{L}v\rangle_{L^2(\lambda\otimes\pi)} = -\langle f, \mathscr{L}\partial_t v\rangle_{L^2(\lambda\otimes\pi)}.
\end{aligned}$$

Therefore,

$$\|Mf\|^2_{L^2(\lambda\otimes\pi)} = -\langle f, \mathscr{L}u + \mathscr{L}\partial_t v\rangle_{L^2(\lambda\otimes\pi)} - \langle f - Mf, \mathscr{L}^2 v\rangle_{L^2(\lambda\otimes\pi)}.$$

Next we replace $\mathscr{L}^2 v$ by its expression and we observe that some terms cancel out. We finally obtain

$$\|Mf\|^2_{L^2(\lambda\otimes\pi)} = -\langle f, \mathscr{L}u + \mathscr{L}\partial_t v + \beta\mathscr{L}v\rangle_{L^2(\lambda\otimes\pi)} - \langle f - Mf, h\rangle_{L^2(\lambda\otimes\pi)}, \quad (29)$$

where h is the function given by

$$h(t,x,y) = \langle \nabla_x^2 v(t,x) y, y\rangle. \quad (30)$$

Now we will start writing inequalities. Recall that if w is a function not depending on y, then $\mathscr{L}w = \langle \nabla_x w, y\rangle$. Together with the Gaussian integration by parts formula:

$$\int_{\mathbb{R}^n} f(x,y) y\, \gamma(dy) = \int_{\mathbb{R}^n} \nabla_y f(x,y)\, \gamma(dy)$$

and Cauchy–Schwarz, we get

$$\langle f, \mathscr{L} w\rangle_{L^2(\lambda\otimes\pi)} = \langle \nabla_y f, \nabla_x w\rangle_{L^2(\lambda\otimes\pi)} \leq \|\nabla_y f\|_{L^2(\lambda\otimes\pi)} \cdot \|\nabla_x w\|_{L^2(\lambda\otimes\mu)}.$$

This allows to upper bound the first term of the right-hand side of (29). For the second term we use the Gaussian Poincaré inequality again. Note that if A is a fixed symmetric matrix, then the gradient of $\langle Ay, y\rangle$ is $2Ay$, and thus

$$\mathrm{var}_\gamma\big(\langle Ay, y\rangle\big) \leq 4 \int_{\mathbb{R}^n} |Ay|^2 \, d\gamma = 4\|A\|_F^2$$

(actually if we were tracking down constants we could save a factor 2 here). As a result, using Cauchy–Schwarz and the definition (30) of the function h, we get

$$\begin{aligned}\langle f - Mf, h\rangle_{L^2(\lambda\otimes\pi)} &\leq \int_{[0,T]\times\mathbb{R}^n} \mathrm{var}_\gamma\big(f(t,x,\cdot)\big)^{1/2} \cdot \mathrm{var}_\gamma\big(h(t,x,\cdot)\big)^{1/2} \, dt \, d\mu \\ &\leq \int_{[0,T]\times\mathbb{R}^n} \|\nabla_y f\|_{L^2(\gamma)} \cdot 2\|\nabla_x^2 v\|_F \, dt \, d\mu \\ &\leq 2\|\nabla_y f\|_{L^2(\lambda\otimes\pi)} \cdot \|\nabla_x^2 v\|_{L^2(\lambda\otimes\mu)}.\end{aligned}$$

Putting everything together, we obtain

$$\|Mf\|_{L^2(\lambda\otimes\mu)}^2 \leq D \cdot \|\nabla_y f\|_{L^2(\lambda\otimes\pi)}$$

with

$$D = \|\nabla_x u\|_{L^2(\lambda\otimes\mu)} + \|\nabla_x \partial_t v\|_{L^2(\lambda\otimes\mu)} + \beta\|\nabla_x v\|_{L^2(\lambda\otimes\mu)} + 2\|\nabla_x^2 v\|_{L^2(\lambda\otimes\mu)}.$$

