CONCENTRATION INEQUALITIES

STEVEN P. LALLEY
UNIVERSITY OF CHICAGO

1. The Martingale Method

1.1. **Azuma-Hoeffding Inequality.**

Concentration inequalities are inequalities that bound probabilities of deviations by a random variable from its mean or median. Our interest will be in concentration inequalities in which the deviation probabilities decay exponentially or super-exponentially in the distance from the mean. One of the most basic such inequality is the *Azuma-Hoeffding* inequality for sums of *bounded* random variables.

Theorem 1.1. *(Azuma-Hoeffding) Let S_n be a martingale (relative to some sequence $Y_0, Y_1, \dots$) satisfying $S_0 = 0$ whose increments $\xi_n = S_n - S_{n-1}$ are bounded in absolute value by 1. Then for any $\alpha > 0$ and $n \geq 1$,*

$$P\{S_n \geq \alpha\} \leq \exp\{-\alpha^2/2n\}. \tag{1}$$

More generally, assume that the martingale differences ξ_k satisfy $|\xi_k| \leq \sigma_k$. Then

$$P\{S_n \geq \alpha\} \leq \exp\left\{-\alpha^2/2\sum_{j=1}^{n}\sigma_j^2\right\}. \tag{2}$$

In both cases the denominator in the exponential is the maximum possible variance of S_n subject to the constraints $|\xi_n| \leq \sigma_n$, which suggests that the worst case is when the distributions of the increments ξ_n are as spread out as possible. The following lemma suggests why this should be so.

Lemma 1.2. *Among all probability distributions on the unit interval $[0,1]$ with mean p, the most spread-out is the Bernoulli-p distribution. In particular, for any probability distribution F on $[0,1]$ with mean p and any convex function $\varphi : [0,1] \to \mathbb{R}$,*

$$\int_0^1 \varphi(x)\,dF(x) \leq p\varphi(1) + (1-p)\varphi(0).$$

Proof. This is just another form of Jensen's inequality. Let X be a random variable with distribution F, and let U be an independent uniform-[0,1] random variable. The indicator $1\{U \leq X\}$ is a Bernoulli r.v. with conditional mean

$$E(1\{U \leq X\} \mid X) = X,$$

and so by hypothesis the *unconditional* mean is $EE(1\{U \leq X\} \mid X) = EX = p$. Thus, $1\{U \leq X\}$ is Bernoulli-p. The conditional version of Jensen's inequality implies that

$$E(\varphi(1\{U \leq X\}) \mid X) \geq \varphi(E(1\{U \leq X\} \mid X)) = \varphi(X).$$

Taking unconditional expectation shows that

$$E\varphi(1\{U \leq X\}) \geq E\varphi(X),$$

which, since $1\{U \leq X\}$ is Bernoulli-p, is equivalent to the assertion of the lemma. □

Rescaling gives the following consequence (exercise).

Corollary 1.3. *Among all probability distributions on the interval $[-A, B]$ with mean zero, the most spread out is the two-point distribution concentrated on $-A$ and B that has mean zero. In particular, if φ is convex on $[-A, B]$ then for any random variable X satisfying $EX = 0$ and $-A \leq X \leq B$,*

$$E\varphi(X) \leq \varphi(-A)]\frac{B}{A+B} + \varphi(B)\frac{A}{A+B}. \tag{3}$$

In particular, if $A = B > 0$ then for any $\theta > 0$,

$$Ee^{\theta X} \leq \cosh(\theta A) \tag{4}$$

Proof. The first statement follows from Lemma 1.2 by rescaling, and the cosh bound in (4) is just the special case $\varphi(x) = e^{\theta x}$. □

Lemma 1.4. $\cosh x \leq e^{x^2/2}$.

Proof. The power series for $2\cosh x$ can be gotten by adding the power series for e^x and e^{-x}. The odd terms cancel, but the even terms agree, so

$$\cosh x = \sum_{n=0}^{\infty} \frac{x^{2n}}{(2n)!} \leq \sum_{n=0}^{\infty} \frac{x^{2n}}{2^n(n!)} = \exp\{x^2/2\}.$$

□

Conditional versions of Lemma 1.2 and Corollary 1.3 can be proved by virtually the same arguments (which you should fill in). Here is the conditional version of the second inequality in Corollary 1.3.

