Understanding the Generalization Error of Markov algorithms through Poissonization

Benjamin Dupuis BENJAMIN.DUPUIS@INRIA.FR
INRIA - Département d'Informatique de l'Ecole Normale Supérieure
PSL Research University
Paris, France

Maxime Haddouche MAXIME.HADDOUCHE@INRIA.FR
INRIA - Département d'Informatique de l'Ecole Normale Supérieure
PSL Research University
Paris, France

George Deligiannidis GEORGE.DELIGIANNIDIS@STATS.OX.AC.UK
Department of Statistics, University of Oxford
Oxford, UK

Umut Simsekli UMUT.SIMSEKLI@INRIA.FR
INRIA - Département d'Informatique de l'Ecole Normale Supérieure
PSL Research University
Paris, France

Editors: Nika Haghtalab and Ankur Moitra

Abstract

Using continuous-time stochastic differential equation (SDE) proxies to stochastic optimization algorithms has proven fruitful for understanding the generalization abilities of the optimizers. These approaches are mainly based on the so-called 'entropy flows', which significantly simplify the generalization analysis. Unfortunately, such well-structured entropy flows cannot be obtained for most discrete-time algorithms, and the existing SDE approaches remain limited to specific noise and algorithmic structures. We aim to alleviate this issue by introducing a generic framework for analyzing the generalization error of Markovian algorithms through 'Poissonization', a continuous-time approximation of discrete-time processes with formal approximation guarantees. Through this approach, we first develop a novel entropy flow, which directly leads to PAC-Bayesian generalization bounds. We then draw novel links to *modified* versions of the celebrated logarithmic Sobolev inequalities (LSI), identify cases where such LSIs are satisfied, and obtain improved bounds. Beyond its generality, our framework allows exploiting specific properties of learning algorithms. In particular, we incorporate the noise structure of different algorithm types—namely, those with additional noise injections (noisy) and those without (non-noisy)—through various technical tools. This illustrates the capacity of our methods to achieve known (yet, Poissonized) and new generalization bounds.

Keywords: Generalization Bounds, Markov Algorithms, Poissonization, Log-Sobolev Inequalities.

1. Introduction

Understanding the generalization ability of machine learning algorithms remains a crucial challenge. We model such learning problems by a tuple $(\ell, \mathcal{Z}, \mu_z, \mathcal{H})$ where $\mathcal{H}$ is a parameter space ($\mathcal{H} = \mathbb{R}^d$ in our study), $\mathcal{Z}$ is a data space, μ_z a data distribution and $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$ is a loss function. The aim is to minimize the *population risk* $\mathcal{R}(w) := \mathbb{E}_{z\sim\mu_z}\left[\ell(w, z)\right]$ over the parameter space

$\mathcal{H}$. Unfortunately, as μ_z is unknown, practitioners resort to the minimization of the *empirical risk* $\widehat{\mathcal{R}}_S(w) = \frac{1}{n}\sum_{i=1}^n \ell(w, z_i)$, where $S := (z_1, \dots, z_n) \sim \mu_z^{\otimes n}$ is a dataset sampled from μ_z.

Modern machine learning systems achieve this minimization through the use of *stochastic optimization algorithms*. In our study, we focus on iterative algorithms with a Markov chain structure: $X_{k+1}^S = F(X_k^S, U_k, S)$ where $S \in \mathcal{Z}^n$ and U_k denotes the external randomness of the algorithm, independent of S. This encompasses many popular algorithms, including stochastic gradient descent (SGD) with constant step size (Dieuleveut et al., 2018) and stochastic gradient Langevin dynamics (SGLD) (see (Camuto et al., 2021; Hodgkinson et al., 2022) for a more detailed list). To assess the learning quality beyond S, it is classical to provide *generalization bounds* on the learned parameter X_k^S, *i.e.* upper bounds on the quantity $G_S(X_k^S)$ where $G_S(w) := \mathcal{R}(w) - \widehat{\mathcal{R}}_S(w)$. To provide computable guarantees, a popular approach is to derive high probability bounds of the form[1]:

$$\mathbb{P}_{S\sim\mu_z^{\otimes n}}\left(\mathbb{E}\left[G_S(X_k^S)|S\right] \lesssim \sqrt{\frac{\texttt{Complexity} + \log(1/\zeta)}{n}}\right) \geq 1-\zeta, \tag{1}$$

where the term `Complexity` translates a certain facet of the learning problem, for instance, the Rademacher complexity (Bartlett and Mendelson, 2002) or the VC dimension (Vapnik, 2000).

Generalization bounds for iterative algorithms. The classical algorithm- and data-independent approaches cannot fully exploit the problem structure. Thus, other techniques have been proposed, such as algorithmic stability (Bousquet, 2002), yielding generalization bounds for SGD under certain assumptions on the loss (convexity, Lipschitz, and/or Lipschitz gradient) (Hardt et al., 2016; Feldman and Vondrak, 2019). Unfortunately, these bounds might not be time-uniform in non-convex settings with constant step size (Bassily et al., 2020). Recently, Zhu et al. (2023) proved Wasserstein stability bounds by relying on the Markov properties of SGD. This iterative structure was also exploited by Camuto et al. (2021); Hodgkinson et al. (2022); Andreeva et al. (2024) through geometric properties, at the cost of having non-explicit mutual information terms in the bounds. Another prospect is that of information-theoretic bounds, which provided *expected* bounds for noisy algorithms (e.g., SGLD) (Xu and Raginsky, 2017; Negrea et al., 2019; Haghifam et al., 2020) and have been extended to SGD by Neu et al. (2021) at the expense of potential time and dimension-dependence.

Of particular interest to us are PAC-Bayesian bounds (McAllester, 1999; Catoni, 2007) where the term `Complexity` in Equation (1) is typically expressed as $\mathrm{KL}\left(\mathrm{Law}(X_k^S)||\pi\right)$, where π is a data-independent 'prior' distribution, $\mathrm{Law}(X_k^S)$ is called the 'posterior' distribution and $\mathrm{KL}\left(\cdot||\cdot\right)$ is the Kullback-Leibler divergence (KL). In particular, Clerico et al. (2023) proved PAC-Bayesian bounds which can handle even deterministic algorithms as well as SGD under gradient-Lipschitz losses and small learning rate, but with a potentially diverging dependence on the number of iterations.

Continuous-time analysis. Another popular route is to analyze the generalization error of 'continuous-time algorithms' $(Y_t^S)_{t\geq 0}$, typically represented by stochastic differential equations (SDE), for their structure is often easier to understand (e.g., existence of a Fokker-Planck equation) and they may provide insights for discrete-time algorithms. Indeed, for specific noise distributions, continuous-time analysis can be used to derive bounds on the discrete-time counterparts (Mou et al., 2017; Dupuis and Simsekli, 2024) and continuous-time algorithms are often viewed as approximations of discrete ones. A fundamental example is SGD, approximated by Langevin processes (Li et al., 2018; Cheng

1. We use $\lesssim$ for informal statements omitting absolute constants or weakly relevant terms.

et al., 2020; Li et al., 2021; Arous et al., 2023; Mandt et al., 2016; Anastasiou et al., 2019; Xie et al., 2021) and heavy-tailed SDEs (Simsekli et al., 2019; Gürbüzbalaban et al., 2021; Raj et al., 2023a,b).

Among these continuous-time algorithms, the generalization error of the continuous Langevin dynamics (CLD) (and its discrete-time counterpart SGLD) has been widely studied through a variety of approaches (Raginsky et al., 2017; Farghly and Rebeschini, 2021; Dupuis et al., 2024). Several methods rely on PAC-Bayesian theory and information-theoretic tools (Mou et al., 2017; Li et al., 2020; Futami and Fujisawa, 2023) where the goal is typically to upper bound the KL divergence $\mathrm{KL}\left(\mathrm{Law}(Y_T^S)||\pi_T\right)$, where π_T is a potentially time-dependent prior distribution. These techniques often rely on the so-called (relative) *entropy flow*, which has proven very useful in numerous settings:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}\left(\mathrm{Law}(Y_t^S)||\pi_t\right) = (F_1 - J) \leq \left(F_2 - \gamma \mathrm{KL}\left(\mathrm{Law}(Y_t^S)||\pi_t\right)\right), \tag{2}$$

where F_1 and F_2 are quantities usually dependent on the stochastic gradient norms, J is a Fisher information term (Chafai and Lehec, 2017) and the inequality above is a consequence of the celebrated logarithmic Sobolev inequality (LSI) (Gross, 1975; Bakry et al., 2014). When combined with classical information-theoretic bounds (Germain et al., 2009; Pensia et al., 2018), Equation (2) leads to time-uniform generalization bounds. This entropy flow technique was recently extended to α-stable noise by Dupuis and Simsekli (2024) and has been adapted to analyze the differential privacy of noisy algorithms (Chourasia et al., 2022). One of its advantages is to open the door to time-uniform bounds through a more flexible set of assumptions, compared to other approaches mentioned above.

Despite its success, this approach remains essentially limited to Gaussian (or α-stable) noises. In particular, the interpolation techniques used by Mou et al. (2017); Dupuis and Simsekli (2024) to deduce discrete-time bounds from continuous-time analysis rely on this noise structure. Moreover, the approximation of discrete-time optimizers by continuous-time dynamics remains largely disputed (Li et al., 2021; Wojtowytsch, 2021) and restricted to rather unrealistic settings, like small learning rates (Li et al., 2018) or high-dimensional limits (Ben Arous et al., 2022).

Extending the scope of continuous-time analysis. In this work, we aim to alleviate these issues and extend the entropy flow technique by utilizing a new class of continuous approximations of discrete-time Markov algorithms[2] with formal approximation guarantees, which we now describe. For a given Markov algorithm $X_{k+1}^S = F(X_k^S, U_k, S)$, we define the *Poissonization* of $(X_k^S)_{k\in\mathbb{N}}$ as the continuous-time process $Y_t^S := X_{N_t}^S$, where N_t is a Poisson process (Lasota and Mackey, 1994) (see Definition 1) . This technique has been classically used in the analysis of the convergence of Markov chains (Diaconis and Saloff-Coste, 1996; Chen et al., 2008; Caputo et al., 2024; Del Moral et al., 2003; Wang and Wu, 2020), notably by relying on modified versions of LSIs. Moreover, contrary to all continuous-time approximations of discrete-time algorithms, Poissonization is not problem-specific, meaning that it can be applied similarly to all Markov algorithms. Recently, Poissonization[3] emerged in optimization theory in Even et al. (2021) to study Nesterov acceleration.

A major innovation in our work is to connect Poissonization with the theory of generalization and thus, unveiling new links between generalization and convergence of Markov algorithms. Through an elegant formulation of the entropy flow, Poissonization acts as a tractable method to leverage the continuous-time machinery to analyze discrete-time algorithms.

2. Such a focus on Markov algorithms is not new in the generalization field (Camuto et al., 2021; Hodgkinson et al., 2022; Chandramoorthy et al., 2022; Zhu et al., 2023).
3. It is called *continuization* by Even et al. (2021) and *continuous-time semigroup* by Diaconis and Saloff-Coste (1996).

Contributions. We summarize our contributions as follows:

1. We propose in Section 3 a framework to analyze the generalization error of Poissonized Markov algorithms by *(i)* deriving a closed-form expression of the associated entropy flow and *(ii)* showing that the Poissonized generalization error is a sound approximation in certain cases.
2. We show that if the algorithm $(X_k^S)_{k\in\mathbb{N}}$ is convergent as $k \to \infty$, the Poissonized generalization error $G_S(Y_k^S)$ (at time $t = k$) will coincide with $G_S(X_k^S)$ at a rate matching the convergence of $(X_k^S)_{k\in\mathbb{N}}$. In addition to the existing literature on "depoissonization", which suggests that Poissonized generalization bounds can be informative in numerous cases, our result provides an alternative sufficient condition, which further highlights this fact.
3. Our entropy flow formula is formally similar to the previously studied Equation (2) and can be written as $\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}\left(\mathrm{Law}(Y_t^S)||\pi_t\right) = \Delta(t) - D_\Phi(t)$, where the first term $\Delta(t)$ is a new notion of 'local distance' between a prior and a posterior dynamics. We unveil the structure of the entropy flow in Section 4 by showing that the second term $D_\Phi(t)$ is connected to a class of modified LSIs that have been initially introduced for the convergence analysis of discrete Markov chains. We further prove that these modified LSIs can lead to time-uniform bounds and are satisfied by a certain class of probability distributions.
4. In Section 5, we apply our methods to two types of Markov algorithms, given that the noise distribution is continuous (e.g., SGLD) or singular (e.g., SGD). This allows us to recover known results and propose new generalization bounds under specific assumptions.

2. Technical Background

Markov kernels. Let $\mathcal{B}(\mathbb{R}^d)$ and $\mathcal{P}(\mathbb{R}^d)$ denote the Borel sets and the Borel probability measures on $\mathbb{R}^d$. Given a time-homogeneous Markov process $(X_k)_{k\in\mathbb{N}}$ in $\mathbb{R}^d$, the *Markov kernel* $P(x, A)$ describes the probability of observing X_{k+1} in a Borel set A, given that $X_k = x$. More precisely, it is a map $P : \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \to \mathbb{R}_+$ such that $\forall x \in \mathbb{R}^d,\ P(x, \cdot) \in \mathcal{P}(\mathbb{R}^d)$ and for all $A \in \mathcal{B}(\mathbb{R}^d)$, the map $x \mapsto P(x, A)$ is measurable. Classically P induces maps $P : \mathcal{P}(\mathbb{R}^d) \to \mathcal{P}(\mathbb{R}^d)$ and $P : L^\infty(\mathbb{R}^d) \to L^\infty(\mathbb{R}^d)$ defined for $\mu \in \mathcal{P}(\mathbb{R}^d),\ A \in \mathcal{B}(\mathbb{R}^d)$ and $f \in L^\infty(\mathbb{R}^d)$ by:

$$\mu P(A) := \mathbb{E}_{x\sim\mu}\left[P(x, A)\right], \quad Pf(x) := \mathbb{E}_{y\sim P(x,\cdot)}\left[f(y)\right].$$

The operator $P : L^\infty(\mathbb{R}^d) \to L^\infty(\mathbb{R}^d)$ may be extended outside of $L^\infty(\mathbb{R}^d)$, when it is well-defined.

A probability measure π is said to be *invariant* under P (or stationary) if $\pi P = \pi$ and it is said to be *reversible* if for all $f, g \in L^\infty(\mathbb{R}^d)$, one has $\mathbb{E}_\pi\left[fPg\right] = \mathbb{E}_\pi\left[gPf\right]$.

Let $\mu \in \mathcal{B}(\mathbb{R}^d)$ having density u with respect to the Lebesgue measure $\mathrm{Leb}(\mathbb{R}^d)$. If also $P\mu \ll \mathrm{Leb}(\mathbb{R}^d)$, we will denote its density by $P^\star u$, so that for $f \in L^1(\mu P)$, we have:

$$\int Pf(x)u(x)\mathrm{d}x = \int f(x)P^\star u(x)dx. \tag{3}$$

Poissonization. In all the following, we fix a Poisson process $(N_t)_{t\geq 1}$ with intensity 1, as defined below (Lasota and Mackey, 1994). It is assumed to be independent of all the other random variables.

Definition 1 (Poisson process) *A Poisson process $(N_t)_{t\geq 0}$ with intensity $\lambda > 0$ is a Lévy process with values in $\mathbb{N}$, almost surely increasing, such that $N_0 = 0$ and $\forall k \in \mathbb{N},\ \mathbb{P}\left(N_t = k\right) = e^{-\lambda t}\frac{(\lambda t)^k}{k!}$.*

We refer to (Schilling, 2016) for the definition of a Lévy process. Given a time-homogeneous Markov process $(X_k)_{k\in\mathbb{N}}$, we define the *Poissonized process* as $Y_t := X_{N_t}$ (Teugels, 1972; Lasota and Mackey, 1994). Note that equivalent definitions of Poissonization exist. For instance, Even et al. (2021) used a stochastic integral formulation in their study of Nesterov acceleration and other authors favored an approach based on semigroups (Diaconis and Saloff-Coste, 1996). One of the main features of Poissonized processes is that their probability density function (PDF) $u_t(\cdot)$ satisfies a simple differential equation, sometimes referred to as the "Boltzmann equation" Lasota and Mackey (1994, Equation 8.3.7). This equation can be written as (a rigorous proof is provided in Section B):

$$\frac{\partial u_t}{\partial t} = (P^\star - I)u_t. \tag{4}$$

Depoissonization. A discrete-time Markov chain and its Poissonized version are known to have comparable properties (Jacquet and Szpankowski, 1998; Levin and Peres, 2017; Caputo et al., 2024). That being said, reconstructing the depoissonized distribution of X_k from Y_t is a technical and long-standing problem (Teugels, 1972; Vallée, 2018; Jacquet and Szpankowski, 1998) for which we give a quick introduction in Section F. In Theorem 3 below, we further show that Poissonization provides a sound approximation of the generalization error of convergent Markov algorithms.

PAC-Bayesian bounds. Analyzing the generalization error of stochastic optimization algorithms leads to the consideration of randomized predictors, which have been classically studied by PAC-Bayesian theory (see (Alquier, 2024) for an introduction). More precisely, given $S \in \mathcal{Z}^n$, we define the posterior $\rho_S \in \mathcal{P}(\mathbb{R}^d)$ to be the distribution[4] of the random output of the algorithm (e.g., SGD). Given a prior (*i.e.*, data-independent) distribution $\pi \in \mathcal{P}(\mathbb{R}^d)$, a wide variety of PAC-Bayesian bounds have related the generalization error to the KL divergence $\mathrm{KL}\left(\rho_S||\pi\right)$ (McAllester, 1999, 2003; Maurer, 2004; Catoni, 2007; Germain et al., 2009; Seeger, 2002) (to name a few).

