2015 IEEE 56th Annual Symposium on Foundations of Computer Science

Talagrand's Convolution Conjecture on Gaussian Space

[An extended abstract]

Ronen Eldan
Mathematics
Weizmann Institute

James R. Lee
Computer Science
University of Washington

Abstract

Smoothing properties of the noise operator on the discrete cube and on Gaussian space have played a pivotal role in many fields. In particular, these smoothing effects have seen a broad range of applications in theoretical computer science. We exhibit new regularization properties of the noise operator on Gaussian space. More specifically, we show that the mass on level sets of a probability density decays uniformly under the Ornstein-Uhlenbeck semigroup. This confirms positively the Gaussian case of Talagrand's convolution conjecture (1989) on the discrete cube.

A major theme is our use of an Itô process (the "Föllmer drift") which can be seen as an entropy-optimal coupling between the Gaussian measure and another given measure on Gaussian space. To analyze this process, we employ stochastic calculus and Girsanov's change of measure formula.

The ideas and tools employed here provide a new perspective on hypercontractivity in Gaussian space and the discrete cube. In particular, our work gives a new way of studying "small" sets in product spaces (e.g., sets of size $2^{o(n)}$ in the discrete cube) using a form of regularized online gradient descent.

Keywords

Gaussian space; hypercontractivity; the Ornstein-Uhlenbeck semigroup; stochastic calculus

I. Introduction

Fix $n \geq 1$ and consider a function $f : \{-1,1\}^n \to \mathbb{R}$. For $\rho \in [0,1]$, the *noise operator* T_ρ maps f to the function

$$T_\rho f(x) = \mathbb{E}[N_\rho(x)]\,,$$

where $N_\rho(x) \in \{-1,1\}^n$ is a random variable whose coordinates are independently distributed such that

$$N_\rho(x)_i = \begin{cases} x_i & \text{with prob. } \rho \\ \varepsilon_i & \text{with prob. } 1-\rho\,, \end{cases}$$

where $\{\varepsilon_i\}$ is a sequence of i.i.d. uniform ± 1 random variables.

One might expect that such a diffusion operator should serve to smoothen the function f. Indeed, one has the following *hypercontractive* estimate [Bon70], [Gro75]:

$$\|T_\rho f\|_q \leq \|f\|_p \qquad \text{for all } q \geq p > 1 \text{ such that } \rho \leq \sqrt{\frac{q-1}{p-1}}. \tag{1}$$

Here we work in the function space $L^p(\{-1,1\}^n)$ equipped with the norm

$$\|f\|_p = \left(2^{-n} \sum_{x\in\{-1,1\}^n} |f(x)|^p\right)^{1/p}.$$

This is an extended abstract of the full paper "Regularization under diffusion and anti-concentration of temperature," math arXiv: 1410.3887.

This theorem has many applications in areas like hardness of approximation, communication complexity, computational learning theory, data structures, social choice theory, and quantum information. We refer to the book [O'D14] and the surveys [DW08], [Mon12].

Talagrand's convolution conjecture. Observe that (1) asserts a smoothing effect assuming a non-trivial a priori smoothness bound (on the L^p norm for $p > 1$). In 1989, Talagrand conjectured that a related effect should be possible given only L^1 information about f. Let $\mathbb{E}$ and $\mathbb{P}$ denote expectation and probability with respect to the uniform distribution on $\{-1,1\}^n$. Given a non-negative $f : \{-1,1\}^n \to \mathbb{R}_+$, Markov's inequality immediately yields that for every $\alpha \geq 1$,

$$\mathbb{P}(f > \alpha \mathbb{E} f) \leq \frac{1}{\alpha}\,.$$

Talagrand conjectured that for smoothed functions, Markov's inequality is not tight. (Informally, smoothed functions cannot look like step functions.)

Conjecture I.1 ([Tal89])**.** *For every $0 \leq \rho < 1$, there exists a function $\varphi_\rho : [1,\infty) \to [1,\infty)$ with $\lim_{\alpha\to\infty} \varphi_\rho(\alpha) = \infty$ such that for every $f : \{-1,1\}^n \to \mathbb{R}_+$ and any $\alpha > 1$,*

$$\mathbb{P}\left(T_\rho f(x) > \alpha \|f\|_1\right) \leq \frac{1}{\alpha\, \varphi_\rho(\alpha)}\,.$$

Note that for non-negative f, we have $\|T_\rho f\|_1 = \|f\|_1$ since T_ρ preserves positivity and the mean value. The conjecture is still open for any value $\rho \in (0,1)$ and even in the case when $f = \mathbf{1}_S$ is merely the characteristic function of a subset $S \subseteq \{-1,1\}^n$. Talagrand noted that for any fixed $\rho \in (0,1)$, the best asymptotic improvement one can hope for is $\varphi_\rho(\alpha) \asymp \sqrt{\log \alpha}$. This would be tight for a threshold function $f(x) = \mathbf{1}_{\{\sum x_i > \tau\}}$ when $\tau = \tau(\rho)$ is chosen appropriately.

He also showed that a similar inequality for the averaged operator $A = \int_{1/2}^{1} T_\rho \, dt$ does hold. Specifically, there is a constant $C > 0$ such that for all $\alpha > e^3$,

$$\mathbb{P}\left(Af \geq \alpha \|f\|_1\right) \leq \frac{C \log\log \alpha}{\log \alpha}\,.$$

His proof makes clever use of $\approx \log \alpha$ invocations of (1).

A. The Gaussian limiting case

It is often the case that conjectures for the discrete cube have an analog over Gaussian space. For dimension-free inequalities, the Gaussian version is often a special case (obtained via approximately embedding Gaussian space into the discrete cube using the central limit theorem). In the present work, we prove a general result about anti-concentration of functions on level sets in Gaussian space which, in particular, resolves the Gaussian limiting case of Conjecture I.1. We hope the ideas and tools employed here provide a new perspective on the geometry of Gaussian space and the discrete cube. Given the importance of the noise operator to a variety of applications, the potential utility appears significant.