Corollary 1.5. *Let X be a random variable satisfying $-A \leq X \leq A$ and $E(X \mid Y) = 0$ for some random variable or random vector Y. Then for any $\theta > 0$,*

$$E(e^{\theta X} \mid Y) \leq \cosh(\theta A) \leq e^{\theta^2 A^2/2}. \tag{5}$$

Proof of Theorem 1.1. The first inequality (1) is obviously a special case of the second, so it suffices to prove (2). By the Markov inequality, for any $\theta > 0$ and $\alpha > 0$,

$$P\{S_n \geq \alpha\} \leq \frac{Ee^{\theta S_n}}{e^{\theta\alpha}}. \tag{6}$$

Since $e^{\theta x}$ is a convex function of x, the expectation on the right side can be bounded using Corollary 1.5, together with the hypothesis that $E(\xi_k | \mathscr{F}_{k-1}) = 0$. (As usual, the notation $\mathscr{F}_k$ is just shorthand for conditioning on $Y_0, Y_1, Y_2, \dots, Y_k$.) The result is

$$\begin{aligned} Ee^{\theta S_n} &= EE(e^{\theta S_n} \mid \mathscr{F}_{n-1}) \\ &\leq Ee^{\theta S_{n-1}} E(e^{\theta \xi_n} \mid \mathscr{F}_{n-1}) \\ &\leq Ee^{\theta S_{n-1}} \cosh(\theta \sigma_n) \\ &\leq Ee^{\theta S_{n-1}} \exp\{\theta^2 \sigma_n^2 / 2\} \end{aligned} \tag{7}$$

Now the same procedure can be used to bound $Ee^{\theta S_{n-1}}$, and so on, until we finally obtain

$$Ee^{\theta S_n} \leq \prod_{k=1}^{n} \exp\{\theta^2 \sigma_n^2 / 2\}$$

Thus,

$$P\{S_n \geq \alpha\} \leq e^{-\theta \alpha} \prod_{k=1}^{n} \exp\{\theta^2 \sigma_n^2 / 2\}$$

for *every* value $\theta > 0$. A sharp inequality can now be obtained by choosing the value of θ that minimizes the right side, or at least a value of θ near the min. A bit of calculus shows that the minimum occurs at

$$\theta = \alpha / \sum_{k=1}^{n} \sigma_k^2.$$

With this value of θ, the bound becomes

$$P\{S_n \geq \alpha\} \leq \exp\left\{-\alpha^2 / 2 \sum_{j=1}^{n} \sigma_j^2\right\}.$$

□

1.2. **McDiarmid's Inequality.** One of the reasons that the Azuma-Hoeffding inequality is useful is that it leads to concentration bounds for *nonlinear* functions of bounded random variables. A striking example is the following inequality of McDiarmid.

Theorem 1.6. *(McDiarmid) Let $X_1, X_2, \dots, X_n$ be independent random variables such that $X_i \in \mathscr{X}_i$, for some (measurable) sets $\mathscr{X}_i$. Suppose that $f : \prod_{i=1}^{n} A_i \to \mathbb{R}$ is "Lipschitz" in the following sense: for each $k \leq n$ and any two sequences $x, x' \in \prod_{i=1}^{n} \mathscr{X}_i$ that differ only in the kth coordinate,*

$$|f(x) - f(x')| \leq \sigma_k. \tag{8}$$

Let $Y = f(X_1, X_2, \dots, X_n)$. Then for any $\alpha > 0$,

$$P\{|Y - EY| \geq \alpha\} \leq 2 \exp\left\{-2\alpha^2 / \sum_{k=1}^{n} \sigma_k^2\right\}. \tag{9}$$

Proof. We will want to condition on the first k of the random variables X_i, so we will denote by $\mathscr{F}_k$ the $\sigma-$algebra generated by these r.v.s. Let

$$Y_k = E(Y \mid \mathscr{F}_k) = E(Y \mid X_1, X_2, \dots, X_k).$$

Then the sequence X_k is a martingale (by the "tower property" of conditional expectations). Furthermore, the successive differences satisfy $|Y_k - Y_{k-1}| \leq \sigma_k$. To see this, let $Y' = f(X')$, where

