Modified Log-Sobolev Inequalities, Mixing and Hypercontractivity

[Extended Abstract]

Sergey Bobkov*
Department of Mathematics
University of Minnesota, Minneapolis, MN
bobkov@math.umn.edu

Prasad Tetali†
School of Mathematics
and College of Computing
Georgia Tech, Atlanta, GA
tetali@math.gatech.edu

ABSTRACT

Motivated by (the rate of information loss or) the rate at which the entropy of an ergodic Markov chain relative to its stationary distribution decays to zero, we study modified versions of the standard logarithmic Sobolev inequality in the discrete setting of finite Markov chains and graphs. These inequalities turn out to be weaker than the standard log-Sobolev inequality, but stronger than the Poincare' (spectral gap) inequality. We also derive a hypercontractivity formulation equivalent to our main modified log-Sobolev inequality which might be of independent interest. Finally we show that, in contrast with the spectral gap, for bounded degree expander graphs various log-Sobolev-type constants go to zero with the size of the graph.

Categories and Subject Descriptors

G.3 [**Mathematics of Computing**]: Probability & Statistics—*Markov processes, Probabilistic Algorithms*

General Terms

Algorithms, Theory

Keywords

Spectral gap, Entropy decay, Sobolev Inequalities

1. INTRODUCTION

Let (M, P, π) denote an ergodic Markov chain with a finite state space M, transition probability matrix P and stationary distribution π. For $f, g : M \to \mathbf{R}$, let $\mathcal{E}(f, g)$ denote the Dirichlet form defined by

$$\mathcal{E}(f,g) = -E_\pi(fLg) = -\sum_{x\in M} f(x)Lg(x)\pi(x), \qquad (1.1)$$

where $-L = I - P$ is the so-called Laplacian matrix. Then the spectral gap of P or the smallest non-zero eigenvalue of $-L$ can be defined as the optimal positive constant in

$$\lambda_1 \mathrm{Var}_\pi f \le \mathcal{E}(f,f), \qquad (1.2)$$

over all $f : M \to \mathbf{R}$. As usual, $\mathrm{Var}_\pi f = E_\pi f^2 - (E_\pi f)^2$. Note that one arrives at such a functional (or variational) definition of the spectral gap in a natural way, when one considers the rate of decay of variance of the distribution of the chain with respect to the stationary distribution. More formally, working in the technically-easier continuous-time, let $\mu_t = \mu_0 P_t$ be the distribution of the chain at time t, for $t \ge 0$, where we use P_t to denote the semi-group generated by L: $e^{tL} = \sum_{n=0}^{\infty} \frac{t^n L^n}{n!}$. Let $f_t = \mu_t/\pi$ denote the density of μ with respect to π, i.e., $f_t(x) = \mu_t(x)/\pi(x)$, for all $x \in M$. Then it is a classical fact that

$$\frac{d}{dt}\,\mathrm{Var}_\pi(f_t) = -2\mathcal{E}(f_t, f_t), \qquad (1.3)$$

which motivates the above definition of λ_1. On the other hand, little attention seems to have been given (particularly in the context of finite Markov chains) to the following equally natural property: for all $t \ge 0$,

$$\frac{d}{dt}\,D(\mu_t\|\pi) = -\mathcal{E}(f_t, \log f_t), \qquad (1.4)$$

where $D(\mu\|\pi) = \sum_{x\in X} \mu(x)\log(\mu(x)/\pi(x))$ denotes (the informational divergence or) the relative entropy of μ with respect to π. Using the standard notation that $\mathrm{Ent}_\pi f = E_\pi(f\log f) - (E_\pi f)\log(E_\pi f)$, one is then motivated in studying the inequality,

$$\rho_0 \mathrm{Ent}_\pi f \le \frac{1}{2}\mathcal{E}(f, \log f), \qquad (1.5)$$

over all $f : M \to \mathbf{R}^+$, since one is then able to conclude (after observing that $\mathrm{Ent}_\pi f = D(\mu\|\pi)$, whenever $f = \mu/\pi$) that for all $t \ge 0$,

$$\frac{d}{dt}\,D(\mu_t\|\pi) \le -2\rho_0\, D(\mu_t\|\pi), \quad t > 0.$$

*Research supported in part by NSF Grant DMS-0103929

†Research supported in part by NSF Grant No. DMS-0100289; research done while visiting Microsoft Research

STOC'03, June 9–11, 2003, San Diego, California, USA.

If one would rather study convergence to stationarity using the more popular total variation norm: $\|\mu_t - \pi\|_{\rm TV} = \frac{1}{2}\sum_{x\in M} |\mu_t(x) - \pi(x)|$, a well-known inequality (see (2.5)) between the total variation norm and the relative entropy could lead the above discussion further to (see Corollary 2.6)): for every initial distribution μ_0 on M, for all $t \geq 0$,

$$\|\mu_t - \pi\|^2_{\rm TV} \leq 2 \log \frac{1}{\pi_*}\, e^{-2\rho_0 t}, \tag{1.6}$$

where $\pi_* = \min_{x\in M} \pi(x)$, thus recovering and in fact improving upon a similar bound (see Remark 2.5 below) employing the standard logarithmic Sobolev constant. In Section 4, we consider a further generalization of (1.3) and (1.4) using Sobolev-type inequalities, which interpolate between the modified log-Sobolev inequality and the Poincaré inequality.

Recall that the standard logarithmic Sobolev inequality is of the form

$$\rho {\rm Ent}_\pi f^2 \leq 2\mathcal{E}(f,f),$$

for all $f : M \to \mathbf{R}$. Also recall that it is shown in [12] that $\frac{1}{4\rho} \leq \tau_2 \leq \left(1 + \frac{1}{4}\log\log(1/\pi_*)\right)\frac{1}{2\rho}$, where $\tau_2 = \inf\{t > 0 : \sup_{\mu_0} E_\pi[|\mu_t/\pi - 1|^2]^{1/2} \leq 1/e\}$. Thus while ρ captures rather accurately the convergence to stationarity using $\sup_{\mu_0} E_\pi[|\mu_t/\pi - 1|^2]^{1/2}$, which in general is larger than $D(\mu_t\|\pi)$, it seems better to use ρ_0 when one wants to work with either the relative entropy or the total variation norm.

We observe that $\rho \leq \rho_0 \leq \lambda_1$ (see Proposition 3.6) and provide examples which show that either inequality could be tight up to universal constants. Some standard examples on which ρ_0 is in fact the order of λ_1, while $\rho = o(\rho_0)$, (thus providing tight bounds on convergence to stationarity using the total variation norm), include simple random walks on the complete graph, on the set of permutations using random transpositions, and a biased random walk on the n-cube.

It is natural to wonder how the relative entropy decays for random walks on expander graphs. Using a (yet another) modified log-Sobolev inequality of the gradient type (see the definition of ρ_1 in Section 3), we show that in fact both ρ and ρ_0 are of the order of $1/(\log|G|)$ for bounded degree expanders G (see Section 5), while by definition λ_1 is bounded away from zero.

On the computational side, we remark that ρ_0 of a graph could be computed efficiently up to arbitrary accuracy, in similarity to λ_1, while we can only argue that ρ can be estimated efficiently up to a factor of at most five.

An independent and interesting aspect (as explained in [3], [12] etc.) of the classical logarithmic Sobolev inequality is its equivalence to the so-called hypercontractivity. The often-used Bonami-Beckner hypercontractivity estimate for functions on $\{0,1\}^n$ is known (due to [15]) to be equivalent to ρ being 4 (in our notation). In this case it turns out that $\rho = \rho_0 = \rho_1 = \lambda_1 = 4$. In addition, for the general case (including the continuous setting) we derive a hypercontractivity characterization of ρ_0 (see Section 6).

2. CONVERGENCE TO STATIONARITY

Elaborating on the introduction, we start with a stochastic matrix P on a finite set M, and define a Markov process $\{X\}_{t\geq 0}$ in M with initial distribution, say, μ_0 and transition matrices

$$P_t = e^{-t(I-P)}, \quad t \geq 0,$$

with the generator $-L = I - P$. To study the asymptotic behavior of the probability distributions μ_t of random variables $X(t)$ for large time, we will assume that:

a) There is a stationary distribution π for P, i.e., $\pi P = P$.
b) $\pi(x) > 0$, for all $x \in M$.
c) For all $x, y \in M$ such that $x \neq y$, there exists $n \geq 1$ with $P^n(x,y) > 0$.