Plugging in the bounds (28), we finally get

$$\|Mf\|_{L^2(\lambda\otimes\mu)} \lesssim \big(T + T^{-1} + \beta + \sqrt{1+\kappa}\big)\|\nabla_y f\|_{L^2(\lambda\otimes\pi)}.$$

This implies (27) and finishes the proof of the proposition. ■

2.4. Proof of the hypocoercive estimate

We can now finally prove the hypocoercive estimate, Theorem 3. Again that part of the argument is pretty much the same as in the Cao, Lu, Wang paper and we include it for completeness.

As we already mentioned there is some homogeneity in the problem which allows us to rescale the measure μ. In particular, it is enough to prove the result assuming $C_P = 1$, say. By the previous proposition, given $f \in L^2(\pi)$, we then have

$$\int_0^T \mathrm{var}_\pi(P_t^* f) \, dt \lesssim \Big(T^2 + \frac{1}{T^2} + \beta^2 + \kappa\Big) \int_{[0,T]\times\mathbb{R}^{2n}} |\nabla_y P_t^* f|^2 \, dt \, d\pi.$$

Since

$$\frac{d}{dt}\operatorname{var}_\pi(P_t^* f) = -2\beta \int_{\mathbb{R}^{2n}} |\nabla_y P_t^* f|^2 \, d\pi \le 0,$$

the previous inequality implies that

$$\begin{aligned} T \cdot \operatorname{var}_\pi(P_T^* f) &\le \int_0^T \operatorname{var}_\pi(P_t^* f)\, dt \\ &\lesssim \frac{1}{\beta}\Big(T^2 + \frac{1}{T^2} + \beta^2 + \kappa\Big)\big(\operatorname{var}_\pi(f) - \operatorname{var}_\pi(P_T^* f)\big). \end{aligned}$$

Hence,

$$\operatorname{var}_\pi(P_T^* f) \le \frac{1}{1 + x(T)} \operatorname{var}_\pi(f),$$

where

$$x(T) = \frac{\beta T}{C(T^2 + 1/T^2 + \beta^2 + \kappa)},$$

and where C is a universal constant, which we can assume to be larger than 1. Since $\beta^2 + T^2 \ge 2\beta T$, we have $x(T) \le 1/(2C) \le 1/2$, no matter what the value of T and β. For $x \le 1/2$, we certainly have $1/(1+x) \le \mathrm{e}^{-x/2}$, and thus

$$\operatorname{var}_\pi(P_T^* f) \le \mathrm{e}^{-x(T)/2} \operatorname{var}_\pi(f)$$

for all $T > 0$. Applying this inequality to $P_T^* f$, then $P_{2T}^* f$, and so on, and using the semigroup property, we get

$$\operatorname{var}_\pi(P_{nT}^* f) \le \mathrm{e}^{-n \cdot x(T)/2} \operatorname{var}_\pi(f)$$

for every $T > 0$ and every integer n. Equivalently,

$$\operatorname{var}_\pi(P_t^* f) \le \mathrm{e}^{-n \cdot x(t/n)/2} \operatorname{var}_\pi(f)$$

for every $t > 0$ and every integer $n \neq 0$. It remains to optimize on n. Since

$$n \cdot x\Big(\frac{t}{n}\Big) = \frac{t\beta}{C(t^2/n^2 + n^2/t^2 + \beta^2 + \kappa)},$$

assuming $t \ge 1$ and choosing n to be the integer part of t yields

$$\operatorname{var}_\pi(P_t^* f) \le \exp\Big(-c \cdot \frac{t\beta}{1 + \beta^2 + \kappa}\Big) \operatorname{var}_\pi(f), \quad \forall t \ge 1,$$

and where c is a small universal constant. For $t \le 1$ we can just use the trivial bound $\operatorname{var}_\pi(P_t^* f) \le \operatorname{var}_\pi(f)$, so that the last inequality implies

$$\operatorname{var}_\pi(P_t^* f) \le 2 \cdot \exp\Big(-c'' \cdot \frac{t\beta}{1 + \beta^2 + \kappa}\Big) \operatorname{var}_\pi(f), \quad \forall t \ge 0,$$

which is the result.