X' is obtained from X by replacing the kth coordinate X_k with an independent copy X'_k and leaving all of the other coordinates alone. Then

$$E(Y' \,|\, \mathscr{F}_k) = E(Y \,|\, \mathscr{F}_{k-1}) = Y_{k-1}$$

But by hypothesis, $|Y' - Y| \leq \sigma_k$. This implies that

$$|Y_k - Y_{k-1}| = |E(Y - Y' \,|\, \mathscr{F}_k)| \leq \sigma_k.$$

Given this, the result follows immediately from the Azuma-Hoeffding inequality, because $Y = E(Y \,|\, \mathscr{F}_n)$ and $EY = E(Y \,|\, \mathscr{F}_0)$. □

In many applications the constants σ_k in (8) will all be the same. In this case the hypothesis (8) is nothing more than the requirement that f be Lipschitz, in the usual sense, relative to the *Hamming metric* d_H on the product space $\prod_{i=1}^n \mathscr{X}_i$. (Recall that the Hamming distance $d_H(x,y)$ between two points $x,y \in \prod_{i=1}^n \mathscr{X}_i$ is just the number of coordinates i where $x_i \neq y_i$. A function $f : \mathscr{Y} \to \mathbb{Z}$ from one metric space $\mathscr{Y}$ to another $\mathscr{Z}$ is any function for which there is a constant $C < \infty$ such that $d_{\mathscr{Z}}(f(y), f(y')) \leq C d_{\mathscr{Y}}(y, y')$ for all $y, y' \in \mathscr{Y}$. The minimal such C is the *Lipschitz constant* for f.) Observe that for any set $A \subset \prod_{i=1}^n \mathscr{X}_i$, the distance function

$$d_H(x, A) := \min_{y \in A} d_H(x, y)$$

is itself Lipschitz relative to the metric d_H, with Lipschitz constant ≤ 1. Hence, McDiarmid's inequality implies that if X_i are independent $\mathscr{X}_i$−valued random variables then for any set $A \subset \prod_{i=1}^n \mathscr{X}_i$,

$$P\{|d_H(\mathbf{X}, A) - E d_H(\mathbf{X}, A)| \geq t\} \leq 2\exp\{-2t^2/n\}, \tag{10}$$

where $\mathbf{X} = (X_1, X_2, \dots, X_n)$. Obviously, if $x \in A$ then $d_H(x, A) = 0$, so if $P\{\mathbf{X} \in A\}$ is large then the concentration inequality implies that $E d_H(\mathbf{X}, A)$ cannot be much larger than $\sqrt{n}$. In particular, if $P\{\mathbf{X} \in A\} = \varepsilon > 0$ then (10) implies that

$$E d_H(\mathbf{X}, A) \leq \sqrt{-(n/2)\log(\varepsilon/2)}. \tag{11}$$

Substituting in (10) gives

$$P\{d_H(\mathbf{X}, A) \geq \sqrt{n}(t + \alpha)\} \leq 2\exp\{-2t^2\} \quad \text{where} \quad \alpha = \sqrt{-\frac{1}{2}\log(P\{\mathbf{X} \in A\}/2)}. \tag{12}$$

2. Gaussian Concentration

2.1. **McDiarmid's inequality and Gaussian concentration.**

McDiarmid's inequality holds in particular when the random variables X_i are Bernoulli, for *any* Lipschitz function $f : \{0,1\}^n \to \mathbb{R}$. There are lots of Lipschitz functions, especially when the number n of variables is large, and at the same time there are lots of ways to use combinations of Bernoullis to approximate other random variables, such as normals. Suppose, for instance, that $g : \mathbb{R}^n \to \mathbb{R}$ is continuous (relative to the usual Euclidean metric on $\mathbb{R}^n$) and Lipschitz in each variable separately, with Lipschitz constant 1 (for simplicity), that is,

$$|g(x_1, x_2, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n) - g(x_1, x_2, \dots, x_{i-1}, x'_i, x_{i+1}, \dots, x_n)| \leq |x_i - x'_i|. \tag{13}$$