In our study, we assume that ℓ is s^2-subgaussian, *i.e.*, $\forall\lambda,\ \mathbb{E}\left[e^{\lambda(\ell(w,z)-\mathbb{E}_{z'}[\ell(w,z')])}\right] \leq e^{\frac{\lambda^2 s^2}{2}}$. To utilize this assumption, we use the PAC-bayesian bound proposed by Dupuis and Simsekli (2024), which is similar to that of McAllester (2003); Germain et al. (2009) in the case of bounded losses.

Theorem 2 *Assume that for all $S \in \mathcal{Z}^n$, we have $\rho_S \ll \pi$ and that ℓ is s^2-subgaussian. Then:*

$$\mathbb{P}_{S\sim\mu_z^{\otimes n}}\left(\mathbb{E}_{\rho_S}\left[G_S(w)\right] \leq 2s\sqrt{\frac{\mathrm{KL}\left(\rho_S||\pi\right)+\log(3/\zeta)}{n}}\right) \geq 1-\zeta.$$

3. Poissonized Markov Algorithms

In this section, we express the entropy flow between Poissonized processes, which underlies all our main results. We introduce our framework in Section 3.1 and derive the entropy flow in Section 3.2.

3.1. A framework for the Poissonization of generalization bounds

In our paper, we apply PAC-Bayesian bounds on distributions that are induced by posterior dynamics (*i.e.*, the learning algorithm) and prior dynamics (*i.e.*, data-independent). The precise definitions are given below and we provide a summary of these notations in Table 1.

4. More precisely, $(\rho_S)_{S\in\mathcal{Z}^n}$ is a Markov kernel on $\mathcal{Z}^n \times \mathbb{R}^d$.

- **Posterior dynamics.** It is a *data-dependent* time-homogeneus Markov process $(X_k^S)_{k\geq 0}$ with kernel denoted P_S for a dataset $S \in \mathcal{Z}^n$. The Poissonization (see Section 2) of $(X_k^S)_{k\geq 0}$ is denoted $(Y_t^S)_{t\geq 0}$. We will write ρ_t^S for the probability distribution of Y_t^S and μ_k^S for the probability distribution of X_k^S. The PDF of Y_t^S is denoted $u_t^S(\cdot)$ (when it is defined). We assume that $(X_k^S)_{k\geq 0}$ is initialized from a smooth probability density p_0 and denote $X_0 \sim p_0$.
- **Prior dynamics.** It is a *data-independent* Markov process $(X_k)_{k\geq 0}$ with kernel P. We denote by μ_k the probability distribution of X_k. The prior Poissonized process is denoted $(Y_t)_{t\geq 0}$, its probability distribution is denoted π_t, and its PDF is denoted u_t.

We use the *Poissonized* distributions ρ_t^S and π_t as the posterior and prior distributions in Theorem 2 to provide generalization bounds for the Poissonized algorithm, *i.e.*, $\mathbb{E}_{w\sim\rho_t^S}[G_S(w)]$. Whether this provides pertinent information about the non-Poissonized iterates is a legitimate concern. Beyond the classical depoissonization results mentioned in Sections 2 and F, we show in Theorem 3 that $\mathbb{E}_{w\sim\rho_k^S}[G_S(w)]$ (with $t=k$) approximates $\mathbb{E}_{w\sim\mu_k^S}[G_S(w)]$ for certain Markov processes. Below TV denotes the total variation distance and the proof can be found in Section C.

Theorem 3 *Assume that $|\ell| \leq B < \infty$ and that $\mathrm{TV}(\mu_k^S, \mu^S) \longrightarrow 0$ for some $\mu^S \in \mathcal{P}(\mathbb{R}^d)$, a.s. for S. Then, a.s., $\mathbb{E}\left[|G_S(X_k^S) - G_S(Y_k^S)||S\right] \longrightarrow 0$. If moreover there exists $C > 0$ and $a_S \in (0,1)$ such that, a.s., $\mathrm{TV}(\mu_k^S, \mu^S) \leq C_S a_S^k$, then, a.s., $\mathbb{E}\left[|G_S(X_k^S) - G_S(Y_k^S)||S\right] \leq 4BC_S e^{-(1-a_S)k}$. If ℓ is L-Lipschitz, then we can replace TV by the 1-Wasserstein distance W_1 (and $2B$ by L) in these statements, e.g., if $W_1(\mu_k^S, \mu^S) \leq C_S a_S^k$, then $\mathbb{E}\left[|G_S(X_k^S) - G_S(Y_k^S)||S\right] \leq 2LC_S e^{-(1-a_S)k}$.*

The conditions $\mathrm{TV}(\mu_k^S, \mu^S) \leq C_S a_S^k$ (resp. $W_1(\mu_k^S, \mu^S) \leq C_S a_S^k$) used above are related to *geometric* (resp. *Wasserstein*) ergodicity (Meyn and Tweedie, 1993; Gallegos-Herrada et al., 2023) that has been widely studied in the context of convergence of Markov chains (Rudolf and Schweizer, 2017). Note that, our condition is weaker than geometric ergodicity: we do not assume the uniqueness of the invariant distribution. These concepts have received growing attention in learning theory for their connections with SGD (Zhu et al., 2023) and differential privacy (Şimşekli et al., 2024). While our study is not specific to ergodic Markov chains, Theorem 3 provides a *sufficient* condition for ensuring that Poissonization is a relevant continuous-time approximation of discrete dynamics.

3.2. Poissonized entropy flow

When it is defined, we denote v_t for the Radon-Nykodym derivative between ρ_t^S and π_t:

$$v_t := \frac{u_t^S}{u_t} = \frac{\mathrm{d}\rho_t^S}{\mathrm{d}\pi_t} \quad \text{(we omit the dependence on } S \in \mathcal{Z}^n \text{ in } v_t\text{)}.$$

Our theory relies on regularity conditions on v_t, which we now explain. In all the following, we consider the convex function $\Phi(x) = x\log(x)$ (and $\Phi(0) = 0$). We also fix a time horizon $T > 0$.

Assumption 1 (Regularity) *Let $t \in [0,T]$ and $k \in \mathbb{N}$, we have $\ell \in L^1(\rho_t^S \otimes \mu_z)$. Moreover, $\mu_k, \mu_k^S \ll \mathrm{Leb}(\mathbb{R}^d)$, we have $\rho_t^S \ll \pi_t$, the function $v_t = \mathrm{d}\rho_t^S/\mathrm{d}\pi_t$ is positive, $v_t \in \mathcal{C}^2(\mathbb{R}^d)$, and:*

- ***H.1*** *(Domination) In every compact time interval I, there exists a positive function $\psi_I \in L^1(\mathrm{d}x)$ such that $\forall t \in I$, $\left|\frac{\mathrm{d}}{\mathrm{d}t}(u_t\Phi(v_t))\right| \leq \psi_I$.*
- ***H.2*** *(Logarithmic regularity) For $x \in \mathbb{R}^d$, $t \in [0,T]$, we have $v_t \in L^1(\delta_x P)$, $P(\Phi(v_t)) \in L^1(\pi_t)$, and $\log(v_t) \in L^1(\rho_t^S) \cap L^1(\rho_t^S P) \cap L^1(\rho_t^S P_S)$.*

The assumptions $\rho_t^S \ll \pi_t$ and $\ell \in L^1(\rho_t^S \otimes \mu_z)$ are natural in our PAC-Bayesian approach. We suppose that μ_k and μ_k^S are absolutely continuous mainly for simplicity as this can be relaxed, as briefly discussed in Remark 19. Requiring that v_t is positive and twice continuously differentiable is a relatively mild assumption. Indeed, if the initialization $X_0 \sim p_0$ is everywhere positive, then this property is preserved for the Poissonized distributions u_t^S. The fact that $v_t \in \mathcal{C}^2(\mathbb{R}^d)$ implies that the algorithm is not creating singularities during training. It can be satisfied even by non-noisy algorithms such as for SGD with a gradient-Lipschitz loss and small learning rate (Clerico et al., 2023).

	Posterior	Prior
Markov kernels	P_S	P
Discrete Markov process	$(X_k^S)_{k\in\mathbb{N}}$	$(X_k)_{k\in\mathbb{N}}$
Poissonized Markov process	$(Y_t^S)_{t>0}$	$(Y_t)_{t>0}$
Dist. of the discrete process	$\mu_k^S,\, k\in\mathbb{N}$	$\mu_k,\, k\in\mathbb{N}$
Dist. of the Poissonized process	$\rho_t^S,\, t>0$	$\pi_t,\, t>0$
Density of the Poissonized process	u_t^S	u_t

Table 1: Notations for posterior and prior dynamics.

Conditions **H.1** and **H.2** regard the minimal integrability properties of v_t to ensure the existence of various terms defined below. In particular, Condition **H.1** allows us to differentiate the relative entropy $\mathrm{KL}\left(\rho_t^S||\pi_t\right)$, which is the purpose of our framework. These conditions can be expected to be mild in practice. For instance, $\log(v_t) \in L^1(\rho_t^S)$ is equivalent to $\mathrm{KL}\left(\rho_t^S||\pi_t\right) < +\infty$ and the other integrability conditions on $\log v_t$ are satisfied as soon as $\exists K > 0,\ |\log(v_t)(y)| = \mathcal{O}\left(\|y\|^K\right)$ and $\rho_t^S P$ and $\rho_t^S P_S$ have finite moments of order K (it typically holds for Gaussian distributions). Finally, in the case where $\pi_t = \pi$ is an invariant measure for P, the conditions $v_t \in L^1(\delta_x P)$ and $P(\Phi(v_t)) \in L^1(v_t)$ are implied by the other conditions (Rudolf, 2012, Lemma 3.6). Assumption 1 is similar to (Dupuis and Simsekli, 2024, Assumption 3.3) made in the case of heavy-tailed SDEs.

Based on Assumption 1, we can now state the main result of this section, which is a closed-form expression for the entropy flow between the Poissonized processes Y_t^S and Y_t.

Theorem 4 (Poissonized entropy flow) *Under Assumption 1, the entropy flow is given by:*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}\left(\rho_t^S||\pi_t\right) = \Delta_{P,P_S}(v_t) - \mathbb{E}_{x\sim\pi_t, y\sim\delta_x P}\left[D_\Phi\left(v_t(x), v_t(y)\right)\right], \tag{5}$$

*where $D_\Phi(a,b) := \Phi(a) - \Phi(b) - \Phi'(b)(a-b)$ is the Bregman divergence. We call the first term $\Delta_{P,P_S}(v_t) := \mathbb{E}_{\rho_t^S}\left[(P_S - P)\log(v_t)\right]$ the **expansion term** and the second the **Bregman term**.*

Proof (Sketch, see Section C) By Item **H.1**, we have $\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}\left(\rho_t^S||\pi_t\right) = \int \frac{\partial}{\partial t}(\Phi(v_t)u_t)\mathrm{d}x$. The crucial step is to use the Boltzmann Equation (4), along with Equation (3) and Item **H.2** to make the Markov operators P and P_S appear in the expression. The result follows by rearranging the terms. Let us note that this proof technique is not specific to $\Phi(x) = x\log(x)$ and can be seamlessly extended to the so-called Φ-entropies (Bakry et al., 2014, Section 7.6.1). ■

Theorem 4 expresses the entropy flow as the difference between the *expansion* term and the *Bregman* term. The expansion term represents a discrepancy between the posterior dynamics P_S and the prior one P and Section 5 is dedicated to its analysis. By convexity of Φ, the Bregman term has a non-positive contribution to the entropy flow, analogously to the role of the Fisher information appearing in (Mou et al., 2017, Proposition 2) and the "Bregman integral" considered by Dupuis and Simsekli (2024) in their study of heavy-tailed SDEs. The Bregman term crucially connects our framework to *modified* logarithmic Sobolev inequalities, as we explain in Section 4.

4. Towards Time-uniform Generalization Bounds

In this section, we study the Bregman term appearing in Theorem 4. We first show in Section 4.1 that it can be related to well-studied *modified* logarithmic Sobolev inequalities (LSI) (Diaconis and Saloff-Coste, 1996) and that such modified LSIs can notably improve our generalization bounds. In Section 4.2, we show that such modified LSIs are satisfied by a certain class of prior dynamics.

4.1. From Bregman integral to modified LSIs

We recall the notion of (classical) LSI (see Bakry et al., 2014; Chafai and Lehec, 2017 for modern introductions). Let ν be a Borel probability measure, we associate to ν an *entropy* functional, defined as $\mathrm{Ent}_\nu(f) = \mathrm{Ent}_\nu^\Phi(f) := \mathbb{E}_\nu[\Phi(f)] - \Phi(\mathbb{E}_\nu[f])$, with $\Phi(x) = x\log(x)$. The entropy generalizes the KL divergence, in the sense that $\mathrm{Ent}_\nu(\mathrm{d}\mu/\mathrm{d}\nu) = \mathrm{KL}(\mu||\nu)$ as soon as $\mu \ll \nu$.

A probability measure π is said to satisfy the β-LSI if for all positive $f \in L^1(\pi) \cap \mathcal{C}^1(\mathbb{R}^d)$ we have the inequality $\mathrm{Ent}_\pi(f) \leq \frac{2}{\beta}\mathbb{E}_\pi\left[\left\|\nabla\sqrt{f}\right\|^2\right]$.

For instance, the Gaussian distribution $\mathcal{N}(0, \sigma^2 I_d)$ satisfies a $1/\sigma^2$-LSI (Gross, 1975). Such inequalities have been extensively studied for their links with the convergence of Markov processes (Bakry et al., 2014) and geometry (Bobkov, 1996; Otto and Villani, 2000). In learning theory, they have been used for the generalization analysis of noisy algorithms (Mou et al., 2017; Li et al., 2020), differential privacy (Chourasia et al., 2022; Ryffel et al., 2022) and PAC-Bayesian bounds (Haddouche et al., 2024; Casado et al., 2024). In the study of (Poissonized) discrete Markov processes, it has been shown that the β-LSI should be replaced by a functional inequality that takes into account the Markov kernel. This leads to the following notion of *modified* LSI.

Definition 5 (Modified LSI) *An invariant*[5] *measure π of the Markov kernel P satisfies a modified γ-LSI if for any positive f s.t. $\forall x,\ f, \log f \in L^1(\delta_x P)$ and $f\log f,\ fP\log f \in L^1(\pi)$, we have:*

$$\mathcal{E}_\pi(\log(f), f) \geq \gamma \mathrm{Ent}_\pi(f), \tag{6}$$

*where $\mathcal{E}_\pi$ is called the **Dirichlet form**[6] and is defined as $\mathcal{E}_\pi(f, g) := \mathbb{E}_\pi[g(I - P)f]$.*

Such inequalities were introduced by Diaconis and Saloff-Coste (1996) and extensively studied by Bobkov and Tetali (2006); Bobkov and Ledoux (1998); Goel (2004); Wu (2000); Ané and Ledoux (2000) to analyze the convergence rate of Markov chains. Bobkov and Tetali (2006) used the term "modified" to avoid confusion with other inequalities, and we adopted this terminology in our study. In order to involve modified LSIs, we remark that the Bregman term of Theorem 4 can be expressed as a Dirichlet form if the prior is an invariant measure of P, as proven in the following corollary.

Corollary 6 *Assume that Assumption 1 holds and that P has an invariant measure π. We have*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}\left(\rho_t^S||\pi\right) = \mathbb{E}_{\rho_t^S}\left[(P_S - P)(\log v_t)\right] - \mathcal{E}_\pi(\log v_t, v_t).$$

Then, in the next theorem, we exploit modified LSIs to reach a novel generalization bound.

5. Bobkov and Tetali (2006) defined this inequality for reversible measures, we seamlessly extend it to the invariant case.
6. This is a Dirichlet form when π is reversible for P, which makes $\mathcal{E}_\pi$ symmetric (Diaconis and Saloff-Coste, 1996).

Theorem 7 (Generalization error of Poissonized algorithms) *Assume that ℓ is s^2-subgaussian, Assumption 1 holds, and the prior dynamics has an invariant measure π which satisfies a modified LSI with constant γ, in the sense of Definition 5. Then, with probability at least $1-\zeta$ under $S \sim \mu_z^{\otimes n}$:*

$$\mathbb{E}_{w\sim\rho_T^S}\left[G_S(w)\right] \leq \frac{2s}{\sqrt{n}}\left\{\int_0^T e^{-\gamma(T-t)}\Delta_{P,P_S}(v_t)\mathrm{d}t + e^{-\gamma T}\mathrm{KL}\left(p_0||\pi\right) + \log\left(\frac{3}{\zeta}\right)\right\}^{\frac{1}{2}}.$$

Proof (Sketch, see Section D) We start by using the entropy flow formula of Corollary 6 to obtain the inequality: $\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}\left(\rho_t^S||\pi\right) = \Delta_{P,P_S}(v_t) - \mathcal{E}_\pi(\log v_t, v_t) \leq \Delta_{P,P_S}(v_t) - \gamma\mathrm{KL}\left(\rho_t^S||\pi\right)$.