Remark. The proof also yields a non-trivial estimate for small values of t, namely

$$\operatorname{var}_\pi(P_t^* f) \leq \exp\Big(-c \cdot \frac{t\beta}{1 + t^{-2} + \beta^2 + \kappa}\Big) \operatorname{var}_\pi(f).$$

3. The discretization argument

It remains to estimate the discretization error. The approach is taken from Dalalyan's article [10] and relies on the Girsanov change of measure formula. The relative entropy (a.k.a. Kullback divergence) plays an important role in this approach. Recall its definition: If μ, ν are probability measures defined on the same space X, then

$$D(\nu \mid \mu) = \int_X \log\Big(\frac{d\nu}{d\mu}\Big)\, d\nu,$$

assuming that ν is absolutely continuous with respect to μ. If this is not the case the convention is that the relative entropy equals $+\infty$. An important feature of relative entropy is that it controls the total variation distance:

$$TV(\mu, \nu) \leq \sqrt{\frac{1}{2} D(\nu \mid \mu)}.$$

This is known as Pinsker's inequality. Another property that we will need is that taking a marginal can only lower the relative entropy. More generally, if $T \colon X \to Y$ is any measurable map, then

$$D(T\#\nu \mid T\#\mu) \leq D(\nu \mid \mu). \tag{31}$$

Here $T\#\mu$ denotes the pushforward of the measure μ by the map T. Note that this contraction property is not specific to the Kullback divergence, the chi-square divergence also has this property. Lastly we will need the following elementary lemma, which can be thought of as an approximate triangle inequality for the relative entropy.

Lemma 10. *Let μ_1, μ_2, μ_3 be probability measures defined on the same space E. Then*

$$D(\mu_3 \mid \mu_1) \leq 2D(\mu_3 \mid \mu_2) + \log\big(1 + \chi_2(\mu_2 \mid \mu_1)\big).$$

Proof. We can assume that $\mu_3 \prec \mu_2 \prec \mu_1$, where the symbol $\prec$ denotes absolute continuity. Indeed, if μ_3 is not absolute continuous with respect to μ_2 or if μ_2 is not absolutely continuous with respect to μ_1 then the inequality trivially holds true. We can also assume that $\mu_1 \prec \mu_2 \prec \mu_3$ (and thus that all three measures are equivalent) by some limiting argument. Then

$$\begin{aligned} D(\mu_3 \mid \mu_1) &= \int_E \log\Big(\frac{d\mu_3}{d\mu_1}\Big)\, d\mu_3 \\ &= D(\mu_3 \mid \mu_2) + \int_E \log\Big(\frac{d\mu_2}{d\mu_1}\Big)\, d\mu_3. \end{aligned} \tag{32}$$

Moreover, by Jensen's inequality, we have

$$\begin{aligned} \int_E \log f \, d\mu_3 &= \int_E \log\Big(f \cdot \frac{d\mu_1}{d\mu_3}\Big)\, d\mu_3 + D(\mu_3 \mid \mu_1) \\ &\le \log\Big(\int_E f \, d\mu_1\Big) + D(\mu_3 \mid \mu_1) \end{aligned}$$

for any positive function f. Applying this to $f = (d\mu_2/d\mu_1)^2$, we get

$$2 \int_X \log\Big(\frac{d\mu_2}{d\mu_1}\Big)\, d\mu_3 \le \log\big(1 + \chi_2(\mu_2 \mid \mu_1)\big) + D(\mu_3 \mid \mu_1).$$

Plugging this back into (32) yields the result. ■

Proposition 11. *Assume that V is gradient Lipschitz, with Lipschitz constant L. Given a probability measure ν on $\mathbb{R}^{2n}$ let ν_t be the law of the discretized kinetic Langevin diffusion (2) starting from ν at time t, and recall that νP_t denotes the law of the true diffusion at time t. Then*

$$TV(\nu_t, \nu P_t) \lesssim \frac{\sqrt{t} \cdot L\eta}{\sqrt{\beta}} \cdot \big(\sqrt{n} + \sqrt{\log(1 + \chi_2(\nu \mid \pi))}\big),$$

where η is the time step and β is the friction parameter.