Define $f:\{-1,1\}^{mn}\to\mathbb{R}$ by setting $f(y)=g(x(y))$ where $x(y)\in\mathbb{R}^n$ is obtained by summing the $y_i s$ in blocks of size m and scaling by $\sqrt{m}$, that is,

$$x(y)_k=\frac{1}{\sqrt{m}}\sum_{i=(k-1)m+1}^{km}y_i. \tag{14}$$

Since g is Lipschitz, so is f (relative to the Hamming metric on $\{-1,1\}^{mn}$), with Lipschitz constant $1/\sqrt{m}$. Therefore, McDiarmid's inequality applies when the random variables Y_i are i.i.d. Rademacher (i.e., $P\{Y_i=\pm1\}=1/2$). Now as $m\to\infty$ the random variables in (14) approach normals. Hence, McDiarmid implies a concentration inequality for Lipschitz functions of Gaussian random variables:

Corollary 2.1. *Let $X_1,X_2,..,X_n$ be independent standard normal random variables, and let $g:\mathbb{R}^n\to\mathbb{R}$ be continuous and Lipschitz in each variable separately, with Lipschitz constant 1. Set $Y=g(X_1,X_2,..,X_n)$. Then*

$$P\{|Y-EY|\geq t\}\leq 2e^{-2t^2/n}. \tag{15}$$

2.2. **The duplication trick.** You might at first think that the result of Corollary 2.1 should be a fairly tight inequality in the special case where g is Lipschitz with respect to the usual Euclidean metric on $\mathbb{R}^n$, because your initial intuition is probably that there isn't much difference between Hamming metrics and Euclidean metrics. But in fact the choice of metrics makes a huge difference: for functions that are Lipschitz relative to the Euclidean metric on $\mathbb{R}^n$ a much sharper concentration inequality than (15) holds.

Theorem 2.2. *(Gaussian concentration) Let γ be the standard Gaussian probability measure on $\mathbb{R}^n$ (that is, the distribution of a $N(0,I)$ random vector), and let $F:\mathbb{R}^n\to\mathbb{R}$ be Lipschitz relative to the Euclidean metric, with Lipschitz constant 1. Then for every $t>0$,*

$$\gamma\{F-E_\gamma F\geq t\}\leq\exp\{-t^2/\pi^2\} \tag{16}$$

Notice that the bound in this inequality does not depend explicitly on the dimension n. Also, if F is 1−Lipschitz then so are $-F$ and $F-c$, and hence (16) yields the two-sided bound

$$\gamma\{|F-E_\gamma F|\geq t\}\leq 2\exp\{-t^2/\pi^2\}. \tag{17}$$

In section **??** below we will show that the constant $1/\pi^2$ in the bounding exponential can be improved. In proving Theorem 2.2 – and a number of other concentration inequalities to come – we will make use of the following simple consequence of the Markov-Chebyshev inequality, which reduces concentration inequalities to bounds on moment generating functions.

Lemma 2.3. *Let Y be a real random variable. If there exist constants $C,A<\infty$ such that $Ee^{\lambda Y}\leq Ce^{A\lambda^2}$ for all $\lambda>0$, then*

$$P\{Y\geq t\}\leq C\exp\left\{\frac{-t^2}{4A}\right\}. \tag{18}$$

Proof of Theorem 2.2. This relies on what I will call the *duplication trick*, which is often useful in connection with concentration inequalities. (The particulars of this argument are due to Maurey and Pisier, but the duplication trick, broadly interpreted, is older.) The basic idea is to build an independent copy of the random variable or random vector that occurs in an inequality and somehow incorporate this copy in an expectation along with the original. By Lemma 2.3, to prove a concentration inequality it suffices to establish bounds on the Laplace transform $Ee^{\lambda F(X)}$. If $EF(X) = 0$, then Jensen's inequality implies that the value of this Laplace transform must be ≥ 1 for all values of $\lambda \in \mathbb{R}$. Consequently, if X' is an independent copy of X then for any $\lambda \in \mathbb{R}$,

$$E\exp\{\lambda F(X) - \lambda F(X')\} \leq Ce^{A\lambda^2} \quad \Longrightarrow \quad E\exp\{\lambda F(X)\} \leq Ce^{A\lambda^2}; \tag{19}$$

hence, to establish the second inequality it suffices to prove the first. If F is Lipschitz, or smooth with bounded gradient, the size of the difference $F(X) - F(X')$ will be controlled by $\mathrm{dist}(X, X')$, which is often easier to handle.