The result follows by solving this differential inequality and plugging the result in Theorem 2. ■

Theorem 7 shows that priors satisfying modified LSIs induce an exponential decay $e^{-\gamma(T-t)}$ in our generalization bound. This is analogous to the role of LSIs for Gaussian (Mou et al., 2017) or heavy-tailed (Dupuis and Simsekli, 2024) SDEs. Thus, determining which measure satisfies a modified LSI is a key question tackled in Section 4.2. Theorem 7 reduces the problem of controlling time-uniformly $\mathbb{E}_{w\sim\rho_T^S}\left[G_S(w)\right]$ for upper-bounding $\Delta_{P,P_S}(v_t)$. Such a conclusion is analogous to Theorem 2, reduces the generalization problem to the control of $\mathrm{KL}\left(\rho_t^S||\pi_t\right)$. This is why in Section 5, we often directly present upper bounds on either $\Delta_{P,P_S}(v_t)$ or $\mathrm{KL}\left(\rho_t^S||\pi_t\right)$.

4.2. Modified logarithmic Sobolev inequalities for diffusive priors

To exploit Theorem 7, it is essential to identify which measures satisfy a modified LSI. Several works proposed modified LSIs for specific Markov chains (Diaconis and Saloff-Coste, 1996; Goel, 2004; Ané and Ledoux, 2000; Bobkov and Tetali, 2006). In particular, Erbar and Maas (2012) obtained modified LSIs under a Ricci curvature condition for discrete Markov chains. However, most of these results are constrained to finite or countable state spaces, which is inconsistent with the continuous distributions involved in this work. Inspired by the theory of classical LSIs, we consider the class of prior Markov kernels P that can be represented by diffusions in the following sense.

Definition 8 *Consider a twice differentiable gradient-Lipschitz potential $V : \mathbb{R}^d \to \mathbb{R}$ such that $e^{-V} \in L^1(\mathbb{R}^d)$, e^{-V} has finite moments of order 2, and the Langevin equation $\mathrm{d}Z_t = -\nabla V(Z_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$, where $(B_t)_{t\geq 0}$ is a standard Wiener process. We say that a Markov kernel P is representable by this equation at time $t_0 > 0$ if for all $x \in \mathbb{R}^d$, we have $P(x,.) = \mathrm{Law}(Z_{t_0}|Z_0 = x)$.*

The main outcome of this section is that prior dynamics satisfying Definition 8 satisfy a modified LSI if the invariant measure of the underlying Langevin equation satisfies a classical LSI.

Theorem 9 (Modified LSI for diffusive priors) *Assume the Markov kernel P to be representable at time t_0 by a diffusion with an ergodic invariant measure π, as in Definition 8. Let $K > 0$ and $f \in \mathcal{C}^2(\mathbb{R}^d)$ a positive function s.t. $\forall x,\ f, \log(f) \in L^1(\delta_x P)$, and $f\log(f),\ Pf\log(f) \in L^1(\pi)$. If π satisfies a LSI with constant K, then we have the modified LSI:*

$$\mathcal{E}_\pi(\log f, f) \geq c_{\mathrm{LSI}}\mathrm{Ent}_\pi\left(f\right), \tag{7}$$

with $c_{\mathrm{LSI}} = \frac{Kt_0}{1+Kt_0}$. If we have $\nabla^2 V \succeq KI_d$, then the constant is improved to $c_{\mathrm{LSI}} = 1 - e^{-Kt_0}$.

We give two examples of diffusive priors of particular interest to this work. First, the following corollary corresponds to the case of Gaussian priors. Similar to (Mou et al., 2017), we use this Gaussian prior in the case of an algorithm featuring ℓ^2-regularization, see Section 5.

Corollary 10 *Consider the Markov process defined by $X_{k+1} = (1-\gamma)X_k + \sigma\mathcal{N}(0, I_d)$ with $\gamma \in (0,1), \sigma > 0$. Then the associated Dirichlet form $\mathcal{E}_\pi$ satisfies a modified LSI with constant γ.*

Proof (Sketch, see Section D.2) We show that the recursion $X_{k+1} = (1-\gamma)X_k + \sigma\mathcal{N}(0, I_d)$ is representable by the SDE $\mathrm{d}Z_t = -cZ_t\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$ at time t_0 for certain values of t_0 and c explicitly given in Section D.2. The result follows from Theorem 9, by the strong convexity of $x \mapsto \frac{c}{2}\|x\|^2$. ■

Second, inspired by (Amir et al., 2022; Dupuis and Viallard, 2023; Dupuis et al., 2024), our framework makes it possible to use the population risk $\mathcal{R}$ as a potential function in Definition 8, leading to the use of the SDE $\mathrm{d}Z_t = -c\nabla\mathcal{R}(Z_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$. Under certain assumptions on $\mathcal{R}$ (see (Li et al., 2020)), the invariant measure satisfies an LSI, making it compatible with Theorem 9. These examples shed new light on the choice of the prior in information-theoretic generalization bounds.

5. Controlling the Discrepancy Between Markov Kernels for Concrete Algorithms

In Section 4, we have controlled the Bregman term appearing in Theorem 4 through modified LSIs and then proved that such inequalities were satisfied by diffusive priors. The last step to apply our theory to practical algorithms is to analyze the expansion term $\Delta_{P,P_S}(v_t)$. We present two sets of tools to achieve this, depending on the structure of the algorithm. First, we consider *noisy* algorithms in Section 5.1 and then turn to non-noisy algorithms in Section 5.2.

5.1. Poissonized bounds for noisy algorithms and SGLD

The generalization error of noisy iterative algorithm has been extensively studied (Haghifam et al., 2020; Xu and Raginsky, 2017; Negrea et al., 2019; Bu et al., 2020). In our work, we say that a Markov algorithm is *noisy* if for all $S \in \mathcal{Z}^n$ and all $x \in \mathbb{R}^d$, we have $P_S(x,\cdot) \ll \mathrm{Leb}(\mathbb{R}^d)$. In all this section, we consider prior dynamics such that for all $x \in \mathbb{R}^d$, $\delta_x P$ is equivalent to the Lebesgue measure $\mathrm{Leb}(\mathbb{R}^d)$ (*i.e.*, $\delta_x p \ll \mathrm{Leb}(\mathbb{R}^d)$ and $\mathrm{Leb}(\mathbb{R}^d) \ll \delta_x P$). Our main example of a noisy algorithm is noisy SGD, for which we introduce some notations in the following example.

Example 1 (SGD and noisy SGD) *We define $X_{k+1} = (1-\lambda\eta)X_k - \eta\widehat{g}_S(X_k, U_k) + \zeta_k$, with learning rate $\eta > 0$, regularization coefficient $\lambda \geq 0$ (potentially 0), stochastic gradient $\widehat{g}_S(x, U_k)$, and added noise ζ_k. The random variable U_k represents the randomness of the batch indices. SGLD corresponds to the case where $\zeta_k \sim \mathcal{N}(0, \sigma^2 I_d)$. We use these notations in several discussions below.*

Warm up. We start with a generic bound of the entropy flow for noisy algorithms, which applies to a general class of noise distributions. The following corollary is a direct consequence of Theorem 4 and an application of Donsker-Varadhan's formula, the proof can be found in Section E.

Corollary 11 *Under the above conditions and Assumption 1, a noisy algorithm satisfies:*

$$\mathrm{KL}\left(\rho_T^S||\pi_T\right) \leq \mathrm{KL}\left(p_0||\pi_0\right) + \int_0^T \mathbb{E}_{\rho_t^S}\left[\mathrm{KL}\left(\delta_x P_S||\delta_x P\right)\right]\mathrm{d}t - \int_0^T \mathbb{E}_{\pi_t}\left[D_\Phi(v_t, Pv_t)\right]\mathrm{d}t. \quad (8)$$

If P has an invariant measure π and we use $\forall t,\ \pi_t = \pi$, then we can simplify the last term as $\mathbb{E}_{\pi_t}\left[D_\Phi(v_t, Pv_t)\right] = \mathrm{KL}\left(\rho_t^S||\rho_t^S P^\dagger\right)$, where $P^\dagger$ is the adjoint of P in $L^2(\pi)$.

To analyze the above proposition, let us compare with a discrete-time (naive) bound of the form $\mathrm{KL}\left(\mu_N^S||\mu_N\right) \leq \mathrm{KL}\left(p_0||\mu_0\right) + \sum_{k=0}^{N-1} \mathbb{E}_{x\sim\mu_k^S}\left[\mathrm{KL}\left(\delta_x P_S||\delta_x P\right)\right]$, which can be obtained by using the data-processing inequality and the chain rule, see (Neu et al., 2021; Negrea et al., 2019) for similar ideas. As we consider a Poisson process $(N_t)_{t>0}$ of intensity 1, the previous sum is analogous to the first integral in Equation (8). We note that Equation (8) provides a better bound, with a negative term that can be seen as an estimate of the error made by the previous reasoning.

The "local" KL divergence $\mathrm{KL}\left(\delta_x P_S||\delta_x P\right)$ can be estimated in numerous cases. For noisy SGD with Gaussian or Laplace-distributed noise, up to a relevant choice of prior P (*i.e.*, corresponding to $X_{k+1} = (1-\eta\lambda)X_k + \zeta_k$ with the notations of Example 1), we have $\mathrm{KL}\left(\delta_x P_S||\delta_x P\right) \lesssim \mathbb{E}_U\left[\|\widehat{g}_S(x,U)\|^2\right]$. Hence, our framework provides informative bounds for various noising schemes, such as Laplace noise, which has been considered for differential privacy (Kuru et al., 2022).

The case of SGLD. The term $\mathbb{E}_{\pi_t}\left[D_\Phi(v_t, Pv_t)\right]$ appearing in Corollary 11 is different (by Jensen's inequality, it is smaller) from the Bregman term featured in Theorem 4. This suggests that modified LSIs using this term instead of $\mathcal{E}_\pi(\log f, f)$ would lead to improved bounds with a form similar to Theorem 7. In the case of SGLD, we were able to circumvent the need for such stronger inequality and applied our Poissonization framework to prove the following result (see the proof in Section E.1).

Theorem 12 (Poissonized SGLD) *Consider the Markov kernel P_S corresponding to SGLD with $\eta\lambda < 1$ and take P and π to be the Markov kernel and the invariant distribution of the recursion $X_{k+1} = (1-\lambda\eta)X_k + \sigma\mathcal{N}(0, I_d)$. Assume that Assumption 1 holds, then we have:*

$$\mathrm{KL}\left(\rho_T^S||\pi\right) \leq \frac{\eta^2(2-\lambda\eta)}{2\sigma^2}\int_0^T e^{-\lambda\eta(T-t)}\mathbb{E}_{x\sim\rho_t^S,U}\left[\|\widehat{g}_S(x,U)\|^2\right]\mathrm{d}t. \tag{9}$$

Theorem 12 is proven in Section E.1. Together with Theorem 2, this result provides a generalization bound for Poissonized SGLD. Because the underlying Poisson process is of intensity 1, we note that the order of magnitude of the different terms in Equation (9) are of the same order of magnitude as the results of Mou et al. (2017) obtained under similar assumptions. Hence, the Poissonization framework is general enough to recover Poissonized counterparts of classical results.

5.2. Analysis of non-noisy algorithms

Unfortunately, various popular procedures, starting with SGD, are not included in the noisy algorithms class. Then, can the Poissonization framework cover non-noisy procedures, *i.e.*, when the absolute continuity property $\delta_x P_S \ll \delta_x P$ does not hold? The answer is positive: We extend the entropy flow technique beyond noisy algorithms (under specific assumptions) and reach informative generalization bounds encompassing both noisy and non-noisy methods, with a focus on SGD.

First-order analysis. To avoid the condition $\delta_x P_S \ll \delta_x P$, our main idea is to perform an expansion of P_S "around P". This is made more precise by the following proposition, which is inspired by (Polyanskiy and Wu, 2016, Proposition 1) and (Raginsky et al., 2017, Lemma 3.5).

Proposition 13 *Assume that, for all $t \in [0,T]$, there exist constants $c_1, c_2 > 0$ such that v_t satisfies the linear growth condition $\forall x \in \mathbb{R}^d$, $\|\nabla \log v_t(x)\| \leq c_1\|x\| + c_2$, then we have:*

$$\Delta_{P,P_S}(v_t) \leq \mathbb{E}_{\rho_t^S}\left[W_2(\delta_x P, \delta_x P_S)^2\right]^{\frac{1}{2}}\left(\frac{c_1}{2}\|P\|_t + \frac{c_1}{2}\|P_S\|_t + c_2\right),$$

with $\|P\|_t^2 := \mathbb{E}_{x\sim\rho_t^S, w\sim\delta_x P}\left[\|w\|^2\right]$ *(resp.* P_S*) and* W_2 *the Wasserstein distance (Section E.2).*

The main feature of Proposition 13, proved in Section E.2, is to relate the expansion term to the Wasserstein distance $W_2(\delta_x P_S, \delta_x P)$. Note that this term is local, *i.e.*, it is computed for every point $x \in \mathbb{R}^d$ and expected over the posterior distribution. In the case of SGD, with a relevant choice of P, we typically have $W_2(\delta_x P, \delta_x P_S)^2 \leq \eta^2 \mathbb{E}_U\left[\|\widehat{g}_S(x,U)\|^2\right]$, with the notations of Example 1. Similar Wasserstein distances between Markov kernels have been extensively studied in the context of convergence and geometry of Markov chains (Rudolf and Schweizer, 2017; Ollivier, 2007) and have been used by (Zhu et al., 2023) to obtain stability-based bounds for SGD. Therefore, Proposition 13 connects our framework and the prior art through these Wasserstein terms.

Proposition 13 relies on a condition of linear growth of $\nabla \log v_t = \nabla \log u_t^S - \nabla \log \pi_t$. This assumption can be seen as an assumption on the tail of the posterior density u_t^S and can hint towards good choices of prior, *i.e.*, with similar tails as the posterior density. For instance, if we use a Gaussian prior as in Corollary 10, it boils down to a linear growth assumption on the score[7] of the posterior density. Finally, note that a related condition was used by Li et al. (2020, Section A.4).

Second-order analysis. Proposition 13 contains kernel norm terms $\|P_S\|_t$ and $\|P\|_t$ possibly large when $c_1 > 0$. In particular, the diffusive priors studied in Section 4.2, make the term $\|P\|_t$ of order $\sqrt{d}$, leading to a multiplicative constant of order $\sqrt{c_1} d^{1/4}/\sqrt{n}$ in the final generalization bound.

To address this potential issue, we observe that the proof of Proposition 13 can be seen as a 1^{st}-order Taylor expansion of the P_S and propose to extend it to a 2^{nd}-order expansion through stronger assumptions. For the sake of simplicity, we focus in Theorem 14 on the case of regularized SGD, *i.e.*, $X_{k+1} = (1-\lambda\eta)X_k - \eta\widehat{g}_S(X_k, U_k)$ with the notations of Example 1. Nevertheless, the methods presented here may be more generally applied to other algorithms, see Section E.2.

Theorem 14 (Generalization bound for regularized SGD) *Assume that* ℓ *is* s^2*-subgaussian and Assumption 1 holds with* π *the invariant measure of the Markov process of Corollary 10 with* $\gamma = \lambda\eta$ *and* $\sigma > 0$ *a noise scale. We further assume that there exists* $\beta \geq 0$ *such that* $0 \preceq \nabla^2 \log(v_t) \preceq \beta I_d$*, for all* $t \in [0, T]$*. Then, we have, with probability at least* $1-\zeta$ *over* $S \sim \mu_z^{\otimes n}$ *that:*

$$\mathbb{E}_{\rho_T^S}\left[G_S\right] \leq \frac{2s}{\sqrt{n}}\left\{\int_0^T e^{-\lambda\eta(T-t)}\eta\mathbb{E}_{x\sim\rho_t^S, U}\left[Q\left(\|\widehat{g}_S(x,U)\|, \|x\|\right)\right]\mathrm{d}t + e^{-\lambda\eta T}K_0 + \log\frac{3}{\zeta}\right\}^{\frac{1}{2}},$$

where $Q(X,Y) := X\left(\beta Y + \left\|\nabla \log u_t^S(0)\right\|\right) + \eta\beta\left(X^2 + \lambda^2 Y^2\right)$ *is a* 2^{nd}*-order bivariate polynomial,* $K_0 := \mathrm{KL}\left(p_0 || \pi\right)$ *and* U *is the randomness of the batches in the stochastic gradient* $\widehat{g}_S$*.*

Let us first note that the assumption $0 \preceq \nabla^2 \log v_t \preceq \beta I_d$ implies the linear growth assumption of Proposition 13 (with $c_1 = \beta$ and $c_2 = \|\nabla \log v_t(0)\|$). The lower bound of this condition ($0 \preceq \nabla^2 \log v_t$) implies that the Radon-Nykodym derivative v_t is convex and non-bounded, which is non-trivial but still a priori compatible with Assumption 1. Moreover, in Theorem 14 the prior π is chosen to be a Gaussian distribution $\mathcal{N}(0, \sigma_\pi^2 I_d)$, thus, $\nabla^2 \log v_t = \nabla^2 \log u_t^S + 1/\sigma_\pi^2$. Therefore, in the case where a condition of the form $-bI_d \preceq \nabla^2 \log(u_t^S) \preceq bI_d$ holds, Theorem 14 suggests to choose π so that $\nabla^2 \log v_t \succeq 0$. Once such a prior is chosen, Assumption 1 remains the sole regularity condition conducing to Theorem 14. As can be seen from the proof in Section E.2, this positive

7. Given a probability density function p, the score function is defined as $x \mapsto \nabla \log p(x)$.

semi-definite condition can be relaxed to $-\beta I_d \preceq \nabla^2 \log(u_t^S) \preceq \beta I_d$, at the cost of introducing dimension-dependent terms in the bound.