Let $n \geq 1$ and equip $\mathbb{R}^n$ with the standard Gaussian measure γ_n. Consider a function $f : \mathbb{R}^n \to \mathbb{R}$ in $L^1(\gamma_n)$. The Ornstein-Uhlenbeck semi-group $\{U_t : t \geq 0\}$ is defined by

$$U_t f(x) = \mathbb{E} f\left(e^{-t} x + \sqrt{1 - e^{-2t}} Z\right)\,,$$

where Z has law γ_n. This is the appropriate analog of the noise operator in the discrete setting (or perhaps more accurately, the corresponding heat diffusion operator). Again, one expects that the action of such a diffusion process serves to smoothen f. Indeed, Nelson's hypercontractivity theorem [Nel73] shows that U_t is a contraction from $L^p(\gamma_n)$ to $L^q(\gamma_n)$ for $1 < p \leq q$ and $t \geq \frac{1}{2} \log \frac{q-1}{p-1}$.

The concept of Gaussian hypercontractivity also plays an important role in several mathematical fields. For example, in quantum field theory hypercontractivity can often be used to show that a Hamiltonian is essentially self-adjoint on its domain, laying the foundation for various constructions (see, e.g., [GRS75]). We refer to the surveys [DGS92], [Gro06]. In the study of partial differential equations, it is a key method in several approaches to establishing the existence and uniqueness of smooth solutions to evolution equations [Bre11].

In the present work, we assert a regularizing effect of U_t merely assuming that $f \in L^1(\gamma_n)$. An important special case is when f is simply the indicator of a measurable subset of $\mathbb{R}^n$. Assume now that $f : \mathbb{R}^n \to \mathbb{R}_+$ is non-negative. Certainly we have Markov's inequality: For any $\alpha \geq 1$,

$$\gamma_n \Big(\{x : f(x) \geq \alpha \|f\|_1\} \Big) \leq \frac{1}{\alpha} \,,$$

where we use $\|f\|_1 = \int |f| \, d\gamma_n$. Of course, this bound is easily seen to be tight for any $\alpha > 0$ by taking $f = \mathbf{1}_S$ for a measurable subset $S \subseteq \mathbb{R}^n$ with $\gamma_n(S) = 1/\alpha$. The "heat content" of f lies on a single level set, i.e. at a single "temperature." A very natural question arises: Can a smoothed version of f, i.e. $U_t f$ for some $t > 0$, have its heat content concentrated near a single high temperature? Of course, this is just the Gaussian analog of Conjecture I.1.

Conjecture I.2 (Gaussian variant)**.** *For every $t > 0$, there exist a function $\psi_t : [1, \infty) \to [1, \infty)$ with $\lim_{\alpha \to \infty} \psi_t(\alpha) = \infty$ such that for any measurable $f : \mathbb{R}^n \to \mathbb{R}_+$ and any $\alpha > 1$,*

$$\gamma_n \Big(\{x : U_t f(x) > \alpha \|f\|_1\} \Big) \leq \frac{1}{\alpha \, \psi_t(\alpha)} \,.$$

It is a straightforward observation that Conjecture I.1 implies Conjecture I.2 with $\varphi_{e^{-t}} = \psi_t$. This is proved by embedding Gaussian space (approximately) into a sequence of discrete cubes of growing dimension via the central limit theorem; we refer to the discussion in [BBB+13].

The conjecture posits a uniform bound on the tail of the smoothed function. Again, one observes that the best rate of decay one can expect is $\psi_t(\alpha) = c(t)\sqrt{\log \alpha}$ where $c(t)$ is some function depending only on t.

Ball, Barthe, Bednorz, Oleszkiewicz, and Wolff [BBB+13] prove that Conjecture I.2 holds in any fixed dimension; they achieve $\psi_t(\alpha) = C(t,n)\sqrt{\log \alpha}/(\log \log \alpha)$ where $C(t,n)$ is a constant depending (exponentially) on the dimension n.

In the remainder of this manuscript, we describe our resolution of Conjecture I.2 achieving the bound

$$\psi_t(\alpha) = \frac{c(t)\sqrt{\log \alpha}}{(\log \log \alpha)^4} \,.$$

See Corollary I.5 for a formal statement. Full details of the proof can be found in [EL14].

The heat analogy. If one imagines the function $f : \mathbb{R}^n \to \mathbb{R}_+$ as assigning an initial distribution of heat to space, Conjecture I.2 asserts that if we allow the heat to diffuse for a short period of time, then the resulting distribution cannot have all the heat concentrated in a narrow range of high temperatures. More formally, we will see that for $t > 0$,

$$\int (U_t f) \mathbf{1}_{\{U_t f \in [\alpha, 2\alpha]\}} \, d\gamma_n \leq c(t) \frac{(\log \log \alpha)^4}{\sqrt{\log \alpha}} \|f\|_1 \,.$$

An isoperimetric perspective. A dual point of view is helpful in understanding the isoperimetric content of Conjecture I.2. Fix $t > 0$, let $S \subseteq \mathbb{R}^n$ be an open subset, and consider the set of non-negative functions

$g : \mathbb{R}^n \to \mathbb{R}_+$ supported on S, and such that $\|U_t g\|_\infty \le 1$. Our goal is to maximize $\int g \, d\gamma_n$ subject to these constraints.

Clearly the choice $g = \mathbf{1}_S$ has $\int g \, d\gamma_n = \gamma_n(S)$. Conjecture I.2 asserts that there should be a strategy that does much better. In fact, the largest function ψ_t achievable in Conjecture I.2 is precisely the same as the largest function ψ_t such that the following holds for every open $S \subseteq \mathbb{R}^n$:

$$\sup_{\mathrm{supp}(g) \subseteq S} \int \frac{g}{\|U_t g\|_\infty} \, d\gamma_n \ge \psi_t \left(1/\gamma_n(S)\right) \gamma_n(S) \,. \tag{2}$$

This dual characterization is a straightforward consequence of Hahn-Banach and self-adjointness of U_t as an operator on $L^2(\gamma_n)$ (i.e., LP duality). We refer to this optimization problem as "isoperimetric" because the intuition is that to make g significantly larger subject to the constraint $\|U_t g\|_\infty \le 1$, one should concentrate g on the "boundary" of the set S where it may be allowed to take larger values due to the smoothing effect of U_t.