Suppose, then, that X and X' are independent $n-$ dimensional standard normal random vectors, and let F be smooth with gradient $|\nabla F| \leq 1$ and mean $EF(X) = 0$. (If (16) holds for smooth functions F with Lipschitz constant 1 then it holds for all Lipschitz functions, by a standard approximation argument.) Our objective is to prove the first inequality in (19), with $A = 1/2$. To accomplish this, we will take a smooth path X_t between $X_0 = X$ and $X_1 = X'$ and use the fundamental theorem of calculus:

$$F(X) - F(X') = \int_0^1 \nabla F(X_t)^T \frac{dX_t}{dt}\, dt$$

The most obvious path is the straight line segment connecting X, X', but it will be easier to use

$$\begin{aligned} X_t &= \cos(\pi t/2)X + \sin(\pi t/2)X' \quad \Longrightarrow \\ dX_t/dt &= -(\pi/2)\sin(\pi t/2)X + (\pi/2)\cos(\pi t/2)X' \\ &=: \frac{\pi}{2} Y_t \end{aligned}$$

because each X_t along this path is a standard normal random vector, and the derivative Y_t is also standard normal and independent of X_t. By Jensen's inequality (using the fact that the path integral is an average),

$$\begin{aligned} E\exp\{\lambda F(X) - \lambda F(X')\} &= E\exp\left\{\lambda \int_0^1 \nabla F(X_t)^T \frac{dX_t}{dt}\, dt\right\} \\ &\leq \int_0^1 E\exp\{(\lambda\pi/2)\nabla F(X_t)^T Y_t\}\, dt. \end{aligned}$$

For each t the random vectors Y_t and X_t are independent, so conditional on X_t the scalar random variable $\nabla F(X_t)^T Y_t$ is Gaussian with mean zero and variance $|\nabla F(X_t)|^2 \leq 1$. Consequently,

$$E\exp\{(\lambda\pi/2)\nabla F(X_t)^T Y_t\} \leq \exp\{\lambda^2\pi^2/4\}.$$

This proves that inequality (19) holds with $C = 1$ and $A = \pi^2/4$. Hence, for any $t > 0$ and λ

$$P\{F(X) \geq t\} \leq e^{-\lambda t} Ee^{\lambda F(X)} \leq e^{-\lambda t} e^{\lambda^2\pi^2/4}.$$

By Lemma 2.3, the concentration bound (16) follows. □

The preceding proof makes explicit use of the hypothesis that the underlying random variables are Gaussian, but a closer look reveals that what is really needed is rotational symmetry and *sub*-Gaussian tails. As an illustration, we next prove a concentration inequality for the uniform distribution $\nu = \nu_n$ on the $n-$sphere

$$\mathbb{S}^n := \{x \in \mathbb{R}^n : |x| = 1\}.$$

Theorem 2.4. *Let $\nu = \nu_n$ be the uniform probability measure on the unit sphere $\mathbb{S}^n$. There exist constants $C, A < \infty$ independent of n such that for any function $F : \mathbb{S}^n \to \mathbb{R}$ that is $1-$Lipschitz relative to the Riemannian metric on $\mathbb{S}^n$,*

$$\nu\{F - E_\nu F \geq t\} \leq Ce^{-nt^2/A} \quad \forall\ t > 0. \tag{20}$$

Proof. Once again, we use the duplication trick to obtain a bound on the moment generating function of F. Let U, U' be independent random vectors, each with the uniform distribution orthonormal basis $\mathbb{S}^n$, and let $\{U_t\}_{t\in[0,1]}$ be the shortest constant-speed geodesic path on $\mathbb{S}^n$ from $U_0 = U'$ to $U_1 = U$ (thus, the path follows the "great circle" on $\mathbb{S}^n$). Since the uniform distribution on the sphere is invariant under orthogonal transformations, for each fixed $t \in [0,1]$ the random vector U_t is uniformly distributed on $\mathbb{S}^n$. Moreover, $V_t :=$ the normalized velocity vector V_t to the path $\{U_t\}_{t\in[0,1]}$ (defined by $V_t = (dU_t/dt)/|dU_t/dt|$) is also uniformly distributed, and its conditional distribution given U_t is uniform on the $(n-2)-$dimensional sphere consisting of all unit vectors in $\mathbb{R}^n$ orthogonal to U_t. Consequently, if $N_t = \nabla F(U_t)/|\nabla F(U_t)|$ is the normalized gradient of F at the point U_t, then