Finally, we see that Theorem 14 relates the generalization error of SGD to the stochastic gradient norms, averaged over the posterior distribution. Similar quantities classically appear in the study of noisy SGD (Mou et al., 2017; Negrea et al., 2019; Haghifam et al., 2020; Dupuis and Simsekli, 2024) and were already involved for non-noisy SGD in the bounds of Neu et al. (2021). Compared to (Neu et al., 2021), the main advantage of Theorem 14 is the presence of the exponential decay $e^{-\lambda\eta(T-t)}$.

6. Conclusion and Future Work

We introduced a framework to understand the generalization error of Markov algorithms through Poissonization. We found a closed-form expression of the associated entropy flow and connected it with a class of modified log-Sobolev inequalities. We showed the relevance of such inequalities in several cases of interest. We further demonstrated the efficiency of our method for both noisy (e.g., SGLD and noisy SGD) and non-noisy algorithms (e.g., SGD).

Future directions. We focused our analysis on KL-based bounds through the function $\Phi(x) = x\log(x)$ and modified LSIs. Another route would be to change Φ to $\Phi_2(x) := \|x\|^2$, in which case our proof technique of Theorem 9 leads to a Poincaré (or spectral gap) inequality. Combined with the theory of Ricci curvature of Markov chains (Ollivier, 2007), this opens new research directions to obtain new generalization bounds and differential privacy guarantees. Finally, the extension of Poissonization to other algorithms (like ADAM) is an important direction for future work.

Acknowledgments

References

Aurélien Alfonsi, Jacopo Corbetta, and Benjamin Jourdain. Evolution of the Wasserstein distance between the marginals of two Markov processes, December 2016.

Pierre Alquier. User-friendly Introduction to PAC-Bayes Bounds. *Foundations and Trends® in Machine Learning*, 2024.

Idan Amir, Roi Livni, and Nathan Srebro. Thinking Outside the Ball: Optimal Learning with Gradient Descent for Generalized Linear Stochastic Convex Optimization, 2022.

Andreas Anastasiou, Krishnakumar Balasubramanian, and Murat A. Erdogdu. Normal approximation for stochastic gradient descent via non-asymptotic rates of martingale clt. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99, pages 115–137, 2019.

Rayna Andreeva, Benjamin Dupuis, Rik Sarkar, Tolga Birdal, and Umut Şimşekli. Topological Generalization Bounds for Discrete-Time Stochastic Optimization Algorithms. In *38th Conference on Neural Information Processing Systems (NeurIPS 2024)*, December 2024.

Cécile Ané and Michel Ledoux. On logarithmic Sobolev inequalities for continuous time random walks on graphs. *Probability Theory and Related Fields*, 116:573–602, 2000.

Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. High-dimensional limit theorems for sgd: Effective dynamics and critical scaling. In *Neurips*, 2023.

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

Peter Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 2002.

Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of Stochastic Gradient Descent on Nonsmooth Convex Losses, June 2020.

Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. High-dimensional limit theorems for SGD: Effective dynamics and critical scaling. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 25349–25362. Curran Associates, Inc., 2022.

S. Bobkov. A Functional Form of the Isoperimetric Inequality for the Gaussian Measure. *Journal of Functional Analysis*, 135(1):39–49, January 1996.

S. G Bobkov and M Ledoux. On Modified Logarithmic Sobolev Inequalities for Bernoulli and Poisson Measures. *Journal of Functional Analysis*, 156(2):347–365, July 1998.

Sergey G. Bobkov and Prasad Tetali. Modified Logarithmic Sobolev Inequalities in Discrete Settings. *Journal of Theoretical Probability*, 19(2):289–336, June 2006.

Vladimir I. Bogachev. *Measure Theory*, volume Volume 1. Springer, 2007.

Olivier Bousquet. Stability and generalization. *Journal of Machine Learning Research*, pages 499–526, 2002.

Yuheng Bu, Shaofeng Zou, and Venugopal V. Veeravalli. Tightening Mutual Information Based Bounds on Generalization Error. *IEEE Journal on Selected Areas in Information Theory*, 1(1): 121–130, May 2020.

Alexander Camuto, George Deligiannidis, Murat A. Erdogdu, Mert Gürbüzbalaban, Umut Şimşekli, and Lingjiong Zhu. Fractal Structure and Generalization Properties of Stochastic Optimization Algorithms. *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, June 2021.

Pietro Caputo, Zongchen Chen, Yuzhou Gu, and Yury Polyanskiy. Entropy Contractions in Markov Chains: Half-Step, Full-Step and Continuous-Time, September 2024.

Ioar Casado, Luis A. Ortega, Aritz Pérez, and Andrés R. Masegosa. PAC-Bayes-Chernoff bounds for unbounded losses, October 2024.

Olivier Catoni. Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning. *IMS Lecture Notes Monograph Series*, 56:1–163, 2007.

Djalil Chafai and Joseph Lehec. Logarithmic sobolev inequalities essentials, 2017.

Nisha Chandramoorthy, Andreas Loukas, Khashayar Gatmiry, and Stefanie Jegelka. On the generalization of learning algorithms that do not converge. *Thirty-Sixth Conference on Neural Information Processing Systems (Neurips 2022)*, August 2022.

Guan-Yu Chen, Wai-Wai Liu, and Laurent Saloff-Coste. The logarithmic Sobolev constant of some finite Markov chains. *Annales de la Faculté des sciences de Toulouse : Mathématiques*, 17(2): 239–290, 2008.

Xiang Cheng, Dong Yin, Peter Bartlett, and Michael Jordan. Stochastic gradient and Langevin processes. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1810–1819. PMLR, 13–18 Jul 2020.

Rishav Chourasia, Jiayuan Ye, and Reza Shokri. Differential Privacy Dynamics of Langevin Diffusion and Noisy Gradient Descent, September 2022.

Eugenio Clerico, Tyler Farghly, George Deligiannidis, Benjamin Guedj, and Arnaud Doucet. Generalisation under gradient descent via deterministic PAC-Bayes, April 2023.

P. Del Moral, M. Ledoux, and L. Miclo. On contraction properties of Markov kernels. *Probability Theory and Related Fields*, 126(3):395–420, June 2003.

P. Diaconis and L. Saloff-Coste. Logarithmic sobolev inequalities for finite markov chains. *The Annals of Applied Probability*, 6(3):695–750, 1996.

Aymeric Dieuleveut, Alain Durmus, and Francis Bach. Bridging the Gap between Constant Step Size Stochastic Gradient Descent and Markov Chains, April 2018.

Benjamin Dupuis and Umut Simsekli. Generalization Bounds for Heavy-Tailed SDEs through the Fractional Fokker-Planck Equation. In *Proceedings of the 41st International Conference on Machine Learning*, pages 12087–12137. PMLR, July 2024.

Benjamin Dupuis and Paul Viallard. From Mutual Information to Expected Dynamics: New Generalization Bounds for Heavy-Tailed SGD, December 2023.

Benjamin Dupuis, Paul Viallard, George Deligiannidis, and Umut Simsekli. Uniform Generalization Bounds on Data-Dependent Hypothesis Sets via PAC-Bayesian Theory on Random Sets, April 2024.

Matthias Erbar and Jan Maas. Ricci curvature of finite markov chains via convexity of the entropy. *Archive for Rational Mechanics and Analysis*, 206(3):997–1038, 2012.

Mathieu Even, Raphaël Berthier, Francis Bach, Nicolas Flammarion, Pierre Gaillard, Hadrien Hendrikx, Laurent Massoulié, and Adrien Taylor. A Continuized View on Nesterov Acceleration for Stochastic Gradient Descent and Randomized Gossip, June 2021.

Tyler Farghly and Patrick Rebeschini. Time-independent Generalization Bounds for SGLD in Non-convex Settings. In *35th Conference on Neural Information Processing Systems (NeurIPS 2021)*. arXiv, November 2021.

Vitaly Feldman and Jan Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate, June 2019.

Eberhard Freitag and Rolf Rusam. *Complex Analysis*. Universitext. Springer-Verlag, Berlin/Heidelberg, 1st edition edition, 2005.

Futoshi Futami and Masahiro Fujisawa. Time-Independent Information-Theoretic Generalization Bounds for SGLD. In *7th Conference on Neural Information Processing Systems (NeurIPS 2023)*. arXiv, November 2023.

M. A. Gallegos-Herrada, D. Ledvinka, and J. S. Rosenthal. Equivalences of geometric ergodicity of markov chains, 2023.

Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. PAC-Bayesian learning of linear classifiers. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 353–360, New York, NY, USA, June 2009. Association for Computing Machinery.

Sharad Goel. Modified logarithmic Sobolev inequalities for some models of random walk. *Stochastic Processes and their Applications*, 114(1):51–79, November 2004.

Leonard Gross. Logarithmic Sobolev Inequalities. *American Journal of Mathematics*, 97(4):1061–1083, 1975.

Gerd Grubb. *Distributions and Operators*, volume 252 of *Graduate Texts in Mathematics*. Springer, New York, NY, 2009.

Mert Gürbüzbalaban, Umut Şimşekli, and Lingjiong Zhu. The heavy-tail phenomenon in SGD. In *International Conference on Machine Learning (ICML)*, 2021.

Maxime Haddouche, Paul Viallard, Umut Simsekli, and Benjamin Guedj. A PAC-Bayesian Link Between Generalisation and Flat Minima, February 2024.

Mahdi Haghifam, Jeffrey Negrea, Ashish Khisti, Daniel M. Roy, and Gintare Karolina Dziugaite. Sharpened Generalization Bounds based on Conditional Mutual Information and an Application to Noisy, Iterative Algorithms, October 2020.

Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent, February 2016.

Liam Hodgkinson, Umut Şimşekli, Rajiv Khanna, and Michael W. Mahoney. Generalization bounds using lower tail exponents in stochastic optimizers, 2022.

Philippe Jacquet and Wojciech Szpankowski. Analytical depoissonization and its applications. *Theoretical Computer Science*, 201(1):1–62, July 1998.

Nurdan Kuru, Ş İlker Birbil, Mert Gurbuzbalaban, and Sinan Yildirim. Differentially Private Accelerated Optimization Algorithms. *SIAM Journal on Optimization*, 32(2):795–821, June 2022.

Serge Lang. *Complex Analysis*, volume 103 of *Graduate Texts in Mathematics*. Springer, New York, NY, fourth edition edition, 1999.

Andrzej Lasota and Michael C. Mackey. *Chaos, Fractals and Noise*. Springer, applied mathematical sciences 97 edition, 1994.

David A. Levin and Yuval Peres. *Markov Chains and Mixing Times*. American Mathematical Society, 2017.

Jian Li, Xuanyuan Luo, and Mingda Qiao. On Generalization Error Bounds of Noisy Gradient Methods for Non-Convex Learning. In *Published as a Conference Paper at ICLR 2020*, February 2020.

Qianxiao Li, Cheng Tai, and E Weinan. Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *Proceedings of the 34th International Conference on Machine Learning*, 2018.

Zhiyuan Li, Sadhika Malladi, and Sanjeev Arora. On the validity of modeling sgd with stochastic differential equations (sdes). In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 12712–12725. Curran Associates, Inc., 2021.

Stephan Mandt, Matthew D. Hoffman, and David M. Blei. A Variational Analysis of Stochastic Gradient Algorithms. In *International Conference on Machine Learning (ICML 2016)*, 2016.

Andreas Maurer. A Note on the PAC Bayesian Theorem, November 2004.

David McAllester. Some PAC-Bayesian theorem. *Machine Learning*, 1999.

David A. McAllester. PAC-Bayesian Stochastic Model Selection. *Machine Learning*, 51(1):5–21, April 2003.

Sean P. Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Springer, 1993.

Wenlong Mou, Liwei Wang, Xiyu Zhai, and Kai Zheng. Generalization Bounds of SGLD for Non-convex Learning: Two Theoretical Viewpoints. In *Proceedings of the 31st Conference On Learning Theory*, 2017.

Jeffrey Negrea, Mahdi Haghifam, Gintare Karolina Dziugaite, Ashish Khisti, and Daniel M Roy. Information-theoretic generalization bounds for SGLD via data-dependent estimates. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Gergely Neu, Gintare Karolina Dziugaite, Mahdi Haghifam, and Daniel M. Roy. Information-Theoretic Generalization Bounds for Stochastic Gradient Descent, August 2021.

Yann Ollivier. Ricci curvature of Markov chains on metric spaces, July 2007.

F. Otto and C. Villani. Generalization of an Inequality by Talagrand and Links with the Logarithmic Sobolev Inequality. *Journal of Functional Analysis*, 173(2):361–400, June 2000.

Ankit Pensia, Varun Jog, and Po-Ling Loh. Generalization Error Bounds for Noisy, Iterative Algorithms. *2018 IEEE International Symposium on Information Theory (ISIT)*, January 2018.

Yury Polyanskiy and Yihong Wu. Wasserstein continuity of entropy and outer bounds for interference channels, February 2016.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via Stochastic Gradient Langevin Dynamics: A nonasymptotic analysis, June 2017.

Anant Raj, Melih Barsbey, Mert Gurbuzbalaban, Lingjiong Zhu, Umut Şim, et al. Algorithmic stability of heavy-tailed stochastic gradient descent on least squares. In *International Conference on Algorithmic Learning Theory*, pages 1292–1342. PMLR, 2023a.

Anant Raj, Lingjiong Zhu, Mert Gurbuzbalaban, and Umut Simsekli. Algorithmic stability of heavy-tailed sgd with general loss functions. In *International Conference on Machine Learning*, pages 28578–28597. PMLR, 2023b.

Daniel Rudolf. Explicit error bounds for Markov chain Monte Carlo. *Dissertationes Mathematicae*, 485:1–93, 2012.

Daniel Rudolf and Nikolaus Schweizer. Perturbation theory for Markov chains via Wasserstein distance, 2017.

Théo Ryffel, Francis Bach, and David Pointcheval. Differential Privacy Guarantees for Stochastic Gradient Langevin Dynamics, February 2022.

René L. Schilling. An Introduction to Lévy and Feller Processes. Advanced Courses in Mathematics - CRM Barcelona 2014, October 2016.

Matthias Seeger. PAC-Bayesian Generalisation Error Bounds for Gaussian Process Classification. *Journal of Machine Learning Research*, 2002.

Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A Tail-Index Analysis of Stochastic Gradient Noise in Deep Neural Networks. In *Proceedings of the 36 Th International Conference on Machine Learning (ICML 2019)*, January 2019.

Umut Şimşekli, Mert Gürbüzbalaban, Sinan Yıldırım, and Lingjiong Zhu. Differential Privacy of Noisy (S)GD under Heavy-Tailed Perturbations, March 2024.

Jozef L. Teugels. A Note on Poisson-Subordination. *The Annals of Mathematical Statistics*, 43(2): 676–680, April 1972.

Brigitte Vallée. The Depoissonisation quintet: Rice-Poisson-Mellin-Newton-Laplace, February 2018.

Vladimir Naumovich Vapnik. *The Nature of Statistical Learning Theory, Second Edition*. Statistics for Engineering and Information Science. Springer, 2000.

Cédric Villani. *Optimal Transport - Old and New*. Springer, 2009.

Neng-Yi Wang and Liming Wu. Transport-information inequalities for Markov chains. *The Annals of Applied Probability*, 30(3):1276–1320, June 2020.

Stephan Wojtowytsch. Stochastic gradient descent with noise of machine learning type. part i: Discrete time analysis, 2021.

Liming Wu. A new modified logarithmic Sobolev inequality for Poisson point processes and several applications. *Probability theory and related fields, 118, 427–438*, 2000.

Yimin Xiao. Random fractals and Markov processes. *Fractal Geometry and Applications: A jubilee of Benoît Mandelbrot - American Mathematical Society*, 72.2:261–338, 2004.

Zeke Xie, Issei Sato, and Masashi Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima, 2021.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, November 2017.

Lingjiong Zhu, Mert Gurbuzbalaban, Anant Raj, and Umut Simsekli. Uniform-in-Time Wasserstein Stability Bounds for (Noisy) Stochastic Gradient Descent, October 2023.

The appendix is organized as follows:

- In Section A, we present additional technical background related to semigroups and their infinitesimal generator, which we use in some of our proofs.
- In Sections B to E, we present all the omitted proofs of our main results.
- Section F presents additional background on depoissonization, to complement the discussion of Sections 2 and 3.

Appendix A. Additional background on semigroup theory

In this subsection, we briefly introduce Markov semigroups, to present concepts and notations that we use in some of our proofs. For elementary introductions, we refer the reader to the tutorials of Schilling (2016, Section 5) and, specifically for diffusion semigroups, Chafai and Lehec (2017). More detailed accounts can be found in (Bakry et al., 2014; Xiao, 2004).

In all this section, we consider the Banach space $\mathcal{C}_\infty$ of functions $\mathbb{R}^d \to \mathbb{R}$ that are continuous and vanish at infinity ($\lim_{\|x\|\to\infty} f(x) = 0$), equipped with the uniform norm $\|\cdot\|_\infty$.

A stochastic process $(X_t)_{t\geq 0}$ is a time-homogeneous Markov process if the law of $(X_s)_{s\geq t}$ given $(X_s)_{s\leq t}$ is the same as the law of $(X_s)_{s\geq t}$ given X_t and the law of $(X_s)_{s\geq 0}$ given X_0.

The semigroup of X is a family of operators $(P_t : L^\infty(\mathbb{R}^d) \to L^\infty(\mathbb{R}^d))_{t\geq 0}$ defined as $P_t f(x) = \mathbb{E}^x[f(X_t)]$, where $\mathbb{E}^x$ means that the process is initialized by $X_0 = x \in \mathbb{R}^d$. This is called a semigroup because of the so-called semigroup property, *i.e.*, $P_t \circ P_s = P_{s+t}$.