Standard fact is that $a)-c)$ imply that such a π is unique. More over, $b)$ also implies that any probability distribution μ on M is absolutely continuous with respect to π, and we are allowed to consider the corresponding density $\frac{d\mu}{d\pi}(x) = \frac{\mu(x)}{\pi(x)}$, $x \in M$. Thus, let $f_t(x) = \dfrac{\mu_t(x)}{\pi(x)}, \quad x \in M$, be the density of μ_t with respect to π at time $t \geq 0$. We wish to show that the measures μ_t approach π, or equivalently, that f_t's are getting close to 1 for large t. A proper quantitative statement may be done, for example, in terms of L^p-distance

$$\|f_t - 1\|^p_{L^p(\pi)} = \int |f_t - 1|^p \, d\pi, \quad 1 \leq p < +\infty,$$

which becomes the total variation norm $\|\mu_t - \pi\|_{\rm TV}$ in case $p = 1$. Another important measure of closeness is the so-called informational divergence, defined by

$$D(\mu_t\|\pi) = {\rm Ent}_\pi(f_t) = \int f_t \log f_t \, d\pi.$$

Recall that $\mu_t = \mu_0 P_t$. Let P^* denote the time-reversal of P given by $\pi(x)P^*(x,y) = \pi(y)P(y,x)$, for $x, y \in M$. Let $P^*_t = e^{tL^*}$, where $-L^* = I - P^*$. Then the following is a useful basic technical fact:

LEMMA 2.1. *For any μ_0 and all $t \geq 0$, we have $f_t = P^*_t f_0$. Consequently, for any $x \in M$,*

$$\frac{df_t(x)}{dt} = L^* f_t(x).$$

Now it is easy to show how the functional $\mathcal{E}(f,g)$ is connected with the L^2-distance $\|f_t - 1\|_{L^2(\pi)}$ and the informational divergence. Indeed, by Lemma 2.1, differentiating the function ${\rm Var}_\pi(f_t) = \int f_t^2 \, d\pi - 1$, we get

$$\begin{aligned} \frac{d}{dt}{\rm Var}_\pi(f_t) &= \int \frac{d}{dt} f_t^2 \, d\pi \\ &= 2\int f_t L^* f_t \, d\pi = 2\int L(f_t) f_t \, d\pi \\ &= -2\mathcal{E}(f_t, f_t). \end{aligned}$$

Assuming the Poincaré-type inequality (1.2) holds true with constant $\lambda_1 > 0$, we get from the above identity, the classical fact: $\dfrac{d}{dt}{\rm Var}_\pi(f_t) \leq -2\lambda_1 {\rm Var}_\pi(f_t)$. Integrating over t, we arrive at the standard estimate:

THEOREM 2.2. *For every initial distribution μ_0,*

$$ {\rm Var}_\pi(f_t) \leq {\rm Var}_\pi(f_0)\, e^{-2\lambda_1 t}, \quad t \geq 0. \tag{2.1}$$

Now, we may repeat a similar argument towards the study of the informational divergence:

LEMMA 2.3. *Under $a)-c)$, for any μ_0 and all $t>0$, the density f_t is strictly positive on M. Furthermore, the function $t \to D(\mu_t\|\pi)$ is differentiable on $(0,+\infty)$, and*

$$\frac{d}{dt}\, D(\mu_t\|\pi) = -\mathcal{E}(f_t, \log f_t), \quad t>0.$$

Proof. Writing Taylor's expansion : for all $x,y \in M$ and $t \geq 0$, $P_t^*(x,y) = e^{-t}\sum_{n=0}^{\infty} \frac{t^n}{n!}\,(P^*)^n(x,y)$. We also have, $\pi(x)(P^*)^n(x,y) = \pi(y)P^n(y,x)$, for $n \geq 1$. Hence, the assumptions $b)-c)$ imply that $P_t^*(x,y)>0$ whenever $t>0$. Since $f_t = P_t^* f_0$ and $\sum_x f_0(x) = 1$, this yields the first statement of the lemma.

Thus, in the range $t>0$, we are allowed to perform differentiation in accordance with Lemma 2.1 and the identity (1.1) for the Dirichlet form:

$$\begin{aligned}\frac{d}{dt}\, D(\mu_t\|\pi) &= \int \frac{d}{dt}\, f_t \log f_t \, d\pi \\ &= \int (\log f_t + 1)\, L^* f_t \, d\pi = \int L(\log f_t) f_t \, d\pi \\ &= -\mathcal{E}(f_t, \log f_t).\end{aligned}$$

□

Now, similarly to Theorem 2.2, we can start from the modified logarithmic Sobolev inequality (1.5) on M.

THEOREM 2.4. *For every initial distribution μ_0,*

$$D(\mu_t\|\pi) \leq D(\mu_0\|\pi)\, e^{-2\rho_0 t}, \quad t \geq 0. \tag{2.2}$$

Proof. By Lemma 2.3,

$$\frac{d}{dt}\, D(\mu_t\|\pi) \leq -2\rho_0\, D(\mu_t\|\pi). \quad t>0.$$

Integrating this inequality over t and since the right hand side is continuous at $t=0$, we arrive at the desired estimate (2.2).

In the general case, let us modify the kernel a little: for small $\varepsilon>0$, consider

$$P^{(\varepsilon)}(x,y) = (1-\varepsilon)P(x,y) + \varepsilon\pi(y), \quad x,y \in M.$$

Evidently, the new Markov kernel satisfies all the properties $a)-c)$ with the same invariant measure π. Hence, we get (2.2) for the corresponding measures $\mu_t^{(\varepsilon)}$:

$$D(\mu_t^{(\varepsilon)}\|\pi) \leq D(\mu_0^{(\varepsilon)}\|\pi)\, e^{-2\rho_0 t}, \quad t \geq 0.$$

It remains to let $\varepsilon \downarrow 0$. □

REMARK 2.5. *Note that we didn't make the assumption of reversibility (namely, the assumption that $\pi(x)P(x,y) = \pi(y)P(y,x)$, for all $x,y \in M$) of the Markov kernel in the above. In the next section, assuming reversibility we show that $\rho \leq \rho_0$, thus showing that the estimate (2.2) improves upon*

$$D(\mu_t\|\pi) \leq D(\mu_0\|\pi)\, e^{-2\rho\, t}, \quad t \geq 0. \tag{2.3}$$

The latter was obtained in [2] and [25], see discussion in [12]. Together with (2.3), Theorem 3.6 in [12] also involves nonreversible Markov kernels in which case ρ is replaced with $\rho/2$ (a result of L. Miclo [22]). The above proof of Lemma 2.3 which led to Theorem 2.4 is implicitly contained in [12] (or as paraphrased in the appendix of [13]). The important difference is that the usual log-Sobolev inequality (3.2) is taken as a starting point in [13] and in all the above-mentioned papers.

The estimates given in Theorems 2.2 and 2.4 are not comparable in general: each may have its own advantages. When $\rho_0 = \lambda_1$ or when these constants are of similar magnitude, the estimate (2.2) can be more useful than the estimate (2.1). First note, there is a general inequality $\int f\, d\pi\, \mathrm{Ent}_\pi f \leq \mathrm{Var}_\pi(f)$, holding true for any measurable function on an arbitrary probability space. Applying this to $f = f_t$, we get

$$D(\mu_t\|\pi) \leq \mathrm{Var}_\pi(f_t).$$

Hence, in the second theorem, a smaller distance (the informational divergence at time t) is estimated from above by a smaller quantity (the informational divergence at the initial time multiplied by an exponentially decreasing factor).

Another natural and typical objective is obtaining the rates of convergence in total variation norm $\|\mu_t - \pi\|_{\mathrm{TV}}$ uniformly over all possible μ_0. Then, in order to apply (2.1), one can use a trivial bound $\|\mu_t - \pi\|_{\mathrm{TV}}^2 \leq \mathrm{Var}_\pi(f_t)$. The right hand side of (2.1) is maximized when μ_0 is one of the Dirac measures δ_x which leads to

$$\|\mu_t - \pi\|_{\mathrm{TV}}^2 \leq \frac{1}{\pi_*}\, e^{-2\lambda_1 t}, \quad \text{where} \quad \pi_* = \min_{x \in M} \pi(x). \tag{2.4}$$

It is also possible to relate the total variation norm to the informational divergence, using the following well known inequality, see e.g. Lemma 12.6.1 in [9], [12] or [13]: for every probability measure μ on M,

$$\|\mu - \pi\|_{\mathrm{TV}}^2 \leq 2\, D(\mu\|\pi). \tag{2.5}$$

With estimate (2.2) this leads to a certain refinement of (2.4) (when ρ_0 is approximately λ_1):

COROLLARY 2.6. *For every initial distribution μ_0 on M, for all $t \geq 0$,*

$$\|\mu_t - \pi\|_{\mathrm{TV}}^2 \leq 2 \log \frac{1}{\pi_*}\, e^{-2\rho_0 t}. \tag{2.6}$$

A general 2-state chain can be used to show that in (2.6), the dependence on t can be sharp; (once again, ρ_0 and λ_1 are of the same order in such an example.)