Proof. We use the following version of Girsanov: Let (W_t) be a standard Brownian motion on $\mathbb{R}^n$, defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ equipped with some filtration $(\mathcal{F}_t)$. Let (X_t) be a process of the form $X_t = W_t + \int_0^t u_s \, ds$, where (u_t) is a progressively measurable process satisfying some integrability conditions. Fix a time horizon T. The new measure $\mathbb{Q}$ defined by

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = \exp\Big(-\int_0^t u_s \, dW_s - \frac{1}{2}\int_0^t |u_s|^2 \, ds\Big) \tag{33}$$

is a probability measure under which the process $(X_t)_{t \le T}$ is a standard Brownian motion.

Now consider the solution (X^η, Y^η) of the discretized Langevin equation (2) initiated at some measure ν, in the sense that the starting point (X_0^η, Y_0^η) has law ν, and consider the process $(\widetilde{W}_t)$ given by

$$\widetilde{W}_t = W_t + \frac{1}{\sqrt{2\beta}} \int_0^t \big(\nabla V(X_s^\eta) - \nabla V(X_{\lfloor s/\eta \rfloor \eta}^\eta)\big)\, ds.$$

Observe that by definition of $\widetilde{W}$, we have

$$\begin{cases} dX_t^\eta = Y_t^\eta \, dt, \\ dY_t^\eta = \sqrt{2\beta}\, d\widetilde{W}_t - \beta Y_t^\eta \, dt - \nabla V(X_t^\eta)\, dt. \end{cases}$$

In other words, if we replace W by $\widetilde{W}$ then (X_t^η, Y_t^η) becomes a genuine kinetic Langevin diffusion. According to Girsanov, if we set

$$u_s = (2\beta)^{-1/2}\big(\nabla V(X_s^\eta) - \nabla V(X_{\lfloor s/\eta \rfloor \eta}^\eta)\big)$$

and define $\mathbb{Q}$ by (33), then $\widetilde{W}$ is standard Brownian motion under $\mathbb{Q}$, so that (X^η, Y^η) is a genuine kinetic Langevin diffusion under $\mathbb{Q}$. Therefore,

$$(X_t^\eta, Y_t^\eta)\#\mathbb{Q} = \nu P_t \quad \text{and} \quad (X_t^\eta, Y_t^\eta)\#\mathbb{P} = \nu_t.$$

Hence, from (31),

$$D(\nu_t \mid \nu P_t) \le D(\mathbb{P} \mid \mathbb{Q}).$$

Moreover, since $\int u_s \cdot dW_s$ is a martingale having expectation 0,

$$D(\mathbb{P} \mid \mathbb{Q}) = \mathbb{E} \log \frac{d\mathbb{P}}{d\mathbb{Q}} = \frac{1}{4\beta} \int_0^t \mathbb{E}|\nabla V(X_s^\eta) - \nabla V(X_{\lfloor s/\eta \rfloor \eta}^\eta)|^2 \, ds.$$

Hence, the estimate

$$D(\nu_t \mid \nu P_t) \le \frac{1}{4\beta} \int_0^t \mathbb{E}|\nabla V(X_s^\eta) - \nabla V(X_{\lfloor s/\eta \rfloor \eta}^\eta)|^2 \, ds. \tag{34}$$