Definition 15 (Infinitesimal generators) *The generator L of the semigroup $(P_t)_{t\geq 0}$ is defined as the following limit in $(\mathcal{C}_\infty, \|\cdot\|_\infty)$:*

$$Lf = \lim_{t\to 0} \frac{P_t f - f}{t}.$$

The domain $\mathcal{D}(L)$ *of L is the set of functions for which the above limit exists.*

Under appropriate conditions (see (Schilling, 2016, Lemma 5.4)), L satisfies the backward Kolmogorov equations $\frac{\mathrm{d}}{\mathrm{d}t} P_t f = L P_t f = P_t L f$, which we use in several places. These equations justify the (informal) exponential notation of the semigroup by $P_t = e^{tL}$.

Remark 16 (About $\mathcal{D}(L)$) *In the above definitions, we used the Banach space $\mathcal{C}_\infty$, because it provides a good framework to define semigroups and generators properly (Schilling, 2016). For the semigroups we consider (Ornstein-Uhlenbeck and Langevin semigroups with regular enough potential) we mostly work in the space $\mathcal{C}_b^2(\mathbb{R}^d)$ of bounded twice continuously differentiable functions with bounded derivatives of order 1 and 2 (Chafai and Lehec, 2017).*

Let us quickly give two examples of semigroups and generators of particular interest: Poissonized and Langevin semigroups.

Discrete-time Markov processes. A stochastic process $(X_k)_{k\in\mathbb{N}}$ is a time-homogeneous Markov process if, for all $k \in \mathbb{N}$, the law of X_{k+1} given $(X_0, \ldots, X_k)$ is the same as the law of X_{k+1} given X_k and is independent of k.

Let us provide a semigroup formulation of the Poissonization procedure displayed in Section 2.

Example 2 (Semigroup formulation of Poissonization) *Let us consider a discrete-time Markov process $(X_k)_{k\in\mathbb{N}}$ with kernel P and its Poissonization $(Y_t)_{t\geq 0}$ as defined in Section 2. Then $(Y_t)_{t\geq 0}$ is a Markov process and we can consider its semigroup $(Q_t)_{t\geq 0}$. As noted by Diaconis and Saloff-Coste (1996, Section 2.1), it can be expressed as:*

$$Q_t f(x) = e^{-t} \sum_{k=0}^{+\infty} \frac{t^k}{k!} P^k f(x)$$

Moreover, the infinitesimal generator of the Poissonized semigroup is $L = P - I$.

Langevin semigroups. Finally, let us consider the SDE $\mathrm{d}Z_t = -\nabla V(Z_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$, as in Definition 8. Then the infinitesimal generator of $(Z_t)_{t\geq 0}$ is $Lf = \Delta f - \langle \nabla f, \nabla V\rangle$. Moreover, if we introduce the so-called "carré du champ" operator[8] $\Gamma(f) := \|\nabla f\|^2$, then L satisfies for smooth functions $\varphi, \psi : \mathbb{R}^d \to \mathbb{R}$ the following *diffusion* property $L\varphi(f) = \varphi'(f)Lf + \varphi''(f)\Gamma(f)$, and the integration by part formula $\int \phi L\psi \mathrm{d}\pi = -\int \langle \nabla\varphi, \nabla\psi\rangle \mathrm{d}\pi$, where $\pi \propto e^{-V}$ is a reversible measure for the process. We refer the reader to (Bakry et al., 2014; Chafai and Lehec, 2017) for more background on diffusion semigroups and, in particular, the associated Poincaré and logarithmic Sobolev inequalities, which we will use in some proofs (with appropriate references).

Appendix B. Omitted proofs of Section 2

In this section, we provide the omitted proofs of the technical results mentioned in Section 2, in particular the "Boltzmann" equation (4). We start with the following technical lemma.

Lemma 17 *let $(X_k)_{k\in\mathbb{N}}$ be a sequence of absolutely continuous random variables in $\mathbb{R}^d$ with probability density functions (PDF) denoted p_k and $(c_k)_{k\in\mathbb{N}}$ a sequence of positive numbers such that $\sum_{k\in\mathbb{N}} c_k = 1$. Let Z be a random variable independent of $(X_k)_{k\in\mathbb{N}}$ s.t. $\forall k \in \mathbb{N}$, $\mathbb{P}(Z = k) = c_k$ and define $Y := X_Z$. Then, the PDF of Y is given by $x \mapsto \sum_{k\in\mathbb{N}} c_k p_k(x)$. In particular, the latter sum is almost surely finite in $\mathbb{R}^d$.*

Proof Let $A \in \mathcal{B}(\mathbb{R}^d)$, we have:

$$\begin{aligned}
\mathbb{P}(Y \in A) &= \mathbb{P}\left(\bigcup_{k=0}^{+\infty} \{X_Z \in A, Z = k\}\right) \\
&= \mathbb{P}\left(\bigcup_{k=0}^{+\infty} \{X_k \in A, Z = k\}\right) \\
&= \sum_{k=0}^{\infty} \mathbb{P}(X_k \in A, Z = k) \\
&= \sum_{k=0}^{\infty} \mathbb{P}(X_k \in A)\,\mathbb{P}(Z = k) \quad \text{(independence)} \\
&= \int_A \sum_{k=0}^{\infty} c_k p_k(x)\mathrm{d}x. \quad \text{(Tonelli's theorem)}
\end{aligned}$$

8. The French term is used even in English (Bakry et al., 2014; Chafai and Lehec, 2017).

By definition of the Radon-Nykodym derivative, we obtain the desired PDF for Y. ■

We can now present the proof of the Boltzmann equation (4). It should be noted that the proof is formally similar to the proof of (Lasota and Mackey, 1994, Equation 8.3.7), which we adapt to our setting. We report it here because both setups are slightly different and to justify our technical assumptions.

Lemma 18 (Boltzman equation) *Let $(X_k)_{k\in\mathbb{N}}$ be a discrete-time Markov chain in $\mathbb{R}^d$ such that for all $k \in \mathbb{N}$, X_k has a PDF denoted p_k. Let $(Y_t)_{t>0}$ be the Poissonized process as defined in Section 2. Then, Y_t has a PDF $u_t(x) = u(t,x)$ which satisfies the following Boltzmann equation:*

$$\frac{\partial u_t}{\partial t} = (P^\star - I)u_t.$$

Proof Let $t > 0$, we first apply Lemma 17 with $Z = N_t$ and $c_k = e^{-t}t^k/k!$. Therefore, we have the following **expression of the density of** Y_t:

$$\boxed{u(t,x) = e^{-t}\sum_{k=0}^{\infty}\frac{t^k}{k!}p_k(x).}$$

Now let $a < b$ and consider $t \in (a,b)$, we apply again Lemma 17 with $c_0 = 0$ and $c_k = e^{-b}b^{k-1}/(k-1)!$ for $k \geq 1$, this gives in particular that the sum $\sum_{k\geq 1} b^{k-1}/(k-1)!p_k(x)$ is finite for almost all $x \in \mathbb{R}^d$. Thus, we can differentiate under the sum (Bogachev, 2007, Corollary 2.8.7) and obtain that for almost all $x \in \mathbb{R}^d$, we have:

$$\begin{aligned}\frac{\partial u}{\partial t}(t,x) &= -e^{-t}\sum_{k=0}^{\infty}\frac{t^k}{k!}p_k(x) + e^{-t}\sum_{k=1}^{\infty}\frac{t^{k-1}}{(k-1)!}p_k(x)\\ &= -u(t,x) + e^{-t}\sum_{k=0}^{\infty}\frac{t^k}{k!}P^\star p_k(x),\end{aligned}$$

where we used that $p_{k+1} = P^\star p_k$ by the notations of Equation (3) (recall that p_k denotes the PDF of X_k). Using Tonelli's theorem twice, we get that for any $A \in \mathcal{B}(\mathbb{R}^d)$, we have:

$$\begin{aligned}\int_A e^{-t}\sum_{k=0}^{\infty}\frac{t^k}{k!}P^\star p_k(x)\mathrm{d}x &= e^{-t}\sum_{k=0}^{\infty}\frac{t^k}{k!}\int_A P^\star p_k(x)\mathrm{d}x\\ &= e^{-t}\sum_{k=0}^{\infty}\frac{t^k}{k!}\int P\mathbb{1}_A(x)p_k(x)\mathrm{d}x\\ &= \int P\mathbb{1}_A(x)u(t,x)\mathrm{d}x\end{aligned}$$

Moreover, the function $x \mapsto \sum_{k\in\mathbb{N}}\frac{t^k}{k!}P^\star p_k(x)$ is Borel measurable as a limit superior of measurable functions. This shows that, if ρ_t denotes the law of Y_t, then $\rho_t P$ has a PDF given by the preceding sum. Thus, according to Equation (3), we can write:

$$\int P\mathbb{1}_A(x)u(t,x)\mathrm{d}x = \int_A P^\star u(t,\cdot)(x)\mathrm{d}x.$$

This implies that for all $A \in \mathcal{B}(\mathbb{R}^d)$, we have:

$$\int_A \frac{\partial u}{\partial t}(t,x)\mathrm{d}x = \int_A \{-u(t,x) + P^\star u(t,x)\}\,\mathrm{d}x.$$

This finally implies the desired equation. ■

Remark 19 (Weak form of the Boltzmann equation) *Let us denote by ρ_t the probability distribution of the Poissonized process Y_t. Then we can write:*

$$\boxed{\frac{\partial \rho_t}{\partial t} = \rho_t(P - I).} \tag{10}$$

Equation (10) should be understood in a weak sense, i.e., for all test function f we have:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int f\mathrm{d}\rho_t = \int f\mathrm{d}(\rho_t P) - \int f\mathrm{d}\rho_t.$$

Appendix C. Omitted proofs of Section 3

We first prove Theorem 3, concerning invariant measures of Poissonized processes.

Theorem 3 *Assume that $|\ell| \leq B < \infty$ and that $\mathrm{TV}(\mu_k^S, \mu^S) \longrightarrow 0$ for some $\mu^S \in \mathcal{P}(\mathbb{R}^d)$, a.s. for S. Then, a.s., $\mathbb{E}\left[|G_S(X_k^S) - G_S(Y_k^S)|\middle|S\right] \longrightarrow 0$. If moreover there exists $C > 0$ and $a_S \in (0,1)$ such that, a.s., $\mathrm{TV}(\mu_k^S, \mu^S) \leq C_S a_S^k$, then, a.s., $\mathbb{E}\left[|G_S(X_k^S) - G_S(Y_k^S)|\middle|S\right] \leq 4BC_S e^{-(1-a_S)k}$. If ℓ is L-Lipschitz, then we can replace TV by the 1-Wasserstein distance W_1 (and $2B$ by L) in these statements, e.g., if $W_1(\mu_k^S, \mu^S) \leq C_S a_S^k$, then $\mathbb{E}\left[|G_S(X_k^S) - G_S(Y_k^S)|\middle|S\right] \leq 2LC_S e^{-(1-a_S)k}$.*

We use the convention: $\mathrm{TV}(\mu,\nu) := \sup_{A\in\mathcal{B}(\mathbb{R}^d)} |\mu(A) - \nu(A)|$.

Proof First part of the statement. Let us fix $S \in \mathcal{Z}^n$ such that the convergences that are assumed almost surely hold (*i.e.*, $\mathrm{TV}(\mu_k^S, \mu^S) \longrightarrow 0$). Let μ_k^S denote the law of X_k^S, we have, for any $A \in \mathcal{B}(\mathbb{R}^d)$ (see Lemma 18):

$$\rho_t^S(A) = e^{-t}\sum_{k=0}^{\infty}\frac{t^k}{k!}\mu_k^S(A).$$

As μ_k^S converges to μ^S in total variation, for all $\varepsilon > 0$ there exists $K \in \mathbb{N}$, such that for all $k \geq K$ and all $A \in \mathcal{B}(\mathbb{R}^d)$, $|\mu_k^S(A) - \mu^S(A)| \leq \varepsilon$ (*i.e.*, K does not depend on A). Therefore, we have:

$$\forall A \in \mathcal{B}(\mathbb{R}^d),\ |\rho_t^S(A) - \mu^S(A)| \leq \varepsilon + 2e^{-t}\sum_{k=0}^{K-1}\frac{t^k}{k!},$$

where the last term is smaller than ε for all t greater than some $t_0(K)$, depending only on K. This shows that $\mathrm{TV}(\rho_t^S, \mu_S) \longrightarrow 0$, hence, by the triangle inequality and boundedness of ℓ, we get:

$$\mathbb{E}\left[|G_S(X_k^S) - G_S(Y_k^S)|\middle|S\right] \leq 2\,\|\ell\|_\infty\,\mathrm{TV}(\mu_k^S, \rho_k^S) \underset{k\to\infty}{\longrightarrow} 0.$$

Now we assume that there exists $C_S > 0$ and $a_S \in (0,1)$ such that $\forall k \in \mathbb{N},\ \mathrm{TV}(\mu_k^S, \mu^S) \leq C_S a_S^k$, then we have, for any $A \in \mathcal{B}(\mathbb{R}^d)$:

$$|\rho_t^S(A) - \mu^S(A)| \leq e^{-t} \sum_{k=0}^{\infty} \frac{t^k}{k!} |\mu_k^S(A) - \mu^S(A)| \leq C_S e^{-t} \sum_{k=0}^{\infty} \frac{(a_S t)^k}{k!} = C_S e^{-(1-a_S)t}.$$

Thus, by the triangle inequality for total variation, we get, for all $k \in \mathbb{N}$:

$$\mathrm{TV}(\mu_k^S, \rho_k^S) \leq \mathrm{TV}(\mu_k^S, \mu^S) + \mathrm{TV}(\rho_k^S, \mu^S) \leq C_S a_S^k + C_S e^{-(1-a_S)k} \leq 2C_S e^{-(1-a_S)k},$$

where we used that $a \leq e^{-(1-a)}$, hence, $\mathbb{E}\left[|G_S(X_k) - G_S(Y_k)| \middle| S\right] \leq 4 \|\ell\|_\infty C_S e^{-(1-a_S)k}$.

Second part of the statement. We now assume that ℓ is L-Lipschitz continuous and use Wasserstein distance instead of TV. We sketch the proof as it is similar to the previous case. The main argument, is that by convexity of the Wasserstein distance W_1 (see (Farghly and Rebeschini, 2021, Lemma 2.3) and (Villani, 2009, Theorem 4.8)), we have:

$$W_1(\rho_t^S, \mu_S) = W_1\left(e^{-t} \sum_{k \in \mathbb{N}} \frac{t^k}{k!} \mu_k^S, e^{-t} \sum_{k \in \mathbb{N}} \frac{t^k}{k!} \mu^S\right) \tag{11}$$

$$\leq e^{-t} \sum_{k \in \mathbb{N}} \frac{t^k}{k!} W_1(\mu_k^S, \mu_S). \tag{12}$$

Then, if we assume that $W_1(\mu_k^S, \mu^S) \longrightarrow 0$ and fix $\varepsilon > 0$, we know that there exists $K \in \mathbb{N}$ such that $\forall k \geq K,\ W_1(\mu_k^S, \mu^S) \leq \varepsilon$ and we obtain that $W_1(\rho_t^S, \mu_S) \longrightarrow 0$ by noting that:

$$W_1(\rho_t^S, \mu_S) \leq \varepsilon + \left(\max_{0 \leq k \leq K} W_1(\mu_k^S, \mu^S)\right) e^{-t} \sum_{k=0}^{K-1} \frac{t^k}{k!} \longrightarrow 0.$$

We conclude by Kantorovith duality (Villani, 2009, Theorem 5.10) and the triangle inequality.

Finally, under a Wasserstein ergodicity assumption, *i.e.*, $W_1(\rho_t^S, \mu_S) \leq C_S a_S^k$ with $C_S > 0$ and $a_S \in (0,1)$, Equation (11) implies that $W_1(\rho_t^S, \mu_S) \leq C_S e^{-(1-a_S)t}$. We conclude again by Kantorovith duality and the triangle inequality for W_1.

■

We can now prove our entropy flow formula, *i.e.*, Theorem 4.

Theorem 4 (Poissonized entropy flow) *Under Assumption 1, the entropy flow is given by:*

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathrm{KL}\left(\rho_t^S || \pi_t\right) = \Delta_{P,P_S}(v_t) - \mathbb{E}_{x \sim \pi_t, y \sim \delta_x P}\left[D_\Phi\left(v_t(x), v_t(y)\right)\right], \tag{5}$$

*where $D_\Phi(a,b) := \Phi(a) - \Phi(b) - \Phi'(b)(a-b)$ is the Bregman divergence. We call the first term $\Delta_{P,P_S}(v_t) := \mathbb{E}_{\rho_t^S}\left[(P_S - P)\log(v_t)\right]$ the **expansion term** and the second the **Bregman term**.*

Proof In this proof, we use the notation ∂_t as a shortcut for $\partial/\partial t$. We also recall the notation $\Phi(x) = x\log(x)$ (the reader may note that this proof is valid for more general convex functions Φ). We also use the notations $L_S := P_S - I$ and $L := P - I$ (which correspond to the infinitesimal generators of the Poissonized processes). We denote accordingly $L_S^\star = P_S^\star - I$ and $L^\star = P^\star - I$.