EXAMPLE 2.7. **I**. *(slices of the n-cube) A fundamental example is a slice $\Omega(n,k)$ of the discrete cube: the vertices being the k-subsets of an n-set. Two subsets are adjacent if and only if they can be obtained from each other by a single swap of a pair of elements. (Note that this is also the so-called uniform matroid.) There is a natural reversible Markov kernel associated with this graph, which assigns $P(x,y) = 1/[k(n-k)]$, whenever x and y are neighbors. In the full version of our paper, using an elementary proof we show that $(n+2)/[4k(n-k)] \leq \rho_0 \leq n/[k(n-k)] = \lambda_1$. (The value of λ_1 is due to [11].)*

II. *(random transpositions) We also prove in the full version that ρ_0 of the chain on permutations using (uniform) random transpositions satisfies: $1/[2(n-1)] \leq \rho_0 \leq 2/(n-1) = \lambda_1$. This in particular implies that the mixing time in total variation is at most $O(n \log n)$, which is the tight,*

whereas only an $O(n\log^2 n)$ bound follows from ρ, since $\rho = \Theta(1/(n\log n))$.

The proof technique uses the chain rule for conditional relative entropy and convexity of $R(a,b) = (a-b)(\log a - \log b)$, for $a, b > 0$, in getting a recurrence for ρ_0 of S_n as a function of n. We have recently found out that ρ_0 has been introduced and estimated in the context of studying precisely the above two examples by Gao and Quastel in a recent paper [14]. The approach in [14] is a bit different, namely, the martingale approach (in the spirit of [20] who estimated the standard log-Sobolev constant for the same examples), while ours is a direct inductive argument. Our bounds are asymptotically equivalent. At the end of Section 4, we rederive the bounds in the above examples using an interpolating Sobolev-type inequality.

It might also be worth mentioning that the bounds (2.4)-(2.6) can be sharpened by virtue of Theorem 2.2 under mild symmetry assumptions of the initial density f_0 about its mean value $\int f_0\, d\pi = 1$. In particular, we have:

COROLLARY 2.8. *For every initial distribution μ_0 such that $\int (f_0 - 1)^3\, d\pi = 0$,*

$$\mathrm{Var}_\pi(f_t) \leq 2\log\frac{1}{\pi_*}\, e^{-2\lambda_1 t}, \quad t \geq 0. \tag{2.7}$$

Proof: See the journal version of the paper.

3. POINCARÉ AND LOG-SOBOLEV IN ABSTRACT SETTINGS

In this section we make a systematic study of various logarithmic Sobolev inequalities and the Poincaré inequality in discrete settings. We begin with a basic definition of a Dirichlet form. Let (M,μ) be a probability space, and let $\mathcal{A}$ be a linear space of bounded measurable functions on M. Further assumptions on $\mathcal{A}$ are:

Axiom 1. If $f, g \in \mathcal{A}$, then $fg \in \mathcal{A}$ (that is, $\mathcal{A}$ is an algebra).

Axiom 2. If $f \in \mathcal{A}$, then $e^f \in \mathcal{A}$.

DEFINITION 3.1. *Any bilinear form $\mathcal{E} : \mathcal{A}\times\mathcal{A} \to \mathbf{R}$ will be called a Dirichlet form.*

Although the definition of Dirichlet forms has nothing to do with the measure μ, it turns out that many standard examples are constructed through a measure. In what follows we will be primarily interested in a discrete setting of finite undirected graphs or finite Markov chains. How ever, since traditionally these functional inequalities have been studied in a continuous setting, we also briefly mention such a setting.

EXAMPLE 3.2. *(a continuous setting). Let M be an open subset of $\mathbf{R}^d$, and let $\mathcal{A}$ be the family of all smooth, compactly supported functions on M. Put*

$$\mathcal{E}(f,g) = \int_M \langle \nabla f(x), \nabla g(x)\rangle\, d\mu(x),$$

where $\langle\cdot,\cdot\rangle$ is a canonical scalar product in $\mathbf{R}^n$, and where $\nabla f(x) = (\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_1})$ denotes the usual gradient of f at the point $x \in M$. This gradient is local in the sense that $\nabla u(f) = u'(f)\nabla f$, for any smooth u such that $u(f) \in \mathcal{A}$.

The example can be generalized by considering for M an arbitrary Riemannian manifold of dimension d. If M is compact, one typically takes for μ the normalized Lebesgue measure on M.

EXAMPLE 3.3. *(a graph setting). Let $G = (M, \mathcal{M})$ be a finite, connected, undirected graph with vertex set M and edge set $\mathcal{M}$. Let $\mu : M \to [0,1]$ be an arbitrary probability measure on the vertices. Given a function f on M, one can define the gradient $\nabla f(x)$ at each vertex $x \in M$ as the vector $\{f(x) - f(y)\}_{y\sim x}$ of the length $d(x)$, the degree of x. Hence, the corresponding Dirichlet form becomes*

$$\mathcal{E}(f,g) = \sum_x \sum_{y\sim x} (f(x) - f(y))(g(x) - g(y))\, \mu(x).$$

Here $\mathcal{A}$ represents the space of all functions on M.

EXAMPLE 3.4. *(an abstract discrete setting and reversible Markov kernels). Again, let (M,μ) be a finite probability space, and let $P : M\times M \to [0,+\infty)$ be a non-negative function, called a kernel in the sequel. For all functions f, g on M, one may define the associated Dirichlet form by*

$$\mathcal{E}(f,g) = \frac{1}{2}\int \sum_{y\in M} (f(x) - f(y))\,(g(x) - g(y))\, P(x,y)\, d\mu(x).$$

It corresponds to the gradient operator $\nabla f(x) = \left\{\frac{1}{\sqrt{2}}\,(f(x) - f(y))\sqrt{P(x,y)}\right\}_{y\in M}$, so again the gradient formula as in Example 3.2 works well. If P is a reversible Markov kernel, then the above definition is also equivalent to the general definition given by (1.1), and we have the additional property that $\mathcal{E}(f,g) = \mathcal{E}(g,f)$. It turns out that the formula suggested by Example 3.2 is particularly suited to study reversible kernels, due to the apparent symmetry, while (1.1) is more general.

In the rest of this section, whenever we assume that P is a Markov kernel, we also assume implicitly that P is in fact reversible.

Consider a probability space (M,μ) and a Dirichlet form $\mathcal{E} : \mathcal{A}\times\mathcal{A} \to \mathbf{R}$, as above. Then we can introduce Poincaré-type (or spectral gap) and logarithmic Sobolev inequalities as

$$\lambda_1 \mathrm{Var}(f) \leq \mathcal{E}(f,f), \quad f \in \mathcal{A}, \tag{3.1}$$

$$\rho\, \mathrm{Ent}(f^2) \leq 2\,\mathcal{E}(f,f), \quad f \in \mathcal{A}. \tag{3.2}$$

As mentioned already in the introduction, by a modified logarithmic Sobolev inequality (of the Dirichlet type), we mean an inequality of the form

$$\rho_0 \mathrm{Ent}(e^f) \leq \frac{1}{2}\,\mathcal{E}(e^f, f), \quad f \in \mathcal{A}. \tag{3.3}$$

If the Dirichlet form comes through a gradient like in all the previous examples, one may also consider modified logarithmic Sobolev inequalities of the gradient type. The most popular versions which appeared in connection with the concentration of measure phenomenon are (see [19]):

$$\rho_1 \mathrm{Ent}(e^f) \leq \frac{1}{2}\int |\nabla f|^2\, e^f\, d\mu, \quad f \in \mathcal{A}, \tag{3.4}$$

$$\rho_2 \mathrm{Ent}(e^f) \leq \frac{1}{2}\int |\nabla e^f|^2\, e^{-f}\, d\mu, \quad f \in \mathcal{A}. \tag{3.5}$$

To be more precise, here one assumes that any point $x \in M$ is assigned with a linear operator $\mathcal{A} \ni f \to \nabla f(x) \in \mathbf{R}^{d(x)}$ such that the functions of the form $x \to \langle \nabla f(x), \nabla g(x) \rangle$ are μ-integrable and bounded, whenever $f, g \in \mathcal{A}$.