Indeed, one can prove Conjecture I.2 for $n = 1$ via the dual (2) as follows: Given $S \subseteq \mathbb{R}$, one should choose g to be a Dirac mass near the point of $\mathbb{R} \setminus S$ which is closest to the origin. (Strictly speaking, one should take a sequence of points in S and a sequence of functions approximating Dirac masses at those points.) From the value $\gamma_n(S)$, one can conclude that S contains a point sufficiently close to the origin. A simple calculation with the Gaussian density yields the desired bound[1].

Brief outline of this abstract. In the next section (Section I-B), we give a formal statement of our results, as well as a short discussion of related work. In Section II, we present a proof of the log-Sobolev inequality in the discrete cube $\{-1, 1\}^n$ as a motivation for studying "adapted couplings" to Brownian motion. Many of the elementary elements in this discrete proof resurface in the Gaussian setting. In Section III, we give a short review and pointers to the literature for various aspects of Brownian motion, Itô processes, stochastic differential equations, and Girsanov's change of measure formula. Finally, Section IV contains an outline of our argument for experts, and those who have familiarized themselves with the tools described in Section III.

B. Semi-log-convexity and anti-concentration of temperature

The resolution of Conjecture I.2 arises from a more general phenomenon for semi-log-convex functions. Our main theorem follows.

Theorem I.3. *Let $f : \mathbb{R}^n \to \mathbb{R}_+$ be a function with continuous second-order partial derivatives. Assume there is a $\beta \ge 1$ such that for all $x \in \mathbb{R}^n$,*

$$\nabla^2 \log f(x) \succeq -\beta \, \mathrm{Id} \,. \tag{3}$$

Then for all $\alpha \ge e^3$,

$$\gamma_n \left(\{ x \in \mathbb{R}^n : f(x) > \alpha \|f\|_1 \} \right) \le \frac{1}{\alpha} \cdot \frac{C\beta (\log\log\alpha)^4}{\sqrt{\log \alpha}} \,, \tag{4}$$

where $C > 0$ is a universal constant.

We first explain how this resolves Conjecture I.2 before moving on to a discussion of Theorem I.3. Let $\{B_t\}$ be an n-dimensional Brownian motion with $B_0 = 0$, and let $P_t f(x) = \mathbb{E}[f(x + B_t)]$ denote the corresponding semigroup. We present a proof of the following well-known fact for the sake of the reader.

[1]Completion of this sketch is Exercise 11.31 in O'Donnell's book [O'D14]. It reflects an observation of O'Donnell and the second-named author from 2010.

Lemma I.4. *For any $f : \mathbb{R}^n \to \mathbb{R}_+$ in $L^1(\gamma_n)$ and $t > 0$, one has $\nabla^2 \log P_t f(x) \succeq -\frac{1}{t}\mathrm{Id}$ for all $x \in \mathbb{R}^n$.*

Proof: The proof is a simple application of the fact that a mixture of log-convex densities is log-convex (see, e.g., [MOA11, p.649]). Observe that for any $y \in \mathbb{R}^n$, the function

$$x \to \frac{|x|^2}{2t} + \log\left(P_t(\delta_y)(x)\right)$$

is convex (here, δ_y denotes a Dirac mass supported on $\{y\}$).

We now apply the aforementioned fact to conclude that

$$x \to \frac{|x|^2}{2t} + \log\left(\int_{\mathbb{R}^n} g(y)\big(P_t(\delta_y)(x)\big)dy\right)$$

must also be convex. In other words, the function

$$x \to \frac{|x|^2}{2t} + \log P_t(g)$$

is convex. We conclude that

$$\nabla^2 \log P_t(g) \succeq -\nabla^2\left(\frac{|x|^2}{2t}\right) = -\frac{1}{t}\mathrm{Id}.$$

■

Together with Theorem I.3, this rather immediately implies the following.

Corollary I.5. *There is a constant $C > 0$ such that for every $\rho \in (0, 1)$, the following holds. For every measurable $g : \mathbb{R}^n \to \mathbb{R}_+$ and every $\alpha \geq e^3$, one has*

$$\mathbb{P}\left(P_{1-\rho}g(B_\rho) > \alpha\|g\|_1\right) \leq \frac{1}{\alpha} \cdot \frac{\rho}{1-\rho} \frac{C(\log\log\alpha)^4}{\sqrt{\log\alpha}}\,. \tag{5}$$

Proof: If we define $f(x) = g(\sqrt{\rho}x)$, then $P_{1-\rho}g(B_\rho)$ and $P_{(1-\rho)/\rho}f(Z)$ have the same law, where Z is a standard n-dimensional Gaussian. Now combining Lemma I.4 and Theorem I.3 yields the desired result. ■

Corollary I.5 yields a resolution to Conjecture I.2 by noting that for any $t > 0$,

$$\gamma_n\left(\{x : U_t f(x) > \alpha\|f\|_1\}\right) = \mathbb{P}\left(P_{1-e^{-2t}}f(B_{e^{-2t}}) > \alpha\|f\|_1\right)\,.$$

Translating the anti-concentration of Brownian motion. Despite the fact that Theorem I.3 is not a stochastic statement, the main theme of our proof is that the variance of Brownian motion can be translated into anti-concentration for the level sets of certain functionals on Gaussian space.

Let $f : \mathbb{R}^n \to \mathbb{R}_+$ be as in the statement of Theorem I.3, and additionally let us assume that $\int f\, d\gamma_n = 1$. Our goal (4) is equivalent to bounding $\mathbb{P}(f(B_1) > \alpha)$, and it is easy to see that it would suffice to give an upper bound on $\mathbb{P}(f(B_1) \in [\alpha, 2\alpha])$.

Very roughly, this will be achieved as follows. We show that if $\mathbb{E}[f(B_1)\mathbf{1}_{\{f(B_1)\in[\alpha,2\alpha]\}}]$ is large enough, then the values $\mathbb{E}[f(B_1)\mathbf{1}_{\{f(B_1)\in[2^k\alpha,2^{k+1}\alpha]\}}]$ are proportionally large for approximately $\sqrt{\log\alpha}$ values of k. Since $\mathbb{E}[f(B_1)] = 1$, this yields the desired conclusion.