We start by noticing that Assumption 1 additionally implies $Pv_t \in L^1(\pi_t)$, indeed, if we let $y_0 = \inf_{y \geq 0} \{\Phi(y) \geq 1\}$, we can show that $0 \leq Pv_t \leq y_0 + \frac{2}{e} + P\Phi(f)$, which is in $L^1(\pi_t)$ by Assumption 1. Thus, by Item **H.1** in Assumption 1 and Equation (3), we obtain:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}\left(\rho_t^S||\pi_t\right) &= \frac{\mathrm{d}}{\mathrm{d}t}\int \Phi(v_t)u_t\mathrm{d}x \\
&= \int \Phi'(v_t)(\partial_t v_t)u_t\mathrm{d}x + \int \Phi(v_t)(\partial_t u_t)\mathrm{d}x \\
&= \int \Phi'(v_t)L_S^\star u_t^S\mathrm{d}x - \int \Phi'(v_t)v_t L^\star u_t\mathrm{d}x + \int \Phi(v_t)L^\star u_t\mathrm{d}x \\
&= \int L_S\left(\Phi'(v_t)\right)u_t^S\mathrm{d}x - \int L\left(\Phi'(v_t)v_t\right)u_t\mathrm{d}x + \int L\left(\Phi(v_t)\right)u_t\mathrm{d}x \\
&= \int L_S\left(\Phi'(v_t)\right)u_t^S\mathrm{d}x \\
&+ \iint u_t(x)\left[v_t(x)\Phi'(v_t(x)) - v_t(y)\Phi'(v_t(y)) + \Phi(v_t(y)) - \Phi(v_t(x))\right]P(x,\mathrm{d}y)\mathrm{d}x.
\end{aligned}$$

Note that by the fact that $Pv_t \in L^1(\pi_t)$ and Assumption 1, all the integrals above are well-defined. By recognizing part of the desired Bregmann divergence, we obtain:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}\left(\rho_t^S||\pi_t\right) = \int L_S\left(\Phi'(v_t)\right)u_t^S\mathrm{d}x - \int P\left(\Phi'(v_t)\right)v_t u_t\mathrm{d}x + \int u_t v_t\Phi'(v_t)\mathrm{d}x \\
- \iint D_\Phi\left(v_t(x), v_t(y)\right)u_t(x)P(x,\mathrm{d}y)\mathrm{d}x,
\end{aligned}$$

which leads to the result by recalling that $L = P - I$ and noting that by definition $L_S - L = P_S - P$ and $u_t v_t = u_t^S$ as well as the obvious fact that $\Phi'(x) = 1 + \log x$. ■

Appendix D. Omitted proofs of Section 4

D.1. Omitted proofs of Section 4.1

Corollary 6 *Assume that Assumption 1 holds and that P has an invariant measure π. We have*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}\left(\rho_t^S||\pi\right) = \mathbb{E}_{\rho_t^S}\left[(P_S - P)(\log v_t)\right] - \mathcal{E}_\pi(\log v_t, v_t).$$

Proof Let $\Phi(x) = x\log(x)$. By Assumption 1 and invariance of π under P, we have (with $f = v_t$):

$$\begin{aligned}
\mathcal{E}_\pi(\Phi'\circ f, f) &= \int f(I-P)(\Phi'\circ f)\mathrm{d}\pi \\
&= \iint f(x)(\Phi'(f(x)) - \Phi'(f(y)))P(x,\mathrm{d}y)\mathrm{d}\pi(x) \\
&= \iint \left[\Phi(f(x)) - \Phi(f(y)) + f(x)(\Phi'(f(x)) - \Phi'(f(y)))\right]P(x,\mathrm{d}y)\mathrm{d}\pi(x) \\
&= \iint \left[\Phi(f(x)) - \Phi(f(y)) + f(y)\Phi'(f(y)) - f(x)\Phi'(f(y))\right]P(x,\mathrm{d}y)\mathrm{d}\pi(x) \\
&= \iint D_\Phi(f(x), f(y))P(x,\mathrm{d}y)\mathrm{d}\pi(x).
\end{aligned}$$

This completes the proof. ■

Theorem 7 (Generalization error of Poissonized algorithms) *Assume that ℓ is s^2-subgaussian, Assumption 1 holds, and the prior dynamics has an invariant measure π which satisfies a modified LSI with constant γ, in the sense of Definition 5. Then, with probability at least $1-\zeta$ under $S \sim \mu_z^{\otimes n}$:*

$$\mathbb{E}_{w\sim\rho_T^S}\left[G_S(w)\right] \leq \frac{2s}{\sqrt{n}}\left\{\int_0^T e^{-\gamma(T-t)}\Delta_{P,P_S}(v_t)\mathrm{d}t + e^{-\gamma T}\mathrm{KL}\left(p_0||\pi\right) + \log\left(\frac{3}{\zeta}\right)\right\}^{\frac{1}{2}}.$$

Proof By the subgaussian assumption on the loss ℓ, we can apply Theorem 2 to get that:

$$\mathbb{P}_{S\sim\mu_z^{\otimes n}}\left(\mathbb{E}_{\rho_t^S}\left[G_S\right] \leq 2s\sqrt{\frac{\mathrm{KL}\left(\rho_t^S||\pi\right)+\log(3/\zeta)}{n}}\right) \geq 1-\zeta. \tag{13}$$

Now by Assumption 1, we apply Corollary 6, which gives:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}\left(\rho_t^S||\pi\right) = \Delta_{P,P_S}(v_t) - \mathcal{E}_\pi(\log(v_t), v_t) \leq \Delta_{P,P_S}(v_t) - \gamma\mathrm{KL}\left(\rho_t^S||\pi\right),$$

where the inequality follows from the modified LSI and noting that $\mathrm{Ent}_\pi\left(v_t\right)$. Now we solve this differential inequality by looking at $F(t) := e^{\gamma t}\mathrm{KL}\left(\rho_t^S||\pi\right)$. A simple calculation provides $F'(t) \leq e^{\gamma t}\Delta_{P,P_S}(v_t)$. By integrating, we immediately obtain:

$$\mathrm{KL}\left(u_t^S||\pi\right) \leq e^{-\gamma t}\mathrm{KL}\left(u_0||\pi\right) + \int_0^t e^{-\gamma(t-s)}\Delta_{P,P_S}(v_s)\mathrm{d}s.$$

The result follows by using this inequality inside Equation (13). ■

D.2. Omitted proofs of Section 4.2

For a probability measure ν and a function f, we recall the notation for the entropy functional:

$$\mathrm{Ent}_\nu^\Phi\left(f\right) := \mathbb{E}_\nu\left[\Phi(f)\right] - \Phi\left(\mathbb{E}_\nu\left[f\right]\right).$$

Theorem 9 (Modified LSI for diffusive priors) *Assume the Markov kernel P to be representable at time t_0 by a diffusion with an ergodic invariant measure π, as in Definition 8. Let $K > 0$ and $f \in \mathcal{C}^2(\mathbb{R}^d)$ a positive function s.t. $\forall x,\ f, \log(f) \in L^1(\delta_x P)$, and $f\log(f),\ Pf\log(f) \in L^1(\pi)$. If π satisfies a LSI with constant K, then we have the modified LSI:*

$$\mathcal{E}_\pi(\log f, f) \geq c_{\mathrm{LSI}}\mathrm{Ent}_\pi\left(f\right), \tag{7}$$

with $c_{\mathrm{LSI}} = \frac{Kt_0}{1+Kt_0}$. If we have $\nabla^2 V \succeq KI_d$, then the constant is improved to $c_{\mathrm{LSI}} = 1 - e^{-Kt_0}$.

Proof In this proof, we use the function $\Phi(x) = x\log(x)$. We will use repeatedly the operator Γ associated with the Langevin equation of Definition 8, $\Gamma(\psi) := \|\nabla\psi\|^2$. It has been briefly introduced in Section A, see (Chafai and Lehec, 2017; Bakry et al., 2014) for more details.

Step 1: Let $f \in \mathcal{C}^2(\mathbb{R}^d)$ satisfy the assumptions of the theorem.

Let $\mathcal{E}_\pi$ be the Dirichlet form associated to P, (see Corollary 6) and (H_t) the diffusion semigroup representing P in the sense of Definition 8. We denote by L_H the infinitesimal generator of (H_t). It is known that the invariant measure π is reversible for the semigroup (H_t) (Chafai and Lehec, 2017, Chapter 5) Using the concavity of Φ' we get:

$$\begin{aligned}
\mathcal{E}_\pi(\Phi' \circ f, f) &= \int f\Phi'(f)\mathrm{d}\pi - \int fP\Phi'(f)\mathrm{d}\pi \\
&= \int f\Phi'(f)\mathrm{d}\pi - \int fH_{t_0}\Phi'(f)\mathrm{d}\pi \quad \text{(representability)} \\
&= \int f\Phi'(f)\mathrm{d}\pi - \int H_{\frac{t_0}{2}} f H_{\frac{t_0}{2}}\Phi'(f)\mathrm{d}\pi \quad \text{(reversibility and semigroup property)} \\
&\geq \int f\Phi'(f)\mathrm{d}\pi - \int H_{\frac{t_0}{2}} f\Phi'(H_{\frac{t_0}{2}} f)\mathrm{d}\pi \quad \text{(Jensen's inequality)} \\
&= \int H_{\frac{t_0}{2}}(f\Phi'(f))\mathrm{d}\pi - \int H_{\frac{t_0}{2}} f\Phi'(H_{\frac{t_0}{2}} f)\mathrm{d}\pi \quad \text{(invariance)} \\
&=: \bar{\mathcal{E}}_\pi(f).
\end{aligned}$$

Note that $H_{\frac{t_0}{2}} f\Phi'(H_{\frac{t_0}{2}} f) \in L^1(\pi)$ in virtue of the inequalities $-1/e \leq H_{\frac{t_0}{2}} f\Phi'(H_{\frac{t_0}{2}} f) \leq H_{\frac{t_0}{2}}\Phi(f) + H_{\frac{t_0}{2}} f$, which is in $L^1(\pi)$ by Assumption 1.

We will now prove an inequality satisfied by $\bar{\mathcal{E}}_\pi(f)$. As a first step, we additionally assume that f is bounded from below by a positive constant (*i.e.* $f \geq \epsilon > 0$) and that f is bounded and has bounded derivatives or order 1 and 2. This ensures that all the derivations below are justified. Let $\Psi(x) := x\Phi'(x) - \Phi(x)$, which is convex (for our choice of function Φ, it is the identity) and obtain:

$$\bar{\mathcal{E}}_\pi(f) = \int \mathrm{Ent}^{\Psi+\Phi}_{H_{\frac{t_0}{2}}}\mathrm{d}\pi = \int \mathrm{Ent}^{\Phi}_{H_{\frac{t_0}{2}}}(f)\,\mathrm{d}\pi + \int \mathrm{Ent}^{\Psi}_{H_{\frac{t_0}{2}}}(f)\mathrm{d}\pi \geq \int \mathrm{Ent}^{\Phi}_{H_{\frac{t_0}{2}}}(f)\,\mathrm{d}\pi.$$

Additionally, let's recall the following two classical computations, which follow from the diffusion property of L_H and the integration by parts formula for L_H (Chafai and Lehec, 2017, Section 5).

$$\frac{\mathrm{d}}{\mathrm{d}t}\int \mathrm{Ent}^{\Phi}_{H_t}(f)(x)\mathrm{d}\pi(x) = -\int \Phi'(H_t f)L_H H_t f\mathrm{d}\pi = \int \Phi''(H_t f)\Gamma(H_t f)\mathrm{d}\pi, \tag{14}$$

and by ergodicity we have $\mathrm{Ent}^{\Phi}_{H_0}(f) = 0 = \lim_{t\to\infty}\mathrm{Ent}^{\Phi}_{\pi}(H_t f)$. Similarly:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Ent}^{\Phi}_{\pi}(H_t f) = \int \Phi'(H_t f)L_H H_t f\mathrm{d}\pi = -\int \Phi''(H_t f)\Gamma(H_t f)\mathrm{d}\pi. \tag{15}$$

From Equation (14) we deduce that:

$$\bar{\mathcal{E}}_\pi(f) \geq \int \mathrm{Ent}^{\Phi}_{H_{\frac{t_0}{2}}}(f)\,\mathrm{d}\pi = \int_0^{\frac{t_0}{2}} \int \Gamma(H_s f)\Phi''(H_s f)\mathrm{d}\pi\mathrm{d}s. \tag{16}$$

Proof of Equation (7): Now we can apply the logarithmic Sobolev inequality for π to obtain:

$$\begin{aligned}
\bar{\mathcal{E}}_\pi(f) &\geq \int_0^{\frac{t_0}{2}} \int \frac{\Gamma(H_s f)}{H_s f} \mathrm{d}\pi \mathrm{d}s \\
&\geq 2K \int_0^{\frac{t_0}{2}} \mathrm{Ent}_\pi\left(H_s f\right) \mathrm{d}s \\
&= 2K \int_0^{\frac{t_0}{2}} \int_s^{+\infty} \int \frac{\Gamma(H_u f)}{H_u f} \mathrm{d}\pi \mathrm{d}u \mathrm{d}s \quad \text{(Equation (15) and ergodicity)} \\
&= 2K \int_0^{+\infty} \min\left(u, \frac{t_0}{2}\right) \int \frac{\Gamma(H_u f)}{H_u f} \mathrm{d}\pi \mathrm{d}u \\
&\geq K t_0 \int_{\frac{t_0}{2}}^{+\infty} \int \frac{\Gamma(H_u f)}{H_u f} \mathrm{d}\pi \mathrm{d}u.
\end{aligned}$$

Combining both inequalities gives us that, for any $a \in [0,1]$, we have:

$$\bar{\mathcal{E}}_\pi(f) \geq (1-a) \int_0^{\frac{t_0}{2}} \int \frac{\Gamma(H_u f)}{H_u f} \mathrm{d}\pi \mathrm{d}u + aKt_0 \int_{\frac{t_0}{2}}^{+\infty} \int \frac{\Gamma(H_u f)}{H_u f} \mathrm{d}\pi \mathrm{d}u,$$

which leads to:

$$\begin{aligned}
\bar{\mathcal{E}}_\pi(f) &\geq \sup_{0 \leq a \leq 1} \min(1-a, aKt_0) \int_0^{+\infty} \int \frac{\Gamma(H_u f)}{H_u f} \mathrm{d}\pi \mathrm{d}u \\
&= \frac{Kt_0}{Kt_0 + 1} \mathrm{Ent}_\pi\left(f\right),
\end{aligned}$$

where we used that by Equation (14), we have:

$$\mathrm{Ent}_\pi\left(f\right) = \int_0^{+\infty} \int \frac{\Gamma(H_s f)}{H_s f} \mathrm{d}\pi \mathrm{d}s.$$

Case where $\nabla^2 V \succeq K I_d$: Because of the strong convexity assumption on V, by (Chafai and Lehec, 2017, Lemma 5.6), we know that the semigroup $(H_t)_{t>0}$ satisfies the $\mathrm{CD}(K, \infty)$ conditions (see (Bakry et al., 2014; Chafai and Lehec, 2017)).

By the formula for $\bar{\mathcal{E}}_\pi(f)$ and the reversed local LSI (Bakry et al., 2014, Theorem 5.5.2) we have:

$$\bar{\mathcal{E}}_\pi(f) \geq \int \mathrm{Ent}^{\Phi}_{H_{\frac{t_0}{2}}}\left(f\right) \mathrm{d}\pi \geq \frac{e^{Kt_0} - 1}{2K} \int \frac{\Gamma(H_{\frac{t_0}{2}} f)}{H_{\frac{t_0}{2}} f} \mathrm{d}\pi.$$

By the $\mathrm{CD}(K, \infty)$ condition and ergodicity of π, it is known that π satisfies the (classical) LSI with constant K (Chafai and Lehec, 2017, Theorem 5.10). Thus, by this LSI and Equation (15), we get:

$$\bar{\mathcal{E}}_\pi(f) \geq \left(e^{Kt_0} - 1\right) \mathrm{Ent}_\pi\left(H_{\frac{t_0}{2}} f\right) = \left(e^{Kt_0} - 1\right) \int_{\frac{t_0}{2}}^{+\infty} \int \frac{\Gamma(H_s f)}{H_s f} \mathrm{d}\pi \mathrm{d}s. \tag{17}$$

Combining Equation (16) and Equation (17), we obtain as in the previous case that:

$$\begin{aligned}\bar{\mathcal{E}}_\pi(f) &\geq \sup_{a\in[0,1]} \min\left(a,(1-a)\left(e^{Kt_0}-1\right)\right)\int_0^{+\infty}\int \frac{\Gamma(H_s f)}{H_s f}\mathrm{d}\pi\mathrm{d}s \\ &= \frac{e^{Kt_0}-1}{e^{Kt_0}}\int_0^{+\infty}\int \frac{\Gamma(H_s f)}{H_s f}\mathrm{d}\pi\mathrm{d}s.\end{aligned}$$

Step 2: We have proven two inequalities of the form $\bar{\mathcal{E}}_\pi(f) \geq c_{\mathrm{LSI}}\mathrm{Ent}_\pi(f)$ for functions f satisfying the assumptions of the theorem and bounded from below and with bounded derivatives of order 0, 1 and 2 (the constant $c_{\mathrm{LSI}} > 0$ depends on whether or not we assume $\nabla^2 V \succeq K I_d$). We finish the proof by classical approximation arguments, see for instance (Otto and Villani, 2000).