Formally replacing f with $\log f$, the inequality (3.5) takes a more familiar form

$$\rho_2 \mathrm{Ent}(f) \ \le \ \frac{1}{2} \int \frac{|\nabla f|^2}{f} \, d\mu, \quad f \in \mathcal{A}, \ f \ : \ \text{positive}. \ (3.6)$$

More precisely, we obtain (3.5) from the last inequality (3.6) by applying it to e^f. At this step the axiom 2 is used. For the converse implication, one needs a different assumption, that $\log f \in A$ as long as f belongs to $\mathcal{A}$ and is positive. Thus, in all the examples we considered before, the inequalities (3.5) and (3.6) are equivalent, but we prefer the first, exponential form in order to keep maximal generality and to save more analogs between (3.5) and the other exponential form (3.4).

If the gradient is local like in Example 3.2, all the log-Sobolev inequalities (3.2), (3.3), (3.4) and (3.5) are equivalent to each other, and moreover $\rho = \rho_0 = \rho_1 = \rho_2$ for optimal values. As for the general case, first we show that, under reasonable assumptions, the spectral gap inequality is weaker than any of these inequalities.

Proposition 3.5. *Assume that*

1) *the function $g(x) = 1$ belongs to $\mathcal{A}$ and $\mathcal{E}(f, 1) = \mathcal{E}(1, f) = 0$, for all $f \in \mathcal{A}$;*

2) *for all $f, g \in \mathcal{A}$ and for any uniformly bounded sequence f_n converging to f (μ-almost everywhere), we have $\mathcal{E}(f_n, g) \to \mathcal{E}(f, g)$, as $n \to \infty$.*

Then, for the optimal constants in (3.1) – (3.5), *we have* $\max\{\rho, \rho_0, \rho_1, \rho_2\} \le \lambda_1$.

Proof. To show $\rho \le \lambda_1$, note that, for every c real, $\mathcal{E}(f + c, f + c) = \mathcal{E}(f, f)$. Since $\mathrm{Ent}((f + c)^2) \to 2\,\mathrm{Var}(f)$, as $c \to \infty$, the application of (3.2) to functions of the form $f + c$ yields (3.1) with ρ in the place of λ_1. To prove $\rho_0, \rho_1, \rho_2 \le \lambda_1$, apply the inequalities (3.3)-(3.5) to functions $\frac{1}{n} f$ with $n \to \infty$.

Note that the assumptions 1) and 2) are not needed for deriving $\rho_1, \rho_2 \le \lambda_1$. The assumption 2) is automatically fulfilled as long as there exists a linear operator L associated with the Dirichlet form $\mathcal{E}$. In particular, this is true for an abstract discrete setting, which might be a well-known fact, and we omit the proof from this version.

Now let us specialize the log-Sobolev inequalities to discrete settings where they may differ considerably in terms of the magnitudes of ρ, ρ_0, ρ_1 and ρ_2. (The fact that $\rho \le \rho_1$ for reversible Markov kernels was observed in [17], wherein ρ_1 was called τ_2 (see comments after Lemma 1 in [17].)

Proposition 3.6. *In the reversible Markov kernel setting, for the optimal constants in* (3.1) – (3.5), *we have*

$$0 \le \rho \le \rho_0 \le \rho_1 \le \rho_2 \le \lambda_1.$$

Proof of Proposition 3.6. Let (M, μ) be a finite probability space with a reversible Markov kernel P. First we show that the logarithmic Sobolev inequality (3.2) implies the modified logarithmic Sobolev inequality (3.3) with $\rho_0 = \rho$. Thus, fix a function f on M. Starting from (3.2), apply it to the function $e^{f/2}$ to get

$$\rho \, \mathrm{Ent}(e^f) \le 2 \, \mathcal{E}(e^{f/2}, e^{f/2}).$$

Hence, in order to derive (3.3) with the same constant on the left, it suffices to show that

$$\mathcal{E}(e^{f/2}, e^{f/2}) \le \frac{1}{4} \, \mathcal{E}(e^f, f).$$

This estimate is actually observed in [12]. According to the definition (in Example 3.4) of the discrete Dirichlet form, we need to check that

$$\left(e^{f(x)/2} - e^{f(y)/2}\right)^2 \le \frac{1}{4} \left(e^{f(x)} - e^{f(y)}\right) (f(x) - f(y)),$$

for all $x, y \in M$. Putting $a = e^{f(x)/2}$, $b = e^{f(y)/2}$, we are reduced to the inequality $(a - b)^2 \le \frac{1}{2}(a^2 - b^2) \log \frac{a}{b}$ in the range $a, b > 0$, which can easily be verified to be true.

Now, in view of Proposition 3.5, we need only to show that (3.3) $\Rightarrow$ (3.4) $\Rightarrow$ (3.5) with $\rho_2 = \rho_1 = \rho_0$. Clearly, it suffices to compare the right hand sides in these inequalities and to see that, for every f on M,

$$\mathcal{E}(e^f, f) \le \int |\nabla f|^2 e^f \, d\mu \le \int |\nabla e^f|^2 \, e^{-f} \, d\mu. \qquad (3.7)$$

Since $|\nabla f(x)|^2 = \frac{1}{2} \sum_{y \in M} (f(x) - f(y))^2 P(x, y)$, we have

$$\begin{aligned} \int |\nabla f|^2 e^f \, d\mu &= \frac{1}{2} \sum_{x, y \in M} (f(x) - f(y))^2 e^{f(x)} P(x, y) \, \mu(x) \\ &= \frac{1}{2} \sum_{x, y \in M} (f(x) - f(y))^2 e^{f(y)} P(x, y) \, \mu(x), \end{aligned}$$

by reversibility. So

$$\int |\nabla f|^2 e^f \, d\mu$$

$$= \frac{1}{2} \sum_{x, y \in M} (f(x) - f(y))^2 \, \frac{e^{f(x)} + e^{f(y)}}{2} \, P(x, y) \, \mu(x).$$

Similarly, $\displaystyle\int |\nabla e^f|^2 \, e^{-f} \, d\mu$

$$= \frac{1}{2} \sum_{x, y \in M} (e^{f(x)} - e^{f(y)})^2 \, \frac{e^{-f(x)} + e^{-f(y)}}{2} \, P(x, y) \, \mu(x).$$

On the other hand, $\mathcal{E}(e^f, f)$

$$\begin{aligned} &= \int \left\langle \nabla e^f, \nabla f \right\rangle \, d\mu \\ &= \frac{1}{2} \sum_{x, y \in M} (f(x) - f(y)) \left(e^{f(x)} - e^{f(y)}\right) P(x, y) \, \mu(x). \end{aligned}$$

To establish (3.7), it suffices to compare the corresponding terms in these three representations. Thus, put $a = f(x)$, $b = f(y)$ for fixed $x, y \in M$: we need to show that

$$(a - b) \left(e^a - e^b\right) \le (a - b)^2 \, \frac{e^a + e^b}{2} \le \left(e^a - e^b\right)^2 \, \frac{e^{-a} + e^{-b}}{2}.$$

Since all the three sides are symmetric with respect to (a, b), we may assume $a \ge b$. Putting $a = b + h$, we are reduced to

$$h(e^h - 1) \le h^2 \, \frac{e^h + 1}{2} \le (e^h - 1)^2 \, \frac{e^{-h} + 1}{2}, \quad h \ge 0. \ (3.8)$$

Write the first inequality as $e^h - 1 \le h \, \frac{e^h + 1}{2}$. It turns into an equality at the point $h = 0$, while after differentiation it becomes $e^h \le \frac{1}{2} + \frac{h + 1}{2} e^h$. Again, there is equality at

$h = 0$, and differentiating it, we arrive at $e^h \le \frac{h+2}{2}\, e^h$ which is evidently true. This proves the first inequality in (3.8).

The second inequality is simplified as $h^2 e^h \le (e^h - 1)^2 \Longleftrightarrow h e^{h/2} \le e^h - 1 \Longleftrightarrow \frac{h}{2} \le \mathrm{sh}(\frac{h}{2})$. It readily holds, as well and thus Proposition 3.6 is proved. □

Note that, the normalizing property ($\sum_y P(x,y) = 1$, for all $x \in M$) was not used in the proof of Proposition 3.6. More over, the proof holds good for the graph setting with μ being uniform on the set of vertices.

EXAMPLE 3.7. *(symmetric discrete cube). Let $M = \{0,1\}^n$ be the discrete cube. For $x \in M$, if y is the neighbor of x obtained by flipping coordinate i, then we write $y = s_i(x)$. Then the canonical Dirichlet form on M is defined by*

$$\mathcal{E}(f,g) = \int \sum_{i=1}^{n} \left(f(x) - f(s_i(x))\right)\left(g(x) - g(s_i(x))\right)\, d\mu(x), \tag{3.9}$$

where the measure μ is uniform. In this case,

$$\rho = \rho_0 = \rho_1 = \rho_2 = \lambda_1 = 4. \tag{3.10}$$

(Formally we are not in a Markov kernel setting. However, one may simply multiply the Dirichlet form by $\frac{1}{2n}$ to get the corresponding constants.) That $\rho = 4$ is due to Gross[15]; that $\lambda_1 = 4$ is trivial. Hence $\rho = \lambda_1$ and the remaining equalities in (3.10) follow immediately from Proposition 3.6.