This "transfer of mass" between levels is achieved by carefully perturbing the underlying Brownian motion. The Hessian condition (3) ensures that f behaves predictably under small perturbations. The primary difficulty is to perform the perturbations without changing the measure of the underlying Brownian motion too much. For this purpose, we will employ an appropriate Itô process, and Girsanov's change of measure theorem will be essential.

Related work. Our use of random measures and stochastic calculus to study the geometry of Gaussian space is certainly closely related to the works [Eld13a], [Eld13b]. On the other hand, the idea to study functionals using an "optimal" adapted coupling to Brownian motion (see Section IV) comes from the viewpoint of stochastic control theory [Föl85], [Leh13] and its geometric applications [Leh13]. Other variational perspectives appear in the work [BD98] and in Borell's papers [Bor00], [Bor02] where one of his primary goals is their use in proving functional inequalities. An important distinction between our work and some previous ones involves our use of second-order methods. Specifically, we study the effect of perturbations on the optimal drift.

Finally, we should mention two vast bodies of work closely related to our study: Markov diffusions and semigroup methods (see, e.g., [BGL14]), as well as the theory of optimal transportation. For the latter topic, one might consult [Vil03, Ch. 9] for an excellent review of the literature related to functional inequalities.

II. Discrete warmup: A proof of the log-Sobolev inequality

Equip $\{-1,1\}^n$ with the uniform measure μ. In this section, we are careful to distinguish simple averaging operators from expectations and probabilities. Thus we will use the analyst's notation: For $f : \{-1,1\}^n \to \mathbb{R}$, we denote $\int f\,d\mu = 2^{-n}\sum_{x\in\{-1,1\}^n} f(x)$. Consider $f : \{-1,1\}^n \to \mathbb{R}_+$ with $\int f\,d\mu = 1$ and denote the relative entropy $H_\mu(f) := \int f \log f\,d\mu$. For $i \in \{1,\ldots,n\}$, we define

$$(\partial_i f)(x) = \frac{f(x \mid x_i = 1) - f(x \mid x_i = -1)}{2}\,,$$

where $f(x \mid x_i = b)$ denotes $f(x_1,\ldots,x_{i-1},b,x_{i+1},\ldots,x_n)$.

We will prove first the following "modified" log-Sobolev inequality, using ideas from an argument of Lehec in the Gaussian setting [Leh13] (also see Section IV)

$$H_\mu(f) \le \sum_{i=1}^n \int \frac{(\partial_i f)^2}{f} d\mu\,. \tag{6}$$

Then we will alter the proof slightly to yield a proof of the stronger inequality

$$H_\mu(f) \le 4\sum_{i=1}^n \int (\partial_i \sqrt{f})^2\,d\mu\,. \tag{7}$$

It is known that the constant 4 can be replaced by 2 (see [O'D14]). We remark that these inequalities are equivalent in the continuous setting since $(\sqrt{f})' = \frac{f'}{2\sqrt{f}}$. In discrete settings, (7) is stronger than (6). We refer to [BT06] for a discussion. It is known that, in significant generality, log-Sobolev inequalities of the form (7) are equivalent to hypercontractive estimates like (1); see [MT06] for a discussion.

An adapted sampling from f. Now consider a random variable $B = (b_1,\ldots,b_n)$ with law μ. We will give a way of sampling a random variable $W = (w_1,\ldots,w_n)$ with density $f\,d\mu$ (i.e., $\mathbb{P}(W = x) = 2^{-n}f(x)$). Suppose that $w_1,\ldots,w_t$ have been chosen. We define the quantity

$$v_t = \frac{\mathbb{E}[\partial_{t+1} f(B) \mid b_1 = w_1,\ldots,b_t = w_t]}{\mathbb{E}[f(B) \mid b_1 = w_1,\ldots,b_t = w_t]}\,. \tag{8}$$

Then we define the next bit of W by

$$w_{t+1} = \begin{cases} +1 & \text{with prob. } \frac{1+v_t}{2} \\ -1 & \text{with prob. } \frac{1-v_t}{2}\,. \end{cases} \tag{9}$$

Here is an equivalent way to describe this process: We have sampled w_{t+1} according to the marginal density coming from $f d\mu$, conditioned on our choices for $w_1, \ldots, w_t$. Thus, manifestly, W has the law of $f d\mu$.

Let us verify inductively the following equality

$$M_t \coloneqq \mathbb{E}[f(B) \mid b_1 = w_1, \ldots, b_t = w_t] = \prod_{i=1}^{t} (1 + v_{i-1} w_i)\,. \tag{10}$$

We have

$$\begin{aligned}
&\mathbb{E}[f(B) \mid b_1 = w_1, \ldots, b_t = w_t] \\
&\qquad = \frac{1}{2}\mathbb{E}[f(B) \mid b_1 = w_1, \ldots, b_t = w_t, b_{t+1} = 1] + \frac{1}{2}\mathbb{E}[f(B) \mid b_1 = w_1, \ldots, b_t = w_t, b_{t+1} = -1]\,,
\end{aligned}$$

thus

$$\begin{aligned}
&\mathbb{E}[f(B) \mid b_1 = w_1, \ldots, b_t = w_t, b_{t+1} = 1] \\
&\qquad = 2\,\mathbb{E}[f(B) \mid b_1 = w_1, \ldots, b_t = w_t] - \mathbb{E}[f(B) \mid b_1 = w_1, \ldots, b_{t+1} = -1] \\
&\qquad = (1 + v_t)\mathbb{E}[f(B) \mid b_1 = w_1, \ldots, b_t = w_t] \\
&\qquad = (1 + v_t) \prod_{i=1}^{t-1} (1 + v_{i-1} w_i)\,,
\end{aligned}$$

where the final line follows by the inductive hypothesis. The symmetric calculation holds for $w_{t+1} = -1$, confirming (10).