Let f satisfy the assumptions of the theorem. First, we assume that f is additionally bounded from below by some $\epsilon > 0$. Let $\varphi_n \in \mathcal{C}_c^\infty(\mathbb{R}^d)$ be a sequence of smooth functions with compact support such that $0 \leq \varphi_n \leq 1$ and $\varphi_n \longrightarrow 1$ pointwise (it can be constructed through the theorem of partitions of unity, see (Grubb, 2009, Theorem 2.17)). Let $f_n := \epsilon + \varphi_n(f-\epsilon)$. Then we have $|\Phi(f_n)| \leq f\left(2|\log(\epsilon)| + \log(f)\right) \in L^1(\pi)$, hence by the dominated convergence theorem:

$$\mathrm{Ent}_\pi(f_n) \longrightarrow \mathrm{Ent}_\pi(f).$$

By a similar argument, we have $\int H_{\frac{t_0}{2}} f_n \log H_{\frac{t_0}{2}} f_n \mathrm{d}\pi \longrightarrow \int Pf \log f \mathrm{d}\pi$. This is enough to extend the inequalities to functions f that are bounded from below.

We now extend to general f satisfying the assumptions of the theorem. For $n \geq 1$, let $\bar{f}_n := f_n + \frac{1}{n}$. Using the properties of Φ, we prove that $-\frac{1}{e} \leq 2|\Phi(f)| + 2\log(2)(1+f) \in L^1(\pi)$, hence, by the dominated convergence theorem we have $\mathrm{Ent}_\pi\left(\bar{f}_n\right) \longrightarrow \mathrm{Ent}_\pi(f)$. We also clearly have $H_t f_n \to H_t f$ pointwise. Moreover, by Jensen's inequality, we have:

$$\frac{-1}{e} \leq H_{\frac{t_0}{2}}\bar{f}_n \log H_{\frac{t_0}{2}}\bar{f}_n \leq H_{\frac{t_0}{2}}(\bar{f}_n \log \bar{f}_n) \leq H_{\frac{t_0}{2}}\left(2|\Phi(f)| + 2\log(2)(1+f)\right),$$

hence, $H_{\frac{t_0}{2}}\bar{f}_n \log H_{\frac{t_0}{2}} \leq \frac{2}{e} + P(|\Phi(f)|) \in L^1(\pi)$. Therefore we conclude again by the dominated convergence theorem that $\int H_{\frac{t_0}{2}}\bar{f}_n \log H_{\frac{t_0}{2}}\bar{f}_n \mathrm{d}\pi \longrightarrow \int Pf \log f \mathrm{d}\pi$. This concludes the proof. ■

Corollary 10 *Consider the Markov process defined by $X_{k+1} = (1-\gamma)X_k + \sigma\mathcal{N}(0, I_d)$ with $\gamma \in (0,1), \sigma > 0$. Then the associated Dirichlet form $\mathcal{E}_\pi$ satisfies a modified LSI with constant γ.*

Proof First, let us note that the Markov process $(X_k)_{k\in\mathbb{N}}$ admits an invariant distribution $\pi = \mathcal{N}(0, \sigma_\pi^2)$ with $\sigma_\pi := \sigma/\sqrt{1-(1-\gamma^2)}$. Consider an Ornstein-Uhlenbeck process $\mathrm{d}Z_t = -V(Z_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$ with $V(x) := \frac{c}{2}\|x\|^2$ and set:

$$c := \frac{1}{\sigma_\pi^2} = \frac{1-(1-\gamma)^2}{\sigma^2}, \quad t_0 := \frac{1}{c}\log\left(\frac{1}{1-\gamma}\right).$$

Let (H_t) be the semigroup of (Z_t). By Mehler's formula (see (Chafai and Lehec, 2017)), one may note that $P = H_{t_0}$ and that the invariant measure of (Z_t) is π. Finally, we note that $\nabla^2 V = cI_d$ and that $ct_0 = -\log(1-\gamma)$. The inequality then follows from Theorem 9. ■

Appendix E. Omitted proofs of Section 5

Corollary 11 *Under the above conditions and Assumption 1, a noisy algorithm satisfies:*

$$\mathrm{KL}\left(\rho_T^S||\pi_T\right) \leq \mathrm{KL}\left(p_0||\pi_0\right) + \int_0^T \mathbb{E}_{\rho_t^S}\left[\mathrm{KL}\left(\delta_x P_S||\delta_x P\right)\right] \mathrm{d}t - \int_0^T \mathbb{E}_{\pi_t}\left[D_\Phi(v_t, Pv_t)\right] \mathrm{d}t. \quad (8)$$

If P has an invariant measure π and we use $\forall t,\ \pi_t = \pi$, then we can simplify the last term as $\mathbb{E}_{\pi_t}\left[D_\Phi(v_t, Pv_t)\right] = \mathrm{KL}\left(\rho_t^S||\rho_t^S P^\dagger\right)$, where $P^\dagger$ is the adjoint of P in $L^2(\pi)$.

Proof Theorem 4 gives us that:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}\left(\rho_t^S||\pi_t\right) = \mathbb{E}_{x\sim\pi_t}\left[v_t(P_S - P)\log v_t\right] - \mathbb{E}_{x\sim\pi_t, y\sim\delta_x P}\left[D_\Phi(v_t(x), v_t(y))\right].$$

By the inequality $a(\log a - \log b) - (a-b) \geq 0$, we have $v_t P(\log v_t) \leq v_t \log Pv_t \leq v_t \log v_t - v_t + Pv_t$, which by Assumption 1 implies that $v_t \log Pv_t \in L^1(\pi_t)$ (we have proven that $Pv_t \in L^1(\pi_t)$ in the proof of Theorem 4). Therefore, by Donsker-Varadhan's formula, absolute continuity property, and the positivity of v_t we have:

$$\forall x \in \mathbb{R}^d,\ P_S \log v_t(x) \leq \mathrm{KL}\left(\delta_x P_S||\delta_x P\right) + \log Pv_t(x).$$

Therefore, using the expression of the Bregman divergence of $\Phi(x) = x\log(x)$, we can write:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}\left(\rho_t^S||\pi_t\right) \leq\ & \mathbb{E}_{x\sim\rho_t^S}\left[\mathrm{KL}\left(\delta_x P_S||\delta_x P\right)\right] + \mathbb{E}_{x\sim\pi_t}\left[v_t(x)\log Pv_t(x) - v_t(x)P\log v_t(x)\right] \\ & - \mathbb{E}_{x\sim\pi_t}\left[v_t(x)(\log v_t(x) - P\log v_t(x)) - (v_t(x) - Pv_t(x))\right] \\ \leq\ & \mathbb{E}_{x\sim\rho_t^S}\left[\mathrm{KL}\left(\delta_x P_S||\delta_x P\right)\right] - \mathbb{E}_{\pi_t}\left[v_t\left(\log v_t - \log Pv_t\right) - (v_t - Pv_t)\right].\end{aligned}$$

The result follows by integrating and recognizing the Bregman divergence of $\Phi : x \mapsto x\log(x)$. Finally, we assume that π is an invariant measure for P. As we assumed $\delta_x P \sim \mathrm{Leb}(\mathbb{R}^d)$, we can introduce the conditional density $p(y|x)\mathrm{d}y = P(x, \mathrm{d}y)$. By Bayes theorem and invariance of π, we can write $\pi(x)p(y|x) = \pi(y)q(x|y)$, where $q(x|y)\mathrm{d}x = P^\dagger(y, \mathrm{d}x)$. We then conclude by:

$$\mathbb{E}_\pi\left[v_t\left(\log v_t - \log Pv_t\right)\right] = \mathbb{E}_{\rho_t^S}\left[\log\frac{u_t^S}{\pi}\right] - \mathbb{E}_{\rho_t^S}\left[\frac{1}{\pi(x)}\int u_t^S(y)q(x|y)\mathrm{d}y\right] = \mathrm{KL}\left(\rho_t^S||\rho_t^S P^\dagger\right),$$

where we recognized the integral to be the expression of the PDF of $\rho_t^S P^\dagger$. ■

E.1. Proof of our bounds for SGLD

Before proving Theorem 12, we start by establishing the following more general result.

Proposition 20 *Let's assume that the prior Markov kernel P is representable in the sense of Definition 8, i.e., there exists a diffusion semigroup $(P_t)_{t\geq 0}$ with reversible ergodic distribution π and $t_0 > 0$ such that $P = P_{t_0}$. Assume further that $\nabla^2 V \succeq K I_d$ ($K > 0$) with the notations of Definition 8 and that Assumption 1 holds for $K = P$, then we have:*

$$\mathrm{KL}\left(\rho_T^S||\pi\right) \leq \frac{1}{q}\int_0^T e^{-\frac{T-t}{\tau_0}}\mathbb{E}_{x\sim\rho_t^S}\left[\mathrm{KL}\left(\delta_x P_S||\delta_x P\right)\right]\mathrm{d}t,$$

with the constants q and τ_0 given by:

$$q = \frac{1 - e^{-Kt_0}}{1 - e^{-2Kt_0}}, \quad \frac{1}{\tau_0} = 1 - e^{-Kt_0}.$$

Proof The first step of the proof is a refinement of the proof of Corollary 11. Let us denote by π the reversible (and invariant) distribution of $(P_t)_{t\geq 0}$ and by $\mathcal{L}$ its infinitesimal generator. As before, we denote $L = P - I$ (L is the generator of the Poissonized semigroup while $\mathcal{L}$ is the generator of the semigroup $(P_t)_{t\geq 0}$). Note that we have proven in the proofs of Theorem 4 and Corollary 11 that Assumption 1 implies $v_t \log P v_t \in L^1(\pi)$, which justifies the computations below.

As π is invariant under P, we can apply Corollary 6 to write:

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\mathbf{KL}\left(\rho_\tau^S || \pi\right) = \mathbb{E}_\pi\left[v_\tau (P_S - P)\log v_\tau\right] - \mathcal{E}_\pi(\log v_\tau, v_\tau),$$

where we used τ as the time variable to avoid later confusion with $(P_t)_{t\geq 0}$.

By Donsker-Varhadan formula, we have, for all $\tau > 0$, $S \in \mathcal{Z}^n$, $q \in (0, 1]$ and $x \in \mathbb{R}^d$:

$$P_S \log v_\tau(x) = \frac{1}{q} P_S \log\left(v_\tau^q\right)(x) \leq \frac{1}{q}\mathbf{KL}\left(\delta_x P_S || \delta_x P\right) + \frac{1}{q}\log\left(P(v_\tau^q)(x)\right). \tag{18}$$

Note that Hölder's inequality implies that $v_\tau^q \in L^1(\delta_x P)$. Consider a differentiable function $q : [t_0/2, t_0] \to (0, 1]$, which will be determined later. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a $\mathcal{C}^2$ positive function s.t. $\forall x \in \mathbb{R}^d,\ \log f \in L^1(\delta_x P) \cap L^1(\delta_x P_S)$ and $f \log f \in L^1(\delta_x P)$. We first assume that f is bounded from below and has bounded derivatives of order 0, 1, and 2. We introduce the quantity:

$$\alpha(t) := \frac{1}{q(t)} \log P_t\left(f^{q(t)}\right).$$

Let us denote $g_t := P_t\left(f^{q(t)}\right)$, by the chain rule we have:

$$\alpha'(t) = -\frac{q'(t)}{q(t)^2}\log g_t + \frac{1}{q(t)}\frac{\mathcal{L} g_t}{g_t} + \frac{q'(t)}{q(t)}\frac{P_t\left(f^{q(t)}\log(f)\right)}{g_t}.$$

Let us denote Γ the carré du champ operator, *i.e.*, in our case $\Gamma(\psi) := \|\nabla\psi\|^2$, see Section A. By the diffusion property (see (Chafai and Lehec, 2017, Section 5.4)), we have:

$$\begin{aligned}
\alpha'(t) &= -\frac{q'(t)}{q(t)^2}\log g_t + \frac{1}{q(t)}\left(\mathcal{L}\log g_t + \frac{\Gamma(g_t)}{g_t^2}\right) + \frac{q'(t)}{q(t)}\frac{P_t\left(f^{q(t)}\log(f)\right)}{g_t} \\
&= -\frac{q'(t)}{q(t)^2}\log g_t + \frac{1}{q(t)}\left(\mathcal{L}\log g_t + \frac{\Gamma(g_t)}{g_t^2}\right) + \frac{q'(t)}{q(t)^2}\frac{\mathrm{Ent}_{P_t}\left(f^{q(t)}\right) + g_t \log g_t}{g_t} \\
&= \frac{1}{q(t)}\left(\mathcal{L}\log g_t + \frac{\Gamma(g_t)}{g_t^2}\right) + \frac{q'(t)}{q(t)^2}\frac{\mathrm{Ent}_{P_t}\left(f^{q(t)}\right)}{g_t}.
\end{aligned}$$

Now we assume that $\forall t \geq 0,\ q'(t) \leq 0$ and we note that by (Chafai and Lehec, 2017, Lemma 5.6) $\nabla^2 V \succeq K I_d$ is equivalent to the semigroup $(P_t)_{t\geq 0}$ satisfying the $\mathrm{CD}(K, \infty)$ condition (see (Bakry

et al., 2014)). Thus, by the reverse local logarithmic Sobolev inequality (Bakry et al., 2014, Theorem 5.5.2 (v)), we have:

$$\alpha'(t) \leq \frac{1}{q(t)} \left(\mathcal{L} \log g_t + \frac{\Gamma(g_t)}{g_t^2} \right) + \frac{q'(t)}{q(t)^2} \frac{e^{2Kt}-1}{2K} \frac{\Gamma(g_t)}{g_t^2}.$$

Based on this inequality, we choose the function q on $t \in [t_0/2, t_0]$ to be (recall that $P = P_{t_0}$):

$$q(t) := \exp\left(- \int_{\frac{t_0}{2}}^{t} \frac{2K}{e^{2Ku}-1} \mathrm{d}u \right).$$

This leads to the following differential inequality, for all $t_0/2 \leq t \leq t_0$:

$$\alpha'(t) \leq \frac{1}{q(t)} \mathcal{L} \log(g_t) = \mathcal{L}\alpha(t).$$

We can now write for $t_0/2 \leq s \leq t_0$ that:

$$\frac{\mathrm{d}}{\mathrm{d}s} \left(P_{t_0-s} \alpha(s) \right) = -\mathcal{L} P_{t_0-s} \alpha(s) + P_{t_0-s} \alpha'(s) \leq -\mathcal{L} P_{t_0-s} \alpha(s) + P_{t_0-s} \mathcal{L} \alpha(s) = 0,$$

where we used that the semigroup $(P_t)_{t \geq 0}$ commutes with its infinitesimal generator $\mathcal{L}$. Therefore, the map $s \to P_{t_0-s}\alpha(s)$ is decreasing. In particular, by using $q(t_0/2) = 1$ and the interpolation condition $P = P_{t_0}$ we have:

$$\frac{1}{q(t_0)} \log P \left(f^{q(t_0)} \right) = \alpha(t_0) \leq P_{\frac{t_0}{2}} \alpha \left(\frac{t_0}{2} \right) = P_{\frac{t_0}{2}} \log P_{\frac{t_0}{2}} f. \tag{19}$$

By using similar classical approximation arguments as at the end of the proof of Theorem 9, we extend the above inequality to positive twice continuously differentiable functions $f : \mathbb{R}^d \to \mathbb{R}$ such that $\forall x \in \mathbb{R}^d,\ \log f \in L^1(\delta_x P) \cap L^1(\delta_x P_S)$ and $f \log f, fP\log(f) \in L^1(\pi)$. By reasoning as in the proof of Corollary 11, this also implies that $f \log P(f^{q(t_0)}) \in L^1(\pi)$. By Assumption 1, v_τ satisfies these conditions. Thus, by reversibility of π under the semigroup $(P_t)_{t>0}$, we have:

$$\frac{1}{q(t_0)} \mathbb{E}_\pi \left[v_\tau \log P \left(v_\tau^{q(t_0)} \right) \right] \leq \mathbb{E}_\pi \left[P_{\frac{t_0}{2}} v_\tau \log P_{\frac{t_0}{2}} v_\tau \right].$$

We now plug this into Equation (18) and get:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\tau} \mathrm{KL}\left(\rho_\tau^S || \pi \right) \leq \frac{1}{q(t_0)} \mathbb{E}_{\rho_t^S} \left[\mathrm{KL}\left(\delta_x P_S || \delta_x P \right) \right] + \mathbb{E}_\pi \left[P_{\frac{t_0}{2}} v_\tau \log P_{\frac{t_0}{2}} v_\tau - v_\tau P \log(v_\tau) \right] \\ + \mathbb{E}_\pi \left[v_\tau P \log(v_\tau) - v_\tau \log v_\tau \right], \end{aligned}$$

which by invariance of π under $P_{t_0/2}$ leads to:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\tau} \mathrm{KL}\left(\rho_\tau^S || \pi \right) &\leq \frac{1}{q(t_0)} \mathbb{E}_{\rho_t^S} \left[\mathrm{KL}\left(\delta_x P_S || \delta_x P \right) \right] + \mathbb{E}_\pi \left[P_{\frac{t_0}{2}} v_\tau \log P_{\frac{t_0}{2}} v_\tau - P_{\frac{t_0}{2}} \left(v_\tau \log(v_\tau) \right) \right] \\ &= \frac{1}{q(t_0)} \mathbb{E}_{\rho_t^S} \left[\mathrm{KL}\left(\delta_x P_S || \delta_x P \right) \right] - \mathbb{E}_\pi \left[\mathrm{Ent}_{P_{\frac{t_0}{2}}} \left(v_\tau \right) \right]. \end{aligned}$$

Finally, we note that $\mathbb{E}_\pi\left[\mathrm{Ent}_{P_{\frac{t_0}{2}}}(v_\tau)\right] = \bar{\mathcal{E}}_\pi(v_\tau)$, where $\bar{\mathcal{E}}_\pi$ has been introduced in the proof of Theorem 9. By the same proof we obtain:

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\mathrm{KL}\left(\rho_\tau^S||\pi\right) \leq \frac{1}{q(t_0)}\mathbb{E}_{\rho_t^S}\left[\mathrm{KL}\left(\delta_x P_S||\delta_x P\right)\right] - \left(1-e^{-Kt_0}\right)\mathrm{KL}\left(\rho_\tau^S||\pi\right).$$

We conclude by solving this differential inequality and using $\exp\left(\int_{\frac{t_0}{2}}^{t_0}\frac{-2K}{e^{2Ku}-1}\mathrm{d}u\right) = \frac{1-e^{-Kt_0}}{1-e^{-2Kt_0}}$. ■

We can now prove Theorem 12 as a corollary of the above proposition.