EXAMPLE 3.8. *(non-symmetric discrete cube). Now, for $p \in (0,1)$, equip M with the product measure $\mu = \mu_p^n$ with marginal μ_p assigning mass p to 1 and mass $q = 1 - p$ to 0. In this case for the Dirichlet form (3.9)*

$$\lambda_1 = \frac{1}{pq}, \qquad \rho = \frac{1}{pq}\ \frac{2\,(p-q)}{\log p - \log q}. \tag{3.11}$$

The first equality is trivial; the second one was obtained in [12] and mentioned in [16] without proof. The constant ρ_2 was studied in [6] where it is proved that

$$\frac{1}{2pq} \le \rho_2 \le \frac{1}{pq}. \tag{3.12}$$

Hence, ρ_2 is of order λ_1. As for the remaining log-Sobolev constants, we have:

PROPOSITION 3.9. *For the discrete cube $M = \{0,1\}^n$ with the product measure $\mu = \mu_p^n$, $0 < p < 1$,*

$$\frac{1}{2pq} \le \rho_0 \le \lambda_1 = \frac{1}{pq}, \qquad \rho_1 \le \frac{2\,(\log p - \log q)}{p - q}. \tag{3.13}$$

Proof. See the journal version of the paper.

Note that in huge contrast with Proposition 3.6, as $pq \to 0$, $\rho_1 \ll \rho \ll \rho_0 \approx \rho_2 \approx \lambda_1$, (although the best value of ρ_1 is not known). This pathological situation concerns only the modified log-Sobolev inequality (3.4) of gradient type. It may be explained with the fact that the gradient is not defined via the Dirichlet form (in contrast with (3.1), (3.2) and (3.3)) and essentially depends on the kernel itself. Indeed, already in dimension one, for any $f : \{0,1\} \to \mathbf{R}$, we have

$$\int |\nabla f|^2 e^f\, d\mu = (f(1) - f(0))^2 \left(p e^{f(1)} + q e^{f(0)}\right). \tag{3.14}$$

Note, if $p \neq q$, the right hand side is not invariant under replacement $f(1) \leftrightarrow f(0)$. On the other hand, in accordance with definition (3.9) in dimension one, we have

$$\mathcal{E}(e^f, f) = (f(1) - f(0))\left(e^{f(1)} - e^{f(0)}\right), \tag{3.15}$$

which is invariant (and does not depend on p, at all).

REMARK 3.10. *Note that the eigenvalue interpretation of λ_1 tells us that there is in fact a function (namely an eigenfunction) which achieves the optimal value λ_1. The same is not necessarily true of ρ and ρ_0 (e.g., as in the symmetric two-point case). How ever, we prove (in the journal version) that if the* inf *in the definition of ρ and ρ_0 is not achieved, then in fact $\rho = \rho_0 = \lambda_1$!*

REMARK 3.11. *All the above constants $(\rho, \rho_0, \rho_1, \rho_2, \lambda_1)$ remain invariant under taking the Cartesian product of a graph with itself. The "standard" argument (e.g., as in [17]) using the tensoring property of entropy and variance works.*

EXAMPLE 3.12. *(the complete graph). Let $(M, \mathcal{M})$ be the complete graph on a non-empty finite set M. Moreover, assume M is equipped with a probability measure μ such that $\mu_* = \min_{x \in M} \mu(x) > 0$, and consider the function $P(x,y) = \mu(y)$. Then, (P, μ) is a reversible Markov kernel, and in accordance with the Markov kernel setting, the Dirichlet form is given by*

$$\mathcal{E}(f,g) = \mathrm{cov}_\mu(f,g). \tag{3.16}$$

In particular, for $M = \{0,1\}$ with $\mu = \mu_p$, this Dirichlet form is $2pq$ times the Dirichlet form (3.15). Since the inequalities (3.1)–(3.3) are defined through the Dirichlet form, we can apply (3.11), (3.13) and then Proposition 3.6 to get:

$$\frac{1}{2} \le \rho_0 \le \rho_1 \le \rho_2 \le \lambda_1 = 1, \tag{3.17}$$

and

$$\rho = \frac{2\,(p-q)}{\log p - \log q}. \tag{3.18}$$

Thus, in contrast with Example 3.8, the optimal constants in all modified log-Sobolev inequalities are of order λ_1. Actually, the set of inequalities (3.17) remains to hold for an arbitrary complete graph M with the remark that, for a single point set M, all the optimal constants are equal to $+\infty$. Indeed, by Jensen's inequality and by (3.16), $\mathrm{Ent}(e^f) \le \mathrm{cov}(f, e^f) = \mathcal{E}(f, e^f)$, so $\rho_0 \ge \frac{1}{2}$. On the other hand, $\lambda_1 = 1$, and it remains to apply Proposition 3.6. As for the constant ρ, every complete graph M satisfies (3.18) with $p = \mu_*$, $q = 1 - \mu_*$. This is shown in [12] on the basis of the two point case (3.11).

4. BETWEEN MODIFIED LOG-SOBOLEV AND POINCARÉ

For reversible kernels, both inequalities (2.1) and (2.2) can be united by a more general scheme under a certain stronger hypothesis. Namely, given (M, P, π) with P being a reversible kernel, for a number $p \in (1, 2]$, one may start with the Sobolev-type inequality

$$\alpha(p)\,\left[\, \|f\|_p^p - \|f\|_1^p \,\right] \ \le\ \frac{p}{2}\,\mathcal{E}(f, f^{p-1}), \tag{4.1}$$

where f is an arbitrary positive function on M, and $\|f\|_p^p = \int f^p \, d\pi$, $\|f\|_1 = \int f \, d\pi$.

If $p = 2$, we are reduced to the Poincaré-type inequality (3.1), so the optimal constant $\alpha(2)$ is just the spectral gap λ_1. For $1 < p < 2$, applying (4.1) to functions of the form $1 + \varepsilon f$ and letting $\varepsilon \to 0$, we obtain the relation

$$\alpha(p) \le \lambda_1.$$

On the other hand, dividing both sides of (4.1) by $p-1$ and letting $p \downarrow 1$, we get in the limit the modified logarithmic Sobolev inequality (3.3), so $\alpha(1) = \rho_0$.

The proofs of Theorems 2.2 and 2.4 are readily extended to the more general statement:

THEOREM 4.1. *Under the hypothesis* (4.1) *with* $p \in (1, 2]$, *for every initial distribution* μ_0 *on* M,

$$\|f_t\|_p^p - 1 \le \left[\|f_0\|_p^p - 1\right] e^{-2\alpha(p)\, t}, \quad t \ge 0. \tag{4.2}$$

In the continuous setting with Dirichlet form $\mathcal{E}(f, g) = \int \langle \nabla f, \nabla g \rangle \, d\pi$, the inequality (4.1) may be rewritten equivalently by replacing p with $2/q$ and putting $f = g^q$. It then takes the form

$$\alpha(2/q) \left[\|g\|_2^2 - \|g\|_q^2 \right] \le (2-q)\, \mathcal{E}(g, g), \quad 1 \le q < 2. \tag{4.3}$$

This inequality was introduced in 1989 by W. Beckner [4] as a kind of sharp interpolation between Poincaré and logarithmic Sobolev inequality: it was established for the canonical Gaussian measure with optimal constants $\alpha(2/q) = 1$ thus generalizing the famous Gross' theorem ((4.3) was also proved there for uniform distributions on Euclidean spheres). Recently, a similar inequality was derived for product measures in $\mathbf{R}^n$ with marginal densities $c_r e^{-|x|^r}$, $1 \le r \le 2$, by R. Latala and K. Oleszkiewicz, cf. [18].