Equation (10) implies that $1/M_t$ is precisely the change of measure under which $(w_1, \ldots, w_t)$ has the law of $(b_1, \ldots, b_t)$. For the sake of analysis, define

$$v_t^i = \frac{\mathbb{E}[\partial_i f(B) \mid b_1 = w_1, \ldots, b_t = w_t]}{\mathbb{E}[f(B) \mid b_1 = w_1, \ldots, b_t = w_t]}\,.$$

Observe that $v_t^{t+1} = v_t$ for all $t \in \{0, 1, \ldots, n-1\}$, and for any fixed $i \in \{1, \ldots, n\}$, the process $M_t^i = \mathbb{E}[\partial_i f(B) \mid b_1, \ldots, b_t]$ is a Doob martingale. Therefore the change of measure formula (10) implies that $\{v_t^i\}$ is a martingale. Let us verify this carefully:

$$\begin{aligned}
&\mathbb{E}[v_t^i \,| w_1, w_2, \ldots, w_{t-1}] \\
&\quad = \mathbb{E}\left[\frac{\mathbb{E}[\partial_i f(B) \mid b_1 = w_1, \ldots, b_t = w_t]}{M_t} \mid w_1, w_2, \ldots, w_{t-1}\right] \\
&\quad = \frac{1}{2}(1 + v_t)\mathbb{E}\left[\frac{\mathbb{E}[\partial_i f(B) \mid b_1 = w_1, \ldots, b_t = w_t]}{M_t} \mid w_1, w_2, \ldots, w_{t-1}, w_t = 1\right] \\
&\qquad + \frac{1}{2}(1 - v_t)\mathbb{E}\left[\frac{\mathbb{E}[\partial_i f(B) \mid b_1 = w_1, \ldots, b_t = w_t]}{M_t} \mid w_1, w_2, \ldots, w_{t-1}, w_t = -1\right] \\
&\quad = \frac{1}{2}\frac{\mathbb{E}[\partial_i f(B) \mid b_1 = w_1, \ldots, b_{t-1} = w_{t-1}, b_t = 1]}{M_{t-1}} + \frac{1}{2}\frac{\mathbb{E}[\partial_i f(B) \mid b_1 = w_1, \ldots, b_{t-1} = w_{t-1}, b_t = -1]}{M_{t-1}} \\
&\quad = \frac{\mathbb{E}[\partial_i f(B) \mid b_1 = w_1, \ldots, b_{t-1} = w_{t-1}]}{M_{t-1}} \\
&\quad = v_{t-1}^i\,.
\end{aligned}$$

Now we may write:

$$\begin{aligned}
H_\mu(f) &= \mathbb{E}[\log f(W)] \\
&\overset{(10)}{=} \sum_{t=1}^{n} \mathbb{E}[\log(1+v_{t-1}w_t)] \\
&\leq \sum_{t=1}^{n} \mathbb{E}\left[v_{t-1}w_t\right] \\
&\overset{(9)}{=} \sum_{t=1}^{n} \mathbb{E}[(v_{t-1})^2] = \sum_{t=1}^{n} \mathbb{E}[(v_{t-1}^t)^2] \leq \sum_{t=1}^{n} \mathbb{E}[(v_n^t)^2]\,,
\end{aligned} \tag{11}$$

where in the final inequality, we have used that $\{v_t^i\}$ is a martingale.

Using the fact that $1/M_n = 1/f(W)$ is the change of measure under which W has the law μ, we have

$$\mathbb{E}[(v_n^t)^2] = \mathbb{E}\left[\frac{(\partial_t f(W))^2}{f(W)^2}\right] = \mathbb{E}\left[\frac{(\partial_t f(B))^2}{f(B)}\right].$$

Combining this with the preceding inequality finally yields our first goal (6).

The full log-Sobolev inequality. Let us now prove (7). For the sake of analysis, define the function $f_i(x) = \frac{1}{2}\left(f(x \mid x_i = 1) + f(x \mid x_i = -1)\right)$, and the value

$$\hat{v}_t^i = \frac{\mathbb{E}[\partial_i f_i(B) \mid b_1 = w_1, \ldots, b_t = w_t]}{\mathbb{E}[f_i(B) \mid b_1 = w_1, \ldots, b_t = w_t]}\,,$$

and note that $v_t = \hat{v}_t^{t+1}$ for all $t \in \{0, 1, \ldots, n-1\}$.

Next, fix $i \in \{1, \ldots, n\}$, and observe that if $\hat{M}_t^i = \mathbb{E}[f_i(B) \mid b_1 = w_1, \ldots, b_t = w_t]$, then $1/\hat{M}_t^i$ is the change of measure that gives $(w_1, \ldots, w_{i-1}, b_i, w_{i+1}, \ldots, w_n)$ the law of B. Thus $\{\hat{v}_t^i\}$ is *also* a martingale, since it does not depend on the value of the ith coordinate.

The proof now proceeds exactly as before to arrive at the inequality

$$H_\mu(f) \leq \sum_{t=1}^{n} \mathbb{E}[(\hat{v}_n^t)^2] \leq \sum_{t=1}^{n} \int \frac{(\partial_t f_t)^2}{f_t} d\mu \leq 4 \sum_{t=1}^{n} \int (\partial_t \sqrt{f})^2 \, d\mu\,.$$

where the final inequality uses the simple numerical fact valid for all $a, b \geq 0$:

$$(a-b)^2 \leq 2(\sqrt{a} - \sqrt{b})^2 (a+b)\,.$$

We remark that this method seems quite powerful. As an example, a similar line of argument recovers the Lee-Yau log-Sobolev inequality in the symmetric group [LY98]. We discuss this and related issues in a forthcoming manuscript.

III. Brownian motion, Itô processes, and Girsanov's formula

A key object in our study will be n-dimensional Brownian motion $\{B_t\}$. One can define the 1-dimensional process by taking an unbiased near-neighbor random walk on the set of points $\{\varepsilon j : j \in \mathbb{Z}\} \subseteq \mathbb{R}$ and then taking an appropriate limit as $\varepsilon \to 0$. We refer to [MP10] for the formal construction and an in-depth study. Now n-dimensional Brownian motion is simply $B_t = (B_t^{(1)}, B_t^{(2)}, \ldots, B_t^{(n)})$ where the processes $\{\{B_t^{(i)}\} : i \in \{1, \ldots, n\}\}$ are independent 1-dimensional Brownian motions. An important point to keep in mind is that the law of $B_t \mid B_0 = x$ is a standard n-dimensional Gaussian with variance t centered at x.