$$\begin{aligned} E\exp\{\lambda F(U) - \lambda F(U')\} &= E\exp\{\lambda \int_0^1 \nabla F(U_t)^T (dX_t/dt)\, dt\} \\ &\leq \int_0^1 E\exp\{\lambda \nabla F(U_t)^T (dX_t/dt)\} \\ &\leq \int_0^1 E\exp\{A\lambda N_t^T V_t\}, \end{aligned}$$

where $0 < A < \infty$ is an upper bound on $|\nabla F(U_t)||dX_t/dt|$. (Here we use the hypothesis that F is $1-$Lipschitz relative to the Riemannian metric. The choice $A = \pi^2 =$ Riemannian diameter^2 of $\mathbb{S}^n$ will work.)

The rest of the argument is just calculus. For each t the normalized gradient vector $N_t = \nabla F(U_t)/|\nabla F(U_t)|$ is a fixed unit vector in the $(n-2)-$dimensional sphere of unit vectors in $\mathbb{R}^n$ orthogonal to U_t. But conditional on U_t the random vector V_t is uniformly distributed on this sphere. Consequently, the joint distribution of N_t and V_t is the same as the distribution of the first coordinate of a random vector uniformly distributed on the $(n-2)-$dimensional sphere in $\mathbb{R}^{n-1}$, and so for any $\lambda > 0$,

$$\begin{aligned} E\exp\{A\lambda N_t^T V_t\} &= \frac{2}{\pi}\int_{-\pi/2}^{\pi/2} e^{A\lambda \sin\theta} \cos^{n-2}\theta\, d\theta \\ &\leq \frac{4}{\pi}\int_0^{\pi/2} e^{A\lambda \sin\theta} \cos^{n-2}\theta\, d\theta \end{aligned}$$

Now use the crude bounds

$$\sin\theta \leq \theta \quad \text{and} \quad \cos\theta \leq 1 - B\theta^2$$

for a suitable constant $B > 0$ to conclude

$$\begin{aligned}\int_0^{\pi/2} e^{A\lambda \sin\theta} \cos^{n-2}\theta \, d\theta &\leq \int_0^{\pi/2} e^{A\lambda\theta}(1 - B\theta^2)^{n-2} \, d\theta \\ &\leq \int_0^{\pi/2} e^{A\lambda\theta} e^{-B(n-2)\theta^2} \\ &\leq \exp\{A\lambda^2/(2B(n-2))\}.\end{aligned}$$

This proves that

$$E\exp\{\lambda F(U) - \lambda F(U')\} \leq \frac{4}{\pi}\exp\{A\lambda^2/(2B(n-2))\},$$

and by Lemma 2.3 the concentration inequality (20) follows. □

2.3. **Johnson-Lindenstrauss Flattening Lemma.** The concentration inequality for the uniform distribution on the sphere has some interesting consequences, one of which has to do with *data compression.* Given a set of m data points in $\mathbb{R}^n$, an obvious way to try to compress them is to project onto a lower dimensional subspace. How much information is lost? If the only features in the original data points of interest are the pairwise distances, then the relevant measure of information is the maximal distortion of (relative) distance under the projection.

Proposition 2.5. *(Johnson-Lindenstrauss) There is a universal constant $D < \infty$ independent of dimension n such that the following is true. Given m points x_j in $\mathbb{R}^n$ and $\varepsilon > 0$, for any $k \geq D\varepsilon^{-2}\log m$ there exists a $k-$dimensional projection $A : \mathbb{R}^n \to \mathbb{R}^k$ that distorts distances by no more than $1 + \varepsilon$, that is, for any two points x_i, x_j in the collection,*

$$(1+\varepsilon)^{-1}|x_i - x_j| \leq \sqrt{n/k}|Ax_i - Ax_j| \leq (1+\varepsilon)|x_i - x_j|. \tag{21}$$

Furthermore, with probability approaching 1 as $m \to \infty$ the projection A can be obtained by choosing randomly from the uniform distribution on $k-$dimensional linear subspaces.