Let us note that, while for $q = 1$ the inequality (4.3) represents the spectral gap, the limiting case $q = 2$ reduces to the usual logarithmic Sobolev inequality (3.2), where the optimal constant may be much smaller than the one in (3.3). Therefore, (4.1) has a correct form to fit the features of the modified log-Sobolev inequality in the discrete setting. The essential difference between (4.1) and (4.3) already appears for complete graphs as we can see from the following:

PROPOSITION 4.2. *For every complete graph* M *on at least two vertices, equipped with an arbitrary probability measure* π, *for every* $p \in (1, 2]$, *we have*

$$\frac{p}{2} \le \alpha(p) \le 1.$$

Proof. The right hand side inequality is immediate since $\lambda_1 = 1$. Recalling that $\mathcal{E}(f, g) = \mathrm{cov}_\pi(f, g)$, the left hand side inequality is just

$$\|f\|_p^p - \|f\|_1^p \le \mathrm{cov}_\pi(f, f^{p-1}) = \|f\|_p^p - \|f\|_1 \|f\|_{p-1}^{p-1},$$

that is, $\|f\|_{p-1} \le \|f\|_1$. The latter holds due to $p - 1 \le 1$. □

EXAMPLE 4.3. *(product spaces) Let* $M = G^n$ *be the Cartesian product of* n *copies of* G, *with product probability measure* μ^n, *where* μ *is arbitrary on the vertices of* G, *and let* $p \in (1, 2]$. *Then we have* $\alpha(p)[G^n] = \alpha(p)[G]$, *for all* $n \ge 1$, *using the tensoring property of the functional* $\mathcal{L}(f) = \|f\|_p^p - \|f\|_1^p$.

In order to study the interpolating inequality (8.1) for non-product graphs, the following is a useful technical lemma; the proof can be found in the journal version.

LEMMA 4.4. *For any* $p \in (1, 2]$, *the function*

$$R(a, b) = (a - b)(a^{p-1} - b^{p-1}), \quad a, b > 0.$$

is convex in the positive octant.

As an illustration, consider the graph $M = \Omega(n, k)$ of slices of the discrete cube. Recall that the statement about the modified log-Sobolev inequality,

$$\mathrm{Ent}_\pi(f) \le \frac{1}{n+2}\, \mathcal{E}(f, \log f), \tag{4.4}$$

was mentioned (in Example 2.7) for the Dirichlet form corresponding to the graph setting, namely,

$$\mathcal{E}(f, g) = \int \sum_{y \sim x} (f(x) - f(y))(g(x) - g(y)) \, d\pi(x), \tag{4.5}$$

where π is uniform probability measure on M. Making use of Proposition 4.2 and the convexity of $R(a, b)$ above, leads to the following generalization of (4.4).

PROPOSITION 4.5. *Let* $1 \le k \le n - 1$ *be integer, and* $p \in (1, 2]$. *For every positive function* f *on* $\Omega(n, k)$, *with repsect to the uniform probability measure*

$$\|f\|_p^p - \|f\|_1^p \le \frac{1}{n+2}\, \mathcal{E}(f, f^{p-1}), \tag{4.6}$$

Equivalently, in the Markov kernel setting,

$$\frac{p\,(n+2)}{4k(n-k)} \le \alpha(p) \le \lambda_1 = \frac{n}{k(n-k)}.$$

The constant on the right, $\frac{1}{n+2}$, is of correct order uniformly over all admissible triples (n, k, p). The particular case $p = 2$ yields the spectral gap inequality

$$\mathrm{Var}_\mu(f) \le \frac{1}{n+2}\, \mathcal{E}(f, f). \tag{4.7}$$

Here, the optimal value of the constant is known and is equal to $\frac{1}{2n}$ (see [10]). For this constant, equality in (4.7) is attained for any linear function f on $\mathbf{R}^n$. On the other hand, dividing both sides of (4.6) by $p-1$, letting $p \to 1$, we arrive at the modified logarithmic Sobolev inequality (4.4) stated above.

Proof of Proposition 4.5. We may identify $\Omega(n, k)$ with a slice of the discrete cube, $\{x \in \{0, 1\}^n : x_1 + x_2 + \cdots + x_n = k\}$, so that $\Omega(n, k)$ inherits the structure of a graph from the discrete cube: neighbors are the pairs of points which differ exactly in two coordinates. The canonical inner metric $\rho = \rho_{n,k}$ in $\Omega(n, k)$ is given by

$$\rho(x, y) = \frac{1}{2}\, \mathrm{card}\{i \le n : x_i \ne y_i\}, \quad x, y \in \Omega(n, k),$$

that is, one half of the Hamming distance.

For $1 \le k \le n - 1$, let $A_{n,k}$ denote the best constant in

$$\|f\|_p^p - \|f\|_1^p \le A_{n,k}\, \mathcal{E}(f, f^{p-1}), \tag{4.8}$$

where f is an arbitrary positive function on $\Omega = \Omega(n, k)$. In terms of the function R this inequality takes the form

$$\int f^p \, d\mu \le \left(\int f \, d\mu \right)^p + A_{n,k}\, \frac{1}{C_n^k} \sum_{\rho(x,y)=1} R(f(x), f(y)), \tag{4.9}$$

where the summation is performed over all ordered pairs $(x,y) \in \Omega \times \Omega$ such that $\rho(x,y) = 1$. By symmetry, $A_{n,k} = A_{n,n-k}$.

We know that $A_{n,1} \le \frac{1}{2n}$. As for $k \ge 2$, we will deduce a recursive inequality relating $A_{n,k}$ to $A_{n-1,k-1}$, and then we may proceed by induction. Thus, fix $k \ge 2$ and a positive function f on Ω with $\int f\, d\mu = 1$ (this can be assumed in view of homogeneity of (4.8)-(4.9)). Introduce the subgraphs

$$\Omega_i = \{x \in \Omega : x_i = 1\}, \quad 1 \le i \le n,$$

and equip them with uniform measures μ_i. Since all Ω_i can be identified with $\Omega(n-1, k-1)$, we may write the definition (4.9) for these graphs:

$$\int_{\Omega_i} f^p\, d\mu_i \le \Big(\int_{\Omega_i} f\, d\mu_i\Big)^p + \frac{A_{n-1,\,k-1}}{C_{n-1}^{k-1}} \sum_{x\in\Omega_i} \sum_{y\in\Omega_i,\, \rho(x,y)=1} R(f(x), f(y)).$$

Setting $\varphi(i) = \int_{\Omega_i} f\, d\mu_i$ and summing these inequalities over all $i \le n$ with weight $\frac{1}{n}$, we obtain

$$\frac{1}{n}\sum_{i=1}^n \int_{\Omega_i} f^p\, d\mu_i \le \frac{1}{n}\sum_{i=1}^n \varphi(i)^p + \frac{A_{n-1,\,k-1}}{nC_{n-1}^{k-1}} \sum_{i=1}^n \sum_{x\in\Omega_i} \sum_{y\in\Omega_i,\, \rho(x,y)=1} R(f(x), f(y)). \tag{4.10}$$

Since $\frac{1}{n}\sum_{i=1}^n \mu_i = \mu$, the first term in (4.10) is equal to $\int f^p\, d\mu$. The second term is estimated from above, according to the case $k = 1$, by

$$\left(\frac{1}{n}\sum_{i=1}^n \varphi(i)\right)^p + \frac{A_{n,1}}{C_n^1} \sum_{i\ne j} R(\varphi(i), \varphi(j)).$$

But $\frac{1}{n}\sum_{i=1}^n \varphi(i) = \int f\, d\mu = 1$. Hence, (4.10) implies

$$\int f^p\, d\mu \le \frac{A_{n,1}}{n} \sum_{i\ne j} R(\varphi(i), \varphi(j)) + \frac{A_{n-1,\,k-1}}{nC_{n-1}^{k-1}} \sum_{i=1}^n \sum_{x\in\Omega_i} \sum_{y\in\Omega_i,\, \rho(x,y)=1} R(f(x), f(y)).$$

To treat the triple sum, fix $x, y \in \Omega$ with $\rho(x,y) = 1$. The number of all i such that $x \in \Omega_i$ and $y \in \Omega_i$ simultaneously is equal to $k-1$. Hence, the triple sum will contribute

$$(k-1) \sum_{x\in\Omega} \sum_{y\in\Omega,\, \rho(x,y)=1} R(f(x), f(y)) = (k-1) C_n^k\, \mathcal{E}(f, \log f).$$

Since $\dfrac{(k-1)\, C_n^k}{nC_{n-1}^{k-1}} = \dfrac{k-1}{k}$, we thus get

$$\int f^p\, d\mu \le \frac{A_{n,1}}{n} \sum_{i\ne j} R(\varphi(i), \varphi(j)) + \frac{(k-1)\, A_{n-1,\,k-1}}{k}\, \mathcal{E}(f, \log f). \tag{4.11}$$

To treat the sum in (4.11), for each pair (i,j), $i \ne j$, define the bijective map $s_{ij} : \{0,1\}^n \to \{0,1\}^n$,

$$(s_{ij}x)_r = x_r, \ \text{ for } r \ne i, j, \quad \text{and} \quad (s_{ij}x)_j = x_i, \ (s_{ij}x)_i = x_j.$$

It acts as a bijection between Ω_i and Ω_j and pushes forward μ_i onto μ_j, provided that $k \ge 2$. In particular, $\varphi(j) \equiv \int f(y)\, d\mu_j(y) = \int f(s_{ij}x)\, d\mu_i(x)$.