Strictly speaking the Girsanov change of measure formula is only valid under some integrability condition on the drift (u_t). For instance (u_t) uniformly bounded on $[0, T]$ is enough. However using some localization argument one can get around this issue and show that inequality (34) remains valid under no further assumption. See [15] for more details. Suppose that t is an integer multiple of η. Fix an integer $k \le N - 1$, where $N = t/\eta$. Since ∇V is Lipschitz with constant L, for every $s \in [k\eta, (k+1)\eta]$, we get from Cauchy–Schwarz

$$\mathbb{E}|\nabla V(X_s^\eta) - \nabla V(X_{k\eta}^\eta)|^2 \le L^2 \mathbb{E}|X_s^\eta - X_{k\eta}^\eta|^2 \le L^2 \eta \int_{k\eta}^s \mathbb{E}|Y_u^\eta|^2 \, du.$$

Integrating this inequality on the interval $s \in [k\eta, (k+1)\eta]$, summing over $k \leq N-1$, and plugging back in (34), we finally get

$$D(\nu_t \mid \nu P_t) \leq \frac{L^2\eta^2}{4\beta} \int_0^t \mathbb{E}|Y_s^\eta|^2 \, ds. \tag{35}$$

It remains to control $\mathbb{E}|Y_s^\eta|^2$. Intuitively the law of Y_s^η should not be too far away from that of the second factor of the equilibrium measure, which is the standard Gaussian γ. Therefore we should have $\mathbb{E}|Y_s^\eta|^2 \lesssim n$. Let $\nu_{s,2}$ be the second marginal of ν_s, namely the law of Y_s^η. Observe that if Y, G is a coupling of $(\nu_{s,2}, \gamma)$, then

$$\mathbb{E}|Y_s^\eta|^2 = \mathbb{E}|Y|^2 \leq 2\mathbb{E}|Y_s - G|^2 + 2\mathbb{E}|G|^2 = 2\mathbb{E}|Y_s - G|^2 + 2n.$$

Taking the infimum over such couplings yields

$$\mathbb{E}|Y_s^\eta|^2 \leq 2T_2(\nu_{s,2}, \gamma) + 2n,$$

where T_2 is the transportation distance from $\nu_{2,s}$ to γ associated to the cost function $|x-y|^2$. Next we invoke Talagrand's inequality [25]:

$$T_2(\nu_{s,2}, \gamma) \leq 2D(\nu_{s,2} \mid \gamma).$$

Combining this with the contraction principle and Lemma 10, we get

$$\begin{aligned}\mathbb{E}|Y_s^\eta|^2 &\leq 2n + 4D(\nu_{s,2} \mid \gamma) \\ &\leq 2n + 4D(\nu_s \mid \pi) \\ &\leq 2n + 8D(\nu_s \mid \nu P_s) + 4\log\bigl(1 + \chi_2(\nu P_s \mid \pi)\bigr).\end{aligned}$$

Recall that $\chi_2(\nu P_s \mid \pi)$ is non-increasing with s (see the previous section). Together with (35), we thus obtain

$$\mathbb{E}|Y_t^\eta|^2 \leq 2n + 4\log\bigl(1 + \chi_2(\nu \mid \pi)\bigr) + \frac{2L^2\eta^2}{\beta} \int_0^t \mathbb{E}|Y_s^\eta|^2 \, ds$$

for all $t > 0$. Using Gronwall's lemma and the convexity on the exponential function, we see that this implies

$$\int_0^t \mathbb{E}|Y_s^\eta|^2 \, ds \lesssim t \cdot \bigl(n + \log\bigl(1 + \chi_2(\nu \mid \pi)\bigr)\bigr)$$

as long as $2tL^2\eta^2 \leq \beta$. Plugging this back into (35) yields

$$D(\nu_t \mid \nu P_t) \lesssim \frac{tL^2\eta^2}{\beta} \cdot \bigl(n + \log\bigl(1 + \chi_2(\nu \mid \pi)\bigr)\bigr).$$

Combining this with Pinsker's inequality, we get the result under the additional hypothesis that $2tL^2\eta^2 \leq \beta$. However, since the total variation is anyways less than 1, if this additional hypothesis is violated then the result trivially holds true. ■

4. Proof of the main result

In this section we wrap up the proof of Theorem 1.