Theorem 12 (Poissonized SGLD) *Consider the Markov kernel P_S corresponding to SGLD with $\eta\lambda < 1$ and take P and π to be the Markov kernel and the invariant distribution of the recursion $X_{k+1} = (1-\lambda\eta)X_k + \sigma\mathcal{N}(0, I_d)$. Assume that Assumption 1 holds, then we have:*

$$\mathrm{KL}\left(\rho_T^S||\pi\right) \leq \frac{\eta^2(2-\lambda\eta)}{2\sigma^2}\int_0^T e^{-\lambda\eta(T-t)}\mathbb{E}_{x\sim\rho_t^S,U}\left[\|\widehat{g}_S(x,U)\|^2\right]\mathrm{d}t. \tag{9}$$

Proof We apply Proposition 20 by introducing the same SDE as in the proof of Corollary 10. The KL divergence $\mathrm{KL}\left(\delta_x P_S||\delta_x P\right)$ is bounded by classical arguments, see (Neu et al., 2021, Lemma 4). ■

E.2. Omitted proofs of Section 5.2

Before stating the proof of Proposition 13, let's recall the definition of Wasserstein distance.

Definition 21 (Wasserstein distance) *let μ and ν be two probability measures. We denote by $\Gamma(\mu,\nu)$ the set of all* couplings *between μ and ν. The Wasserstein's distance W_2 is then defined as:*

$$W_2(\mu,\nu) = \inf_{\gamma\in\Gamma(\mu,\nu)}\left\{\iint\|x-y\|^2\,\mathrm{d}\gamma(x,y)\right\}^{\frac{1}{2}}.$$

Proposition 13 *Assume that, for all $t\in[0,T]$, there exist constants $c_1, c_2 > 0$ such that v_t satisfies the linear growth condition $\forall x\in\mathbb{R}^d,\ \|\nabla\log v_t(x)\| \leq c_1\|x\| + c_2$, then we have:*

$$\Delta_{P,P_S}(v_t) \leq \mathbb{E}_{\rho_t^S}\left[W_2(\delta_x P,\delta_x P_S)^2\right]^{\frac{1}{2}}\left(\frac{c_1}{2}\|P\|_t + \frac{c_1}{2}\|P_S\|_t + c_2\right),$$

with $\|P\|_t^2 := \mathbb{E}_{x\sim\rho_t^S,w\sim\delta_x P}\left[\|w\|^2\right]$ (resp. P_S) and W_2 the Wasserstein distance (Section E.2).

The proof is inspired by (Polyanskiy and Wu, 2016, Proposition 1) and Raginsky et al. (2017).

Proof Let $\gamma_x \in \Gamma(\delta_x P,\delta_x P_S)$ that is optimal (see for instance (Alfonsi et al., 2016)) for the Wasserstein distance $W_2(\delta_x P,\delta_x P_S)$ and $(U,V)\sim\gamma_x$. We denote by $\rho_t^S\otimes\gamma_x$ the joint distribution of $x\sim\rho_t^S$ and $(U,V)\sim\gamma_x$.

By the linear growth condition, we have almost surely:

$$\begin{aligned}\log v_t(U) - \log v_t(V) &= \int_0^1\langle\nabla\log v_t(sU+(1-s)V), U-V\rangle\mathrm{d}s\\ &\leq \|U-V\|\left(\frac{c_1}{2}\|U\| + \frac{c_1}{2}\|V\| + c_2\right). \quad \text{(Cauchy-Schwarz's inequality)}\end{aligned}$$

By integrating over $\rho_t^S \otimes \gamma_x$ and using Cauchy-Schwarz and triangle inequalities we obtain:

$$\begin{aligned}(P_S - P)(\log v_t)(x) &= \mathbb{E}_{(U,V)\sim\gamma_x}\left[\log v_t(U) - \log v_t(V)\right] \\ &\leq \mathbb{E}_{\rho_t^S}\left[W_2(\delta_x P, \delta_x P_S)^2\right]^{\frac{1}{2}} \left\| \frac{c_1}{2}\|U\| + \frac{c_1}{2}\|V\| + c_2 \right\|_{L^2(\rho_t^S\otimes\gamma_x)} \\ &\leq \mathbb{E}_{\rho_t^S}\left[W_2(\delta_x P, \delta_x P_S)^2\right]^{\frac{1}{2}} \left(\frac{c_1}{2}\|P\|_t + \frac{c_1}{2}\|P_S\|_t + c_2\right).\end{aligned}$$

■

E.3. Proof of Theorem 14

Theorem 14 (Generalization bound for regularized SGD) *Assume that ℓ is s^2-subgaussian and Assumption 1 holds with π the invariant measure of the Markov process of Corollary 10 with $\gamma = \lambda\eta$ and $\sigma > 0$ a noise scale. We further assume that there exists $\beta \geq 0$ such that $0 \preceq \nabla^2 \log(v_t) \preceq \beta I_d$, for all $t \in [0, T]$. Then, we have, with probability at least $1-\zeta$ over $S \sim \mu_z^{\otimes n}$ that:*

$$\mathbb{E}_{\rho_T^S}\left[G_S\right] \leq \frac{2s}{\sqrt{n}}\left\{\int_0^T e^{-\lambda\eta(T-t)}\eta\mathbb{E}_{x\sim\rho_t^S,U}\left[Q\left(\|\widehat{g}_S(x,U)\|, \|x\|\right)\right]\mathrm{d}t + e^{-\lambda\eta T}K_0 + \log\frac{3}{\zeta}\right\}^{\frac{1}{2}},$$

where $Q(X,Y) := X\left(\beta Y + \left\|\nabla \log u_t^S(0)\right\|\right) + \eta\beta\left(X^2 + \lambda^2 Y^2\right)$ is a 2^{nd}-order bivariate polynomial, $K_0 := \mathrm{KL}\left(p_0||\pi\right)$ and U is the randomness of the batches in the stochastic gradient $\widehat{g}_S$.

Before providing the proof of this result, we introduce a few notations. We write the posterior and prior Markov operators in the following form:

$$P_S f(x) = \mathbb{E}_{U_1\sim\nu_1}\left[g_S(x,U_1)\right], \quad Pf(x) = \mathbb{E}_{U_2\sim\nu_2}\left[g_S(x,U_2)\right]$$

where g and g_S are called the stochastic gradient functions and the probability measures ν_1 and ν_2 represent the randomness of the stochastic gradients (recall that $S \in \mathcal{Z}^n$ denotes the dataset). To simplify the notations, we introduce an arbitrary coupling $\nu \in \Gamma(\nu_1, \nu_2)$ between ν_1 and ν_2 and, up to a slight abuse of notations, we write $U := (U_1, U_2)$ and:

$$P_S f(x) = \mathbb{E}_U\left[f(x - \eta g_S(x,U))\right] \quad Pf(x) = \mathbb{E}_U\left[f(x - \eta g(x,U))\right].$$

In the case of Theorem 14, we have $g_S(x,U_1) := \widehat{g}_S(x,U_1) + \lambda x$ in the notations of Example 1 (and there is no additional noise, *i.e.*, $\zeta_k = 0$), where $\widehat{g}_S(x,U)$ represents the unbiased stochastic gradient, and $g(x,U_2) = \lambda x + (\sigma/\eta)\mathcal{N}(0, I_d)$.

Equipped with these notations, we derive the proof of Theorem 14. These general notations show that our proof yields a more general result. The proof is based on a Taylor expansion technique of v_t. We note that different Taylor expansions have already been used for SGD (Dieuleveut et al., 2018).

Proof Consider the function $\Phi(x) = x\log(x)$. Let $S \in \mathcal{Z}^n$, we denote the expansion term as before by $\Delta_{P,P_S}(v_t) := \mathbb{E}_\pi\left[(P_S - P)(\Phi' \circ v_t)v_t\right]$. By Assumption 1, v_t is twice continuously differentiable, thus, by Taylor expansions around $\eta = 0$ of the functions $\eta \mapsto v_t(x - \eta g_S(x,U))$ and $\eta \mapsto v_t(x - \eta g(x,U))$, we can write:

$$\Delta_{P,P_S}(v_t) = -\eta\mathbb{E}_\pi\left[v_t\Phi''(v_t)\nabla v_t \cdot \Delta_S\right] + \eta^2\mathbb{E}_\pi\left[v_t(x)\int_0^1 (1-u)\left(H_x^S(u) - H_x(u)\right)\mathrm{d}u\right],$$

with Δ_S the expected gradient difference given by $\Delta_S(x) = \mathbb{E}_{U\sim\nu}\left[g_S(x,U) - g(x,U)\right]$. The quantities H_x^S and H_x correspond to the following Hessian "norms", given by:

$$\begin{cases} H_x^S(u) = \mathbb{E}_{U\sim\nu}\left[g_S(x,U)^T\nabla^2(\Phi'\circ v_t)(x - u\eta g_S(x,U))g_S(x,U)\right] \\ H_x(u) = \mathbb{E}_{U\sim\nu}\left[g(x,U)^T\nabla^2(\Phi'\circ v_t)(x - u\eta g(x,U))g(x,U)\right]. \end{cases}$$

In our case, because $\widehat{g}_S$ is an unbiased stochastic gradient, we simply have $\Delta_S = \nabla\widehat{\mathcal{R}}_S(x)$. By the assumptions on $\nabla^2\log v_t$, this simplifies into:

$$\Delta_{P,P_S}(v_t) \leq -\eta\mathbb{E}_{\rho_t^S}\left[\langle\nabla\log v_t(x), \nabla\widehat{\mathcal{R}}_S(x)\rangle\right] + \frac{\eta^2\beta}{2}\mathbb{E}_{\rho_t^S\otimes\nu}\left[\|g_S(x,U)\|^2\right].$$

By our assumptions on $\nabla^2\log v_t$, we see that $\nabla\log v_t$ is in particular β-Lispchitz-continuous. Moreover, we note that $\nabla\log v_t(0) = \nabla\log u_t^S(0)$, as $v_t = u_t^S/\pi$ and π is a centered Gaussian distribution. Therefore, we can write, by the Cauchy-Schwarz and triangle inequalities (we omit the randomness of the batch indices U in $\widehat{g}_S(x)$ to simplify the notations):

$$\begin{aligned} \Delta_{P,P_S}(v_t) &\leq -\eta\mathbb{E}_{\rho_t^S}\left[\langle\nabla\log v_t(x), \nabla\widehat{\mathcal{R}}_S(x)\rangle\right] + \frac{\eta^2\beta}{2}\mathbb{E}_{\rho_t^S\otimes\nu}\left[\|\widehat{g}_S(x) + \lambda x\|^2\right] \\ &\leq \eta\mathbb{E}_{\rho_t^S}\left[\left(\beta\|x\| + \left\|\nabla\log u_t^S(0)\right\|\right)\left\|\nabla\widehat{\mathcal{R}}_S(x)\right\|\right] + \eta^2\beta\mathbb{E}_{\rho_t^S\otimes\nu}\left[\|\widehat{g}_S(x)\|^2 + \lambda^2\|x\|^2\right] \\ &\leq \eta\mathbb{E}_{\rho_t^S\otimes\nu}\left[\left(\beta\|x\| + \left\|\nabla\log u_t^S(0)\right\|\right)\|\widehat{g}_S(x)\| + \eta^2\beta\left(\|\widehat{g}_S(x)\|^2 + \lambda^2\|x\|^2\right)\right], \end{aligned}$$

where the last line follows from Jensen's inequality.

Finally, by Corollary 10, the prior dynamics satisfies a modified LSI with constant $\gamma = \lambda\eta$. Therefore, we can directly conclude by applying Theorem 7.

■

Appendix F. Additional background on depoissonization

In this short subsection, complement the discussion of Section 2 on depoissonization, *i.e.*, the process of deducing asymptotic properties of a Markov chain from bounds on the Poissonized distribution. We present part of the analysis of Jacquet and Szpankowski (1998) on this matter. See Section 2 for additional references on this well-studied topic.

As in Section 2, we denote $(Y_t)_{t\geq 0}$ the Poissonized version of a discrete-time process $(X_k)_{k\in\mathbb{N}}$.

We first note that, for an integrable function f, by Fubini's theorem and independence we have:

$$\mathbb{E}\left[f(Y_t)\right] = \sum_{k=0}^{\infty}\mathbb{E}\left[f(X_k)\mathbb{1}\left\{N_t = k\right\}\right] = e^{-t}\sum_{k=0}^{+\infty}\mathbb{E}\left[f(X_k)\right]\frac{t^k}{k!}.$$

We call such an expression a "Poisson transform". Our theory typically provides bounds on quantities like $\mathbb{E}\left[f(Y_t)\right]$ (where f is the generalization error G_S). In this context, the goal of depoissonization would be to obtain guarantees on $\mathbb{E}\left[f(X_k)\right]$ from these bounds on $\mathbb{E}\left[f(Y_t)\right]$. To generalize the above formula, we consider a sequence $(g_n)_{n\in\mathbb{N}}$ and extend the Poisson transform to the whole complex plane as follows:

$$\forall z\in\mathbb{C},\ \tilde{G}(z) := e^{-z}\sum_{k=0}^{+\infty} g_k\frac{z^k}{k!}. \tag{20}$$

We make the following assumption on the Poisson transform. For technical background on complex analysis, we refer the reader to (Lang, 1999; Freitag and Rusam, 2005).

Assumption 2 *We assume that the series defining $\tilde{G}(z)$ converges normally for all $z \in \mathbb{C}$ and that $\tilde{G}$ is holomorphic in the entire complex plane, i.e., it is an entire function.*

First note that $(g_n)_{n\in\mathbb{N}}$ can be reconstructed if we have full knowledge of $\tilde{G}(z)$, at least on a countour around 0 in $\mathbb{C}$. Indeed, if C is a circle centered at 0, we have, by Cauchy's formula (Freitag and Rusam, 2005, Theorem 2.3.4):

$$g_n = \frac{n!}{2\pi i}\oint_C \frac{e^z \tilde{G}(z)}{z^{n+1}}\mathrm{d}z = \frac{n!}{n^n 2\pi}\int_{-\pi}^{\pi} \tilde{G}(ne^{it})\exp\left(ne^{it}\right)e^{-nit}\mathrm{d}t.$$

However, in most practical cases we have little information on the entire Poisson transform (*i.e.*, we only know it on the real line). Jacquet and Szpankowski (1998) provided general results to derive the asymptotics of the initial sequence from that of $\tilde{G}$, *i.e.*, to show that $\tilde{G}(k) \approx g_k$ when $k \to \infty$. The main difficulty is that the proof of such results requires asymptotics of the Poisson transform in regions that are bigger than the positive real line. This is formalized by the following definition.

Definition 22 (Linear cones) *Let $|\theta| < \pi/2$, the θ-linear cone is $\mathcal{S}_\theta := \{z \in \mathbb{C},\ |\arg(z)| \leq \theta\}$.*

Equipped with this definition, we can state a basic depoissonization. Note that Jacquet and Szpankowski (1998) also provide more general depoissonization results. For the sake of simplicity, we focus on the basic result.

Theorem 23 (Basic depoissonization lemma, (Jacquet and Szpankowski, 1998)) *Assume that Assumption 2 holds and that there exists $|\theta| < \pi/2$, $A, B, R > 0$, $\beta \in \mathbb{R}$ and $\alpha < 1$ such that the following assumptions hold:*

1. *For all $z \in \mathcal{S}_\theta$, we have $|z| > R \implies |\tilde{G}(z)| \leq B|z|^\beta$.*
2. *For all $z \in \mathbb{C}\backslash\mathcal{S}_\theta$, we have $|z| > R \implies |\tilde{G}(z)e^z| \leq Ae^{\alpha|z|}$.*

Then we have $g_k = \tilde{G}(k) + \mathcal{O}_{k\to\infty}\left(k^{\beta-1}\right)$.

For instance, if $(X_k^S)_{k\in\mathbb{N}}$ is the posterior dynamics (*i.e.*, learning algorithm) for some $S \in \mathcal{Z}^n$ as defined in Section 3, and we consider the associated Poisson transform:first version

$$\forall z \in \mathbb{C},\ \tilde{G}_S(z) := e^{-z}\sum_{k\in\mathbb{N}}^{\infty}\frac{z^k}{k!}\mathbb{E}\left[G_S(X_k^S)|S\right].$$

Then, if for all $S \in \mathcal{Z}^n$, $\tilde{G}_S$ satisfies the assumptions of Theorem 23 with a constant $\beta_S < 1$ and if $(Y_t^S)_{t\geq 0}$ denotes the Poissonized process (see Section 2), then we have the following approximation of the generalization error by its Poissonized counterpart:

$$\mathbb{E}\left[G_S(X_k^S)|S\right] = \mathbb{E}\left[G_S(Y_k^S)|S\right] + \mathcal{O}\left(\frac{1}{k^{1-\beta_S}}\right). \tag{21}$$

This result complements the depoissonization result obtained in Theorem 3.