Now, by Lemma 4.4, the function R is convex in the quandrant $a, b > 0$. Consequently, by Jensen's inequality,

$$\begin{aligned} R(\varphi(i), \varphi(j)) &= R\left(\int f(x)\, d\mu_i(x), \int f(s_{ij}x)\, d\mu_i(x)\right) \\ &\le \int R(f(x), f(s_{ij}x))\, d\mu_i(x). \end{aligned}$$

Therefore,

$$\sum_{i\ne j} R(\varphi(i), \varphi(j)) \le \frac{1}{C_{n-1}^{k-1}} \sum_{i\ne j} \sum_{x\in\Omega_i} R(f(x), f(s_{ij}x)). \tag{4.12}$$

Note that $y = s_{ij}x$ always implies $\rho(x,y) \le 1$, and in case $x \in \Omega_i$, the equality $\rho(x,y) = 1$ is only possible when $x_i = 1$, $x_j = 0$. Hence, the double sum in (4.12) contains only terms $R(f(x), f(y))$ with $\rho(x,y) = 1$ (the cases $\rho(x,y) = 0$ can be excluded). In turn, fixing any pair $(x,y) \in \Omega$ such that $\rho(x,y) = 1$, there is a unique pair (i,j) such that $i \ne j$ and $y = s_{ij}x$. Thus, the right hand side of (4.12) is just

$$\frac{1}{C_{n-1}^{k-1}} \sum_{x\in\Omega} \sum_{y\in\Omega,\, \rho(x,y)=1} R(f(x), f(y)) = \frac{n}{k}\, \mathcal{E}(f, f^{p-1}),$$

and we get from (4.11)

$$\int f^p\, d\mu \le \frac{A_{n,1} + (k-1)\, A_{n-1,\,k-1}}{k}\, \mathcal{E}(f, f^{p-1}).$$

Hence, $A_{n,k} \le \dfrac{A_{n,1} + (k-1)\, A_{n-1,\,k-1}}{k}$, or in terms of $B_{n,k} = kA_{n,k}$,

$$B_{n,k} \le A_{n,1} + B_{n-1,\,k-1}.$$

Applying this inequality successively $k-1$ times and recalling that $A_{r,1} \le \frac{1}{2r}$, we arrive at

$$\begin{aligned} B_{n,k} &\le \frac{1}{2n} + \frac{1}{2(n-1)} + \cdots + \frac{1}{2(n-(k-2))} + B_{n-(k-1),\,1} \\ &\le \frac{1}{2n} + \frac{1}{2(n-1)} + \cdots + \frac{1}{2(n-(k-2))} + \frac{1}{2(n-(k-1))}. \end{aligned}$$

If $k \le \frac{n}{2}$, each of the above k terms does not exceed $\frac{1}{n+2}$, so $B_{n,k} \le \frac{k}{n+2}$. This implies the desired estimate $A_{n,k} \le \frac{1}{n+2}$. In the case $k \ge \frac{n}{2}$, we may use $A_{n,k} = A_{n,n-k}$, and Proposition 4.5 follows. □

A similar statement can be made about the symmetric group $M = S_n$ in which case we have (4.4) with the same constant $\frac{1}{n+2}$ for the Dirichlet form (4.5):

PROPOSITION 4.6. *Let $p \in (1,2]$. For every positive function f on S_n, $n \ge 2$, with respect to the uniform probability measure*

$$\|f\|_p^p - \|f\|_1^p \le \frac{1}{n+2}\, \mathcal{E}(f, f^{p-1}),$$

Equivalently, in the Markov kernel setting,

$$\frac{p\,(n+2)}{2n(n-1)} \le \alpha(p) \le \lambda_1 = \frac{2}{n-1}.$$

Now let us return to Theorem 4.1 and the inequality (4.2) about the mixing time. Since the norm $\|f_0\|_p$ is maximized for Dirac measure $\mu_0 = \delta_x$, for some $x \in M$, we obtain similarly to (2.4) a more general bound

$$\|f_t\|_p^p - 1 \leq \frac{1 - \pi_*^{p-1}}{\pi_*^{p-1}} \; e^{-2\alpha(p)t}, \quad t \geq 0,$$

where $\pi_* = \min_x \pi(x)$. Letting $p \downarrow 1$ helps us recover the previous estimate on the informational divergence, cf. (2.2) and (2.6),

$$\mathrm{Ent}_\pi(f_t) \leq \log \frac{1}{\pi_*} \; e^{-2\rho_0 t}, \quad t \geq 0.$$

5. BOUNDS ON DIAMETER

Throughout this section we assume that $G = (V, E)$ is a finite, connected, undirected graph. For simplicity we also assume that G is d-regular, although typically the weaker assumption, that the maximum degree is at most d, is sufficient. Let μ be uniform over V. More over, let (G, μ) satisfy the inequality: for all $f > 0$ on V,

$$\rho_1 \mathrm{Ent}_\mu(e^f) \leq \frac{1}{2} E_\mu |\nabla f|^2 e^f, \tag{5.1}$$

for some $\rho_1 > 0$; recall that $|\nabla f(x)|^2 = \sum_{y: y \sim x} (f(x) - f(y))^2$, for $x \in V$.

Then we have the following bound on the diameter $D = D(G)$ of G.

PROPOSITION 5.1. $D \leq 2\sqrt{\frac{2d}{\rho_1} \log |V|}$.

REMARK 5.2. *This improves upon similar results by [1] and [8], where the bounds are of the type: $D \leq c\sqrt{\frac{d}{\lambda_1}} \log |V|$, for $c > 0$ some universal constant. Using the general inequalities, $\rho \geq \lambda_1/(\log |V|)$ (see e.g. [24]) and $\rho_1 \geq \rho$ (see comments after the proof of Proposition 3.6), it is clear that the Proposition 5.1 is an improvement. Results in [7] and [23] also provide improvements over [1], but are in general incomparable with ours.*

REMARK 5.3. *The proposition also implies that for bounded degree expander graphs, ρ, ρ_0 and ρ_1 are all of the order of $1/\log |V|$, where the constants would depend on the bounds on the degree and the expansion, or equivalently, the spectral gap. Indeed since we have,*

$$\frac{\lambda_1}{\log |V|} \leq \rho \leq \rho_0 \leq \rho_1 \leq \frac{8d \log |V|}{D^2},$$

and since for graphs with degree at most d_0, the diameter is at least $\log_{d_0} |V|$, up to a universal constant, we verify the above assertion.

Proof of Proposition 5.1. Applying the inequality (5.1) to tf, with $t \in \mathbb{R}$, f:Lipschitz, $E_\mu f = 0$, we get

$$\rho_1 \mathrm{Ent}_\mu(e^{tf}) \leq \frac{dt^2}{2} E_\mu e^{tf}. \tag{5.2}$$

Setting $Ee^{tf} = e^{tu(t)}$, $\mathrm{Ent}(e^{tf})$ becomes $t^2 u'(t) e^{tu(t)}$. Plugging into (5.2) yields that $u'(t) \leq d/(2\rho_1)$, which in turn implies that $u(t) \leq dt/(2\rho_1)$. Hence

$$E_\mu e^{tf} \leq e^{dt^2/2\rho_1}. \tag{5.3}$$

Now let us tensorize (5.3) on $G \square G$, the Cartesian product of G with itself: as mentioned in Remark 3.11, $\rho_1(G \square G) = \rho_1(G)$. Consider g on $V \times V$, with $g(x, y) := f(x) - f(y)$. Then f: Lipschitz on G implies that g: Lipschitz on $G \square G$, and more over, $E_{\mu \times \mu} g = 0$. Thus applying (5.3) with g, and noting $G \square G$ is regular with degree $2d$ yields:

$$E_{\mu \times \mu} e^{t(f(x) - f(y))} \leq e^{dt^2/\rho_1}.$$

On the other hand, letting $M = \max f(x)$ and $m = \min f(x)$, we have

$$E_{\mu \times \mu} e^{t(f(x) - f(y))} \geq \frac{e^{t(M-m)}}{|V|^2}.$$

Thus we may conclude that for all $t \in \mathbb{R}$, we have

$$M - m \leq \frac{dt}{\rho_1} + \frac{2 \log |V|}{t}.$$

Minimizing over t yields, for all Lipschitz f on G,

$$\max f - \min f \leq 2\sqrt{\frac{2d}{\rho_1} \log |V|}. \tag{5.4}$$

To conclude the proof of the proposition, let us take $f(x) = d(x, x_0)$, for $x_0 \in V$, and maximize the left hand side of (5.4) over all choices of x_0:

$$D \leq 2\sqrt{\frac{2d}{\rho_1} \log |V|}. \tag{5.5}$$

□

We refer the reader to the journal version of our paper for other bounds on the diameter of a graph using ρ and ρ_0, and also for related results on concentration of measure on product graphs.