Itô processes. Given a Brownian motion, one can derive other stochastic processes. If we are given processes $\{u_t\}$ and $\{\sigma_t\}$ that are *adapated* to the Brownian motion and predictable (meaning that u_t and σ_t depend "only on the past up to time t"), then one can define a new stochastic process $\{W_t\}$ via the stochastic differential equation

$$dW_t = \sigma_t dB_t + u_t\, dt, \qquad W_0 = x_0\,.$$

Note that u_t is an n-dimensional vector, and σ_t is an $n \times n$ matrix. Under some technical assumptions concerning regularity of $\{u_t\}$ and $\{\sigma_t\}$, the above equation admits a unique solution $\{W_t\}$. One refers to a process constructed in this manner as an *Itô process.* We refer to [Øks03] for a very readable introduction to stochastic processes and Itô calculus. In this work, we will only consider processes with $\sigma_t = \mathrm{Id}$. In analogy with Section II, one can think of this as a biased sampling process.

Itô's formula. Take a real-valued, twice continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ and consider the quantity $df(W_t)$ given intuitively by

$$\text{“}df(W_t) = f(W_{t+dt}) - f(W_t)\text{”}$$

One might expect that the chain rule applies here; we would have $df(W_t) = f'(W_t)dW_t$. But this formula is incorrect, since Brownian motion (and hence W_t) has non-trivial quadratic variation (intuitively, $\mathbb{E}[\|dB_t\|^2] = dt$, hence we cannot simply pass to a first-order approximation). Itô's formula tell us that a second-order approximation suffices:

$$df(W_t) = f'(W_t)dW_t + \frac{1}{2}f''(W_t)\sigma_t^2\, dt\,.$$

In general, if f also depends on the variable t, then we have

$$df(W_t, t) = (\partial_t f)(W_t, t)\, dt + (\partial_x f)(W_t, t)dW_t + \frac{1}{2}(\partial_x^2 f)(W_t, t)\sigma_t^2\, dt\,.$$

As an example of the utility of this formula, recall line (11) from our previous argument in the discrete setting. In the continuous world, our coarse approximation for $\log(\cdot)$ can be replaced by an exact equality. When $f : \mathbb{R}^n \to \mathbb{R}^m$ takes values in a vector space, there is a corresponding multi-dimensional Itô formula (which we employ).

Girsanov's change of measure formula. Itô's formula will wield the most power for us in the form of Girsanov's change of measure formula. We will now restrict our attention to Itô processes which are just Brownian motion plus a drift:

$$dW_t = dB_t + u_t\, dt, \qquad W_0 = 0\,.$$

One might ask: Given a sample path $\{W_t : t \in [0, 1]\}$, what is the probability that an un-drifted Brownian motion $\{B_t : t \in [0, 1]\}$ could have followed the same path? Under some regularity conditions, Girsanov's formula (see [Øks03]) offers a precise answer:

$$\frac{dP}{dQ} = \exp\left(-\int_0^1 \langle u_t, dB_t\rangle - \frac{1}{2}\int_0^1 \|u_t\|^2\, dt\right).$$

Here, P is the law under which B_t is a Brownian motion and Q is the law under which W_t is a Brownian motion. The formula gives us exactly a weighting on paths to transfer between the two measures. Note that this carries a lot of intuition: The quadratic term is in line with the fact that, the stronger the drift, the less likely that the Brownian motion would have followed the same path. The linear term is related to the tail of the Gaussian density: If the Brownian motion itself was correlated with the drift, then there should be an extra "penalty" for this drift.

IV. Entropy, energy, and the Föllmer drift

Fix $n \geq 1$ and consider $\mathbb{R}^n$ with the equipped with the standard Euclidean structures $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$, and the Gaussian measure γ_n defined by

$$\frac{d\gamma_n}{dx} = \frac{1}{(2\pi)^{n/2}} \exp\left(-\|x\|^2/2\right).$$

To conclude this abstract, we present an informal discussion highlighting a stochastic calculus approach to the geometry of Gaussian space.

Suppose now that $f : \mathbb{R}^n \to \mathbb{R}_+$ has continuous second-order partial derivatives and $\int f \, d\gamma_n = 1$. Recall that, given $\alpha > 0$, we are interested in showing that $\mathbb{P}(f(B_1) \in [\alpha, 2\alpha]) \ll 1/\alpha$ as $\alpha \to \infty$, where $\{B_t\}$ is a Brownian motion with $B_0 = 0$. Since f could be concentrated on a set of very small measure, this would necessitate the study of events of very small probability. Instead, we will restrict our attention to the interesting parts of the space by changing the measure of the Brownian motion so that B_1 has law $f d\gamma_n$.

To this end, we define an Itô process $\{W_t\}$ by the stochastic differential equation

$$W_0 = 0, \quad dW_t = dB_t + v_t \, dt$$

for some predictable drift process $\{v_t\}$ with respect to the filtration $\{\mathcal{F}_t\}$ underlying the Brownian motion. Moreover, we will choose this drift as the solution to an energy optimization problem.

The following variational viewpoint is taken from the papers of Föllmer [Föl85] and Lehec [Leh13]. Lehec's work convincingly demonstrates its geometric applicability and it provided us with considerable inspiration. Let us take any predictable drift $\{u_t\}_{t \in [0,1]}$ such that $B_1 + \int_0^1 u_t \, dt$ has law $f d\gamma_n$. Among all such drifts, we will define $\{v_t\}$ to be the one that minimizes the quantity

$$\frac{1}{2} \int_0^1 \mathbb{E} \|u_t\|^2 \, dt.$$

It is quite beneficial to think of $\{v_t\}$ as the minimum-energy adapted coupling between $d\gamma_n$ and $f d\gamma_n$. Furthermore, one can connect this energy to the entropy of f:

$$H_{\gamma_n}(f) = \frac{1}{2} \int_0^1 \mathbb{E} \|v_t\|^2 \, dt, \tag{12}$$

where $H_{\gamma_n}(f) := \int f \log f \, d\gamma_n$ denotes the relative entropy of f with respect to γ_n. As one might expect, this optimality property of v_t implies that $\{v_t\}$ is a martingale with respect to $\{\mathcal{F}_t\}$, a fact that will be central in our study. In particular, the martingale property will imply that the behavior of $\{W_t\}$ at small times must have echoes that reverberate to time 1.