Let (x_k, y_k) be the Markov chain induced by the discretization with time step η of the kinetic Langevin diffusion and assume that the law of the initial point is a product measure whose second factor is γ. We write

$$TV(x_k, \mu) \leq TV\big((x_y, y_k), \pi\big) \leq TV\big((x_k, y_k), (X_t, Y_t)\big) + TV\big((X_t, Y_t), \pi\big),$$

where $t = k\eta$. Recall that the notation $TV((x, y), \pi)$ stands for the total variation between the law of (x, y) and π. Also, as the reader probably guessed already, (X_t, Y_t) denotes the genuine kinetic Langevin diffusion initiated at the same point as the algorithm. To upper bound the first term we invoke the latest proposition. For the second term we use the bound

$$TV\big((X_t, Y_t), \pi\big) \leq \sqrt{\chi_2((X_t, Y_t) \mid \pi)}$$

and we apply Theorem 2. Since

$$\chi_2\big((X_0, Y_0) \mid \pi\big) = \chi_2\big((x_0, y_0) \mid \pi\big) = \chi_2(x_0 \mid \mu),$$

we finally arrive at

$$\begin{aligned} TV(x_k, \mu) &\lesssim \frac{L\eta\sqrt{t}}{\sqrt{\beta}} \cdot \big(\sqrt{n} + \sqrt{\log(1 + \chi_2(x_0 \mid \mu))}\big) \\ &\quad + \exp\Big(-\frac{ct\beta}{1 + (\beta^2 + \kappa)C_P}\Big)\sqrt{\chi_2(x_0 \mid \mu)}. \end{aligned} \tag{36}$$

We still have the freedom to optimize on β. Let us focus on the log-concave case for now. Namely, we assume that $\kappa = 0$. In this situation the best choice is $\beta = C_p^{-1/2}$ in which case the latest displays becomes

$$\begin{aligned} TV(x_k \mid \mu) &\lesssim L\eta C_P^{1/4} t^{1/2} \cdot \big(\sqrt{n} + \sqrt{\log(1 + \chi_2(x_0 \mid \mu))}\big) \\ &\quad + \exp\Big(-\frac{ct}{\sqrt{C_P}}\Big)\sqrt{\chi_2(x_0 \mid \mu)} \end{aligned}$$

Fix $\varepsilon \in (0, 1)$. For the second term to be of order ε we need to take

$$t \approx \sqrt{C_P} \cdot \log\Big(\frac{\chi_2(x_0 \mid \mu)}{\varepsilon}\Big).$$

Then we adjust η so that the first term equals ε, and the corresponding number of steps is

$$\begin{aligned} k = \frac{t}{\eta} &\approx \varepsilon^{-1} C_P^{1/4} L t^{3/2} \cdot \big(\sqrt{n} + \sqrt{\log(1 + \chi_2(x_0 \mid \mu))}\big) \\ &\approx \varepsilon^{-1} L C_P \cdot C(n, \varepsilon, x_0), \end{aligned}$$

where

$$C(n,\varepsilon,x_0) = \big(\sqrt{n} + \sqrt{\log(1+\chi_2(x_0 \mid \mu))}\big) \cdot \log\Big(\frac{\chi_2(x_0 \mid \mu)}{\varepsilon}\Big)^{3/2}.$$

This finishes the proof of the main result in the log-concave case.