The uniform distribution on $k-$dimensional linear subspaces can be defined (and sampled from) using independent standard Gaussian random vectors $Y_1, Y_2, \dots, Y_k$ in $\mathbb{R}^n$. Let V be the linear subspace spanned by these k random vectors. With probability one, the subpsace V will be $k-$dimensional [Exercise: Prove this], and for any fixed orthogonal transformation $U : \mathbb{R}^n \to \mathbb{R}^n$ the distribution of the random subspace UV will be the same as that of V.

Lemma 2.6. *Let A be the orthogonal projection of $\mathbb{R}^n$ onto a random $k-$dimensional subspace. Then for every fixed $x \neq 0 \in \mathbb{R}^n$,*

$$P\left\{-\varepsilon\sqrt{\frac{k}{n}} \leq \left|\frac{|Ax|}{|x|} - \sqrt{\frac{k}{n}}\right| \leq \varepsilon\sqrt{\frac{k}{n}}\right\} \geq 1 - C\exp\{-B'k\varepsilon^2\} \tag{22}$$

for constants C', B' that do not depend on n, k, or the projection A.

Proof. The proof will rely on two simple observations. First, for a *fixed* orthogonal projection A, the mapping $x \mapsto |Ax|$ is $1-$Lipschitz on $\mathbb{R}^n$, so the concentration inequality (20) is applicable. Second, by the rotational invariance of the uniform distribution ν_n, the distribution of $|Ax|$ when A is fixed and x is random (with spherically symmetric distribution) is the same as when x is fixed and A is random. Hence, it suffices to prove (22) when A is a fixed projection

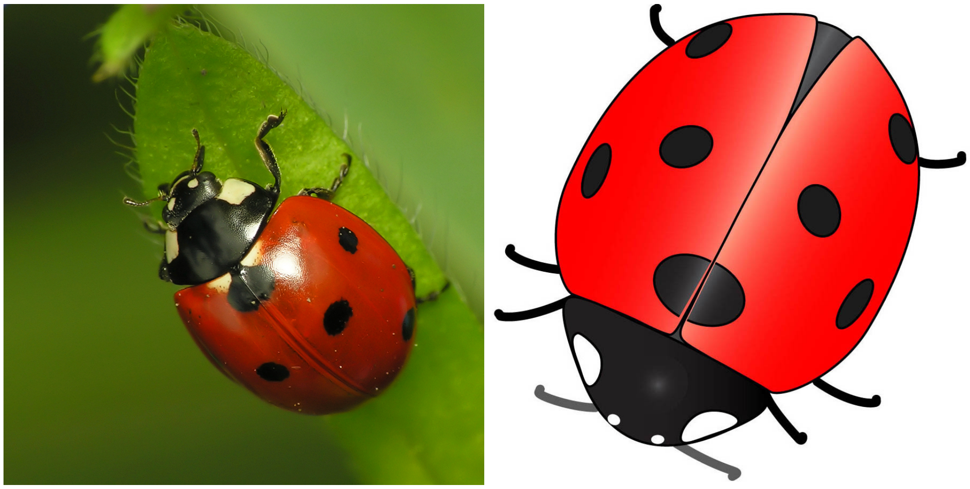

FIGURE 1. Data Compression by Orthogonal Projection

and x is chosen randomly from the uniform distribution on the unit sphere. Since $x \mapsto |Ax|$ is $1-$Lipschitz, the inequality (20) for the uniform distribution ν_n on $\mathbb{S}^n$ implies that for suitable constants $B, C < \infty$ independent of dimension, if $Z \sim \nu_n$

$$P\{||AZ| - E|AZ|| \geq t\} \leq Ce^{-Bnt^2}. \tag{23}$$

To proceed further we must estimate the distance between $\sqrt{k/n}$ and $E|AZ|$, where $Z \sim \nu$ is uniformly distributed on the sphere. It is easy to calculate $E|AZ|^2 = k/n$ (Hint: by rotational symmetry, it is enough to consider only projections A onto subspaces spanned by k of the standard unit vectors.) But inequality (23) implies that the variance of $|AZ|$ can be bounded, using the elementary fact that for a nonnegative random variable Y the expectation EY can be computed by