As we will see below, one can compute explicitly

$$v_t = \nabla \log P_{1-t} f(W_t). \tag{13}$$

Note that $\nabla \log P_{1-t} f(W_t) = \frac{\nabla P_{1-t} f(W_t)}{P_{1-t} f(W_t)}$. Thus $\{v_t\}$ is the analog of (8) in the discrete setting of Section II.

This has a straightforward geometric interpretation. Consider the relative density

$$\phi_t(x) = f(x) e^{-\|x - W_t\|^2 / 2(1-t)},$$

and let $\bar{\phi}_t(x)$ be the normalization of $\phi_t(x)$ such that $\bar{\phi}_t(x) \, dx$ is a probability density. Then,

$$v_t = (1-t)^{-1} \left(\int x \bar{\phi}_t(x) \, dx - W_t \right)$$

is the vector pointing from W_t to the center of mass of f with respect to a Gaussian distribution of variance $1-t$ centered at W_t. The scaling by $(1-t)^{-1}$ stands to reason: The fact that $W_1 \sim f d\gamma_n$ means that as t approaches 1, if W_t is far from the "bulk" of f, the desperation of the drift increases.

It is possible to show that equation (13) implies that for every $t \in [0,1]$,

$$\mathbb{E}\,\|v_t\|^2 = \int \frac{\|\nabla P_{1-t} f\|^2}{P_{1-t} f}\, d\gamma_n\,. \tag{14}$$

The latter quantity is the Fisher information of $P_{1-t}f$ (see [BGL14, Ch. II.5]). (It is the continuous analog of the "total influence" of $\sqrt{P_{1-t}f}$.) Thus v_t reflects the geometry of f seen from many "granularities." Given our discussion so far, it is difficult to avoid stating Lehec's elegant proof [Leh13] of the Gaussian log-Sobolev inequality on which Section II is based:

$$H_{\gamma_n}(f) \overset{(12)}{=} \frac{1}{2}\int_0^1 \mathbb{E}\,\|v_t\|^2\, dt \le \frac{1}{2}\mathbb{E}\,\|v_1\|^2 \overset{(14)}{=} \frac{1}{2}\int \frac{\|\nabla f\|^2}{f}\, d\gamma_n\,, \tag{15}$$

where the only inequality is an immediate consequence of the fact that v_t is a martingale.

Changes of measure and gradient ascent. Recall that our goal is to bound $\mathbb{P}(f(B_1) \in [\alpha, 2\alpha]) \ll 1/\alpha$. To this end, we will study the Doob martingale $P_{1-t}f(B_t)$. As just argued, it will be beneficial to consider instead the process $P_{1-t}f(W_t)$. At least intuitively, our change of measure was helpful: Since W_1 has law $f d\gamma_n$, we have

$$\mathbb{P}(f(W_1) \in [\alpha, 2\alpha]) \approx \alpha \cdot \mathbb{P}(f(B_1) \in [\alpha, 2\alpha])\,.$$

Thus it suffices to prove simply that $\mathbb{P}(f(W_1) \in [\alpha, 2\alpha]) \to 0$ as $\alpha \to \infty$. Now the story comes together, as Itô's formula will tell us that our process $P_{1-t}f(W_t)$ can be related directly to the drift $\{v_t\}$: For all $t \in [0,1]$,

$$P_{1-t}f(W_t) = \exp\left(\int_0^t \langle v_s, dB_s\rangle + \frac{1}{2}\int_0^t \|v_s\|^2\, ds\right).$$

As alluded to in the introduction, we will bound $\mathbb{P}(f(W_1) \in [\alpha, 2\alpha])$ by showing that if the former is large, then so is $\mathbb{P}\left(f(W_1) \in [2^k\alpha, 2^{k+1}\alpha]\right)$ for many values of k. This will be accomplished by perturbing the process $\{W_t\}$ to achieve these larger values, and then arguing that the measure of $\{W_t\}$ is relatively insensitive to such perturbations. The perturbed processes are essentially of the following form. For fixed $\delta > 0$, they are given by the stochastic differential equation

$$dW_t^\delta = dB_t + (1+\delta) v_t\, dt\,.$$

Girsanov's theorem tells us that there is a measure Q_δ under which W_t^δ is a Brownian motion, and furthermore that

$$\frac{dQ_\delta}{dQ} = \exp\left(-\delta\int_0^1 \langle v_t, dB_t\rangle - \left(\delta + \frac{\delta^2}{2}\right)\int_0^1 \|v_t\|^2\, dt\right).$$

This expresses the relative probability of Brownian motion having the sample path $\{W_t^\delta : t \in [0,1]\}$ vs. the sample path $\{W_t : t \in [0,1]\}$. From a high-level perspective, the most important factor here is $\exp\left(-\delta\int_0^1 \|v_t\|^2\, dt\right)$.

On the other hand, this loss in measure will be compensated by an increase in the value f. Here we employ the Hessian condition (3) to conclude that since $W_1^\delta = W_1 + \delta\int_0^1 v_t\, dt$,

$$f(W_1^\delta) \ge f(W_1)\exp\left(\delta\left\langle v_1, \int_0^1 v_t\, dt\right\rangle - \beta\delta^2\left\|\int_0^1 v_t\, dt\right\|^2\right),$$

where we have used the fact that $v_1 = \nabla \log f(W_1)$ from (13).