For the general case, observe that being gradient Lipschitz is a stronger assumption than being semi-convex. We can thus replace κ by L in (36). Note that we always have $LC_P \geq 1$. This follows from Caffarelli's contraction theorem [3]. Therefore,

$$\exp\Big(-\frac{ct\beta}{1+(\beta^2+L)C_P}\Big) \leq \exp\Big(-\frac{c't\beta}{(\beta^2+L)C_P}\Big).$$

Choosing $\beta = \sqrt{L}$ then yields

$$\begin{aligned} TV(x_k,\mu) \lesssim \eta L^{3/4} t^{1/2} \cdot \big(\sqrt{n} + \sqrt{\log(1+\chi_2(x_0 \mid \mu))}\big) \\ + \exp\Big(-\frac{c''t}{\sqrt{LC_P}}\Big)\sqrt{\chi_2(x_0 \mid \mu)}. \end{aligned}$$

We then conclude as in the log-concave case.

Acknowledgments. The author is grateful to Arnaud Guillin, Pierre Monmarché and Matthew Zhang for discussions related to this work and/or pointing out relevant references. We would also like to thank the two anonymous referees for their careful reading of the manuscript and accurate comments.

Funding. This work has been partially supported by the Project CONVIVIALITY ANR-23-CE40-0003 of the French National Research Agency.

References

[1] D. Albritton, S. Armstrong, J.-C. Mourrat, and M. Novack, Variational methods for the kinetic Fokker–Planck equation. *Anal. PDE* **17** (2024), no. 6, 1953–2010 Zbl 1547.35680 MR 4776290

[2] F. Baudoin, Bakry–émery meet Villani. *J. Funct. Anal.* **273** (2017), no. 7, 2275–2291 Zbl 1373.35061 MR 3677826

[3] L. A. Caffarelli, Monotonicity properties of optimal transportation and the FKG and related inequalities. *Comm. Math. Phys.* **214** (2000), no. 3, 547–563 Zbl 0978.60107 MR 1800860

[4] E. Camrud, A. Durmus, P. Monmarché, and G. Stoltz, Second order quantitative bounds for unadjusted generalized Hamiltonian Monte Carlo. [v1] 2023, [v2] 2024, arXiv:2306.09513v2

[5] Y. Cao, J. Lu, and L. Wang, On explicit L^2-convergence rate estimate for underdamped Langevin dynamics. *Arch. Ration. Mech. Anal.* **247** (2023), no. 5, article no. 90 Zbl 07754930 MR 4632836

[6] Z. Chen and S. S. Vempala, Optimal convergence rate of Hamiltonian Monte Carlo for strongly logconcave distributions. *Theory Comput.* **18** (2022), article no. 9 Zbl 1547.68837 MR 4430734

[7] X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan, Underdamped Langevin MCMC: A non-asymptotic analysis. In *Proceedings of the 31st Conference On Learning Theory*, pp. 300–323, Proceedings of Machine Learning Research 75, PMLR, 2018

[8] D. Cordero-Erausquin, M. Fradelizi, and B. Maurey, The (B) conjecture for the Gaussian measure of dilates of symmetric convex sets and related problems. *J. Funct. Anal.* **214** (2004), no. 2, 410–427 Zbl 1073.60042 MR 2083308

[9] A. S. Dalalyan, Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. In *Proceedings of the 30th Conference on Learning Theory*, pp. 678–689, Proceedings of Machine Learning Research 65, PMLR, 2017

[10] A. S. Dalalyan, Theoretical guarantees for approximate sampling from smooth and log-concave densities. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** (2017), no. 3, 651–676 Zbl 1411.62030 MR 3641401

[11] A. S. Dalalyan and L. Riou-Durand, On sampling from a log-concave density using kinetic Langevin diffusions. *Bernoulli* **26** (2020), no. 3, 1956–1988 Zbl 07193949 MR 4091098

[12] J. Dolbeault, C. Mouhot, and C. Schmeiser, Hypocoercivity for linear kinetic equations conserving mass. *Trans. Amer. Math. Soc.* **367** (2015), no. 6, 3807–3828 Zbl 1342.82115 MR 3324910