REMARK 5.4. *It is well known that the spectral gap of a graph can be estimated efficiently up to arbitrary accuracy. Considering the computational complexity of ρ_0 of an undirected graph on n vertices, let us remark that ρ_0 can also be estimated up to arbitrary accuracy using, say, the ellipsoid algorithm : indeed computing ρ_0 corresponds to minimizing the convex functional $\sum_x \sum_{y \sim x} R(f(x), f(y))\mu(x)$ over the convex body $\{f \in \mathbf{R}^n : \mathrm{Ent}_\mu f \leq 1\}$.*

The computational complexity of ρ was raised as an open question in [24]. Note that the above argument can not be made directly regarding ρ, since $\{f : \mathrm{Ent}_\mu f^2 \leq 1\}$ is not a convex body in $\mathbf{R}^n$. However, it is known from [5] that the log-Sobolev inequality in the form (3.2) can equivalently be rewritten as

$$\rho \mathcal{L}(f) \leq \mathcal{E}(f, f),$$

where $L(f) = \sup_a \mathrm{Ent}_\mu \left((f + a)^2\right)$. Furthermore, it is shown in [5] that

$$\frac{2}{3} \|f\|_N^2 \leq \mathcal{L}(f) \leq \frac{13}{4} \|f\|_N^2,$$

where $\|.\|_N$ denotes the Orlicz norm corresponding to the (convex) Young function $N(x) = x^2 \log(1 + x^2)$. Thus over the convex body $\{f \in \mathbf{R}^n : \|f\|_N^2 \leq 1\}$ one can minimize the convex functional $\sum_x \sum_{y \sim x} (f(x) - f(y))^2 \mu(x)$ to estimate ρ to within a factor of 39/8.

6. HYPERCONTRACTIVITY

In this section our treatment is very general, including both the continuous and the discrete cases. Let (M, μ) be a probability space and let $\mathcal{A}$ be a linear space of bounded measurable functions on M. Let $L : \mathcal{A} \to \mathcal{A}$ be a linear operator associated with the Dirichlet form $\mathcal{E} : \mathcal{A} \times \mathcal{A} \to \mathbf{R}$ in the sense that

$$-\int fLg\, d\mu = \mathcal{E}(f, g), \text{ for all } f, g \in \mathcal{A}. \tag{6.1}$$

It is easy to see that whenever such an operator exists, it is unique. Furthermore, assume a family of linear operators $P_t : \mathcal{A} \to \mathcal{A}$, $t \geq 0$, is associated with $\mathcal{E}$, with the properties that:

1) P_t has generator L satisfying the relation (6.1);

2) For all $f \in \mathcal{A}$ and $t_0 > 0$, $\sup_{0 \leq t \leq t_0} \|P_t f\| < +\infty$.

Note that the property of having a generator is understood in the following sense: for all $f \in \mathcal{A}$ and $t \geq 0$,

$$\lim_{\varepsilon \to 0} \frac{P_{t+\varepsilon} f - P_t f}{\varepsilon} = L(P_t f) \tag{6.2}$$

with convergence in the norm of the space $L^1(\mu)$. Then we have:

THEOREM 6.1. *Given a number ρ_0, the following properties are equivalent:*

a) The Dirichlet form $\mathcal{E}$ satisfies the modified logarithmic Sobolev inequality

$$\rho_0 \operatorname{Ent}(e^f) \leq \frac{1}{2}\, \mathcal{E}(e^f, f), \quad f \in \mathcal{A}. \tag{6.3}$$

b) For all $t \geq 0$ and $f \in \mathcal{A}$,

$$\left\| e^{P_t f} \right\|_{e^{2\rho_0 t}} \leq \left\| e^f \right\|_1. \tag{6.4}$$

All the norms here are taken in the Lebesgue spaces $L^q(\mu)$ (although we say "norm" even if $q < 1$). The equivalence of (6.4) and log-Sobolev inequality (3.2) is well known in the continuous setting (cf. [3]). Here we are dealing with the most general formulation fitting both continuous and discrete cases. The main point and motivation is that, in discrete spaces, the constant ρ_0 can be much better than ρ.

Proof. See the journal version of the paper.

7. REFERENCES

[1] Alon, N., Milman, V. λ_1, isoperimetric inequalities for graphs and superconcentrators, *J. Comb. Theory Ser. B* **38** (1985), 73–88.

[2] Bakry, D. L'hypercontractivité et son utilisation en théorie des semigroups. *Ecole d'Été de Saint Flour, 1992. Lect. Notes in Math.*, 1581 (1994), Springer-Berlin.

[3] Bakry, D., Emery, M. Diffusions hypercontractive. *Séminaire de Probabilité* XIX, *Lect. Notes in Math.*, 1123 (1994), 179–206, Springer, Berlin.

[4] Beckner, W. A generalized Poincaré inequality for Gaussian measures. *Proc. of the AMS*, 105 (1989), No 2, 397–400.

[5] Bobkov, S.G., Götze, F. Exponential integrability and transporation cost related to logarithmic transportation inequalities. *J. Funct. Anal.*, 163 (1999), 1–28.

[6] Bobkov, S.G., Ledoux, M. On modified logarithmic Sobolev inequalities for Bernoulli and Poisson measures. *J. Funct. Anal.*, 156 (1998), 347–365.

[7] Chung, F.R.K. Diameters and Eigenvalues, *J. Amer. Math. Soc.* **2** (1989), 187–196.

[8] Chung, F.R.K., Grigor'yan, A., Yau, S.-T. Higher eigenvalues and isoperimetric inequalities on Riemannian manifolds and graphs. *Comm. Anal. Geom.* 8 (2000), no. 5, 969–1026.

[9] Cover, T.M., Thomas, J.A. Elements of Information Theory. John Wiley & Sons, New York (1991).

[10] Diaconis, P., Group representations in Probability and Statistics. IMS, Hayward, CA (1988).

[11] Diaconis, P., Shashahani, M. Time to reach stationarity in the Bernoulli-Laplace diffusion model. *SIAM J. Math. Anal.*, 18 (1987), 208–218.

[12] Diaconis, P., Saloff-Coste, L. Logarithmic Sobolev inequalities for finite Markov chains. *Ann. Appl. Probab.*, 6 (1996), 695–750.

[13] Frieze, A., Kannan, R. Log-Sobolev inequalities and sampling from log-concave distributions. Preprint (1998).

[14] Gao, F., Quastel, J., Exponential decay of entropy in the Random Transposition and Bernoulli-Laplace models, *Ann. Appl. Probab.*, to appear.

[15] Gross, L. Logarithmic Sobolev inequalities. *Amer. J. Math.*, 97 (1975), 1060–1083.

[16] Higuchi, Y., Yoshida, N. Analytic conditions and phase transition for Ising models. *Lect. Notes in Japanese*, 1995.

[17] Houdré, C., Tetali, P. Concentration of measure for products of Markov kernels and graph products via functional inequalities. *Comb. Probab. & Comp.* 10 (2001), 1–28.

[18] Latala, R., Oleszkiewicz, K. Between Sobolev and Poincaré. *Geometric aspects of Funct. Anal.*, *Lect. Notes in Math.*, 1745 (2000), 147–168.

[19] Ledoux, M. The concentration of measure phenomenon. Mathematical Surveys and Monographs, 89. American Mathematical Society, Providence, RI, 2001.

[20] Lee,T.Y., Yau, H.T. Logarithmic Sobolev inequality for some models of random walks. *Ann. Probab.* 26 (1998), no. 4, 1855–1873.

[21] Lieb, E.H. Some convexity and subadditivity properties of entropy. *Bull. Amer. Math. Soc.*, 81 (1975), 1–13.

[22] Miclo, L. Sur les problémes de sortie discrets inhomogénes. *Ann. Appl. Probab.*, 6 (1996), No.4, 1112–1156.

[23] Mohar, B. Eigenvalues, diameter, and mean distance in graphs. *Graphs Combin.* **7** (1991), 53–64.

[24] Saloff-Coste, L. Lectures on finite Markov chains. *Lect. Notes in Math.*, 1665 (1997), 301–413, Springer, Berlin.

[25] Stroock, D. Logarithmic Sobolev inequalities for Gibbs measures. *Lect. Notes in Math.*, 1563 (1993), Springer, Berlin.