In order to accomplish our goal, we need that the loss of measure is almost exactly compensated for by the increase in the value of f. Ignoring another low-order term (which requires δ to be small), this necessitates that

$$\exp\left(-\delta \int_0^1 \|v_t\|^2 \, dt\right) \approx \exp\left(\delta \left\langle v_1, \int_0^1 v_t \, dt \right\rangle\right).$$

Now we use the martingale property of v_t. It implies immediately that

$$\mathbb{E}\left[\left\langle v_1, \int_0^1 v_t \, dt \right\rangle\right] = \int_0^1 \mathbb{E}\,\|v_t\|^2 \, dt\,.$$

Thus the last issue we need to address is the *concentration* of $\langle v_1, \int_0^1 v_t \, dt\rangle$ and how it interacts with the many details and lower-order terms that we have glossed over. Controlling this presents the bulk of the technical difficulties in the proof. We refer to [EL14] for the detailed arguments.

Acknowledgements

The authors would like to thank Elchanan Mossel for encouraging their collaboration, and Joseph Lehec for a careful reading of many early drafts of this manuscript, as well as numerous insightful comments. Both authors were partially supported by NSF grants CCF-1217256 and CCF-1407779.

Much of this work was done at Microsoft Research in Redmond where the vibrant environment and abundance of whiteboard markers were essential. We are also grateful to the Simons Institute in Berkeley for hosting us during the final portion of the project. R.E. acknowledges Jian Ding for introducing him to Talagrand's problem. J.L. points out that he introduced Jian to the problem, and thanks Ryan O'Donnell for enlightening preliminary discussions.

REFERENCES

[BBB+13] K. Ball, F. Barthe, W. Bednorz, K. Oleszkiewicz, and P. Wolff. L^1-smoothing for the Ornstein-Uhlenbeck semigroup. *Mathematika*, 59(1):160–168, 2013.

[BD98] Michelle Boué and Paul Dupuis. A variational representation for certain functionals of Brownian motion. *Ann. Probab.*, 26(4):1641–1659, 1998.

[BGL14] Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 348 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer, Cham, 2014.

[Bon70] Aline Bonami. Étude des coefficients de Fourier des fonctions de $L^p(G)$. *Ann. Inst. Fourier (Grenoble)*, 20(fasc. 2):335–402 (1971), 1970.

[Bor00] Christer Borell. Diffusion equations and geometric inequalities. *Potential Anal.*, 12(1):49–71, 2000.

[Bor02] Christer Borell. Isoperimetry, log-concavity, and elasticity of option prices. In *New directions in Mathematical Finance*, pages 73–91. Wiley, 2002. Edited by P. Wilmott and H. Rasmussen.

[Bre11] Haim Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York, 2011.

[BT06] Sergey G. Bobkov and Prasad Tetali. Modified logarithmic Sobolev inequalities in discrete settings. *J. Theoret. Probab.*, 19(2):289–336, 2006.

[DGS92] E. Brian Davies, Leonard Gross, and Barry Simon. Hypercontractivity: a bibliographic review. In *Ideas and methods in quantum and statistical physics (Oslo, 1988)*, pages 370–389. Cambridge Univ. Press, Cambridge, 1992.

[DW08] Ronald De Wolf. A brief introduction to fourier analysis on the boolean cube. *Theory of Computing, Graduate Surveys*, 1:1–20, 2008.

[EL14] R. Eldan and J. R. Lee. Regularization under diffusion and anti-concentration of temperature. *Preprint:* arXiv: 1410.3887, 2014.

[Eld13a] Ronen Eldan. Thin shell implies spectral gap up to polylog via a stochastic localization scheme. *Geom. Funct. Anal.*, 23(2):532–569, 2013.

[Eld13b] Ronen Eldan. A two-sided estimate for the gaussian noise stability deficit. *Preprint:* arXiv:1307.2781, 2013.

[Föl85] H. Föllmer. An entropy approach to the time reversal of diffusion processes. In *Stochastic differential systems (Marseille-Luminy, 1984)*, volume 69 of *Lecture Notes in Control and Inform. Sci.*, pages 156–163. Springer, Berlin, 1985.

[Gro75] Leonard Gross. Logarithmic Sobolev inequalities. *Amer. J. Math.*, 97(4):1061–1083, 1975.

[Gro06] Leonard Gross. Hypercontractivity, logarithmic Sobolev inequalities, and applications: a survey of surveys. In *Diffusion, quantum theory, and radically elementary mathematics*, volume 47 of *Math. Notes*, pages 45–73. Princeton Univ. Press, Princeton, NJ, 2006.

[GRS75] F. Guerra, L. Rosen, and B. Simon. The $\mathbf{P}(\phi)_2$ Euclidean quantum field theory as classical statistical mechanics. I, II. *Ann. of Math. (2)*, 101:111–189; ibid. (2) 101 (1975), 191–259, 1975.

[Leh13] Joseph Lehec. Representation formula for the entropy and functional inequalities. *Ann. Inst. Henri Poincaré Probab. Stat.*, 49(3):885–899, 2013.

[LY98] Tzong-Yow Lee and Horng-Tzer Yau. Logarithmic Sobolev inequality for some models of random walks. *Ann. Probab.*, 26(4):1855–1873, 1998.

[MOA11] Albert W. Marshall, Ingram Olkin, and Barry C. Arnold. *Inequalities: theory of majorization and its applications*. Springer Series in Statistics. Springer, New York, second edition, 2011.

[Mon12] Ashley Montanaro. Some applications of hypercontractive inequalities in quantum information theory. *J. Math. Phys.*, 53(12):122206, 15, 2012.

[MP10] Peter Mörters and Yuval Peres. *Brownian motion*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2010. With an appendix by Oded Schramm and Wendelin Werner.

[MT06] Ravi Montenegro and Prasad Tetali. Mathematical aspects of mixing times in Markov chains. *Found. Trends Theor. Comput. Sci.*, 1(3):x+121, 2006.

[Nel73] Edward Nelson. The free Markoff field. *J. Functional Analysis*, 12:211–227, 1973.

[O'D14] Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014.

[Øks03] Bernt Øksendal. *Stochastic differential equations*. Universitext. Springer-Verlag, Berlin, sixth edition, 2003. An introduction with applications.

[Tal89] Michel Talagrand. A conjecture on convolution operators, and a non-Dunford-Pettis operator on L^1. *Israel J. Math.*, 68(1):82–88, 1989.

[Vil03] Cédric Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.