[13] A. Durmus and É. Moulines, High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli* **25** (2019), no. 4A, 2854–2882 Zbl 1428.62111 MR 4003567

[14] A. Eberle, A. Guillin, L. Hahn, F. Lörler, and M. Michel, Convergence of non-reversible Markov processes via lifting and flow Poincaré inequality. 2025, arXiv:2503.04238v1

[15] H. Föllmer, An entropy approach to the time reversal of diffusion processes. In *Stochastic differential systems (Marseille–Luminy, 1984)*, pp. 156–163, Lect. Notes Control Inf. Sci. 69, Springer, Berlin, 1985 Zbl 0562.60083 MR 0798318

[16] S. Gadat and L. Miclo, Spectral decompositions and $\mathbb{L}^2$-operator norms of toy hypocoercive semi-groups. *Kinet. Relat. Models* **6** (2013), no. 2, 317–372 Zbl 1262.35134 MR 3030715

[17] A. Guillin and P. Monmarché, Optimal linear drift for the speed of convergence of an hypoelliptic diffusion. *Electron. Commun. Probab.* **21** (2016), article no. 74 Zbl 1354.60084 MR 3568348

[18] B. Klartag, A Berry–Esseen type inequality for convex bodies with an unconditional basis. *Probab. Theory Related Fields* **145** (2009), no. 1-2, 1–33 Zbl 1171.60322 MR 2520120

[19] J. Lehec, The Langevin Monte Carlo algorithm in the non-smooth log-concave case. *Ann. Appl. Probab.* **33** (2023), no. 6A, 4858–4874 Zbl 07789649 MR 4674066

[20] A. Lichnerowicz, *Géométrie des groupes de transformations*. Trav. Rech. Math. 3, Dunod, Paris, 1958 Zbl 0096.16001 MR 0124009

[21] Y. Liu, The Poincaré inequality and quadratic transportation-variance inequalities. *Electron. J. Probab.* **25** (2020), article no. 1 Zbl 1448.60051 MR 4053901

[22] Y.-A. Ma, N. S. Chatterji, X. Cheng, N. Flammarion, P. L. Bartlett, and M. I. Jordan, Is there an analog of Nesterov acceleration for gradient-based MCMC? *Bernoulli* **27** (2021), no. 3, 1942–1992 Zbl 1475.62123 MR 4278799

[23] O. Mangoubi and A. Smith, Mixing of Hamiltonian Monte Carlo on strongly log-concave distributions 2: Numerical integrators. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pp. 586–595, Proceedings of Machine Learning Research 89, PMLR, 2017

[24] O. Mangoubi and A. Smith, Mixing of Hamiltonian Monte Carlo on strongly log-concave distributions: Continuous dynamics. *Ann. Appl. Probab.* **31** (2021), no. 5, 2019–2045 Zbl 1476.60112 MR 4332690

[25] M. Talagrand, Transportation cost for Gaussian and other product measures. *Geom. Funct. Anal.* **6** (1996), no. 3, 587–600 Zbl 0859.46030 MR 1392331

[26] S. Vempala and A. Wibisono, Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. In *Advances in Neural Information Processing Systems 32*, Curran Associates, 2019

[27] C. Villani, Hypocoercivity. *Mem. Amer. Math. Soc.* **202** (2009), no. 950 Zbl 1197.35004 MR 2562709

[28] M. S. Zhang, S. Chewi, M. Li, K. Balasubramanian, and M. A. Erdogdu, Improved discretization analysis for underdamped Langevin Monte Carlo. In *Proceedings of the 36th Conference on Learning Theory*, pp. 36–71, Proceedings of Machine Learning Research 195, PMLR, 2023

Received 27 July 2024; revised 6 March 2025.

Joseph Lehec
Université de Poitiers, CNRS, LMA, 11 boulevard Pierre et Marie Curie, 86000 Poitiers, France; joseph.lehec@univ-poitiers.fr