$$EY = \int_0^\infty P\{Y \geq y\}\, dy.$$

This together with (23) implies that

$$\operatorname{var}(|AZ|) \leq \int_0^\infty Ce^{-Bny}\, dy = \frac{C}{Bn}.$$

But $\operatorname{var}(|AZ|) = E|AZ|^2 - (E|AZ|)^2$, and $E|AZ|^2 = k/n$, so it follows that

$$\left|(E|A|)^2 - \frac{k}{n}\right| \leq \frac{C}{Bn} \quad \Longrightarrow$$
$$\left|E|A| - \sqrt{\frac{k}{n}}\right| \leq \frac{D}{\sqrt{nk}}$$

where $D = C/B$ does not depend on n or k. Using this together with (23) and the triangle inequality, one obtains that for a suitable B',

$$P\{||AZ| - \sqrt{k/n}| \geq t\} \leq P\{||AZ| - E|AZ|| \geq t - D/\sqrt{nk}\} \leq C\exp\{-Bn(t - D/\sqrt{nk})^2\}.$$

The substitution $t = \varepsilon\sqrt{k/n}$ now yields, for any $0 < \varepsilon < 1$,

$$P\{||AZ| - \sqrt{k/n}| \geq \varepsilon\sqrt{k/n}\} \leq C\exp\{-Bn(\varepsilon\sqrt{k/n} - D/\sqrt{nk})^2\} \leq C'\exp\{-B\varepsilon^2 k\}$$

for a suitable constant C'. □

Proof of Proposition 2.5. Let $\mathscr{X}$ be a set of m distinct nonzero points in $\mathbb{R}^n$, and let $\mathscr{Y}$ be the set of $\binom{m}{2}$ pairwise differences (which are all nonzero). Let A be the orthogonal projection onto a $k-$dimensional subspace of $\mathbb{R}^n$, and set $T = \sqrt{n/k}A$. For $y \in \mathscr{Y}$ say that T distorts in direction y if

$$||Ty| - |y|| \geq \varepsilon|y|.$$

Our aim is to show that if $k \leq \varepsilon^{-2}\log m$ and if A is chosen randomly from the uniform distribution on $k-$dimensional projections then with high probability there will be no $y \in \mathscr{Y}$ such that T distorts in direction y. Now by Lemma 2.6, for each $y \in \mathbb{Y}$ the probability that T distorts in direction y is bounded above by $C\exp\{-B'k\varepsilon^2\}$. Consequently, by the Bonferroni (union) bound, the probability that T distorts in the direction of *some* $y \in \mathscr{Y}$ is bounded by $C\binom{m}{2}\exp\{-B'k\varepsilon^2\}$. The proposition follows, because if $k \leq D\varepsilon^{-2}\log m$ then

$$C\tbinom{m}{2}\exp\{-B'k\varepsilon^2\} \leq C\binom{m}{2}m^{-B'D};$$

this converges to 0 as $m \to \infty$ provided $B'D > 2$. □

3. Geometry and Concentration

3.1. The Concentration Function.

Concentration inequalities for Lipschitz functions, such as McDiarmid's inequality and the Gaussian concentration inequality, can be reformulated in geometric terms, using the *concentration function* of the underlying probability measure.

Definition 3.1. Let μ be a Borel probability measure on a metric space $(\mathscr{X}, d)$. The *concentration function* of μ (relative to the metric d) is defined by

$$\alpha_\mu(r) := \sup\{\mu(A_r^c) : \mu(A) \geq \frac{1}{2}\} \quad \text{where} \tag{24}$$

$$A_r := \{x \in \mathscr{X} : d(x, A) \geq r\}. \tag{25}$$

Proposition 3.2. *Let $F : \mathscr{X} \to \mathbb{R}$ be Lipschitz, with Lipschitz constant C. If m_F is a median of F with respect to the Borel probability measure μ then*

$$\mu\{F \geq m_F + Ct\} \leq \alpha_\mu(t) \quad \textit{and} \tag{26}$$

$$\mu\{|F - m_f| \geq Ct\} \leq 2\alpha_\mu(t) \tag{27}$$

Proof. Let $A = \{F \leq m_F\}$. By definition of a median, $\mu(A) \geq 1/2$. The set $\{F \geq m_F + Ct\}$ is contained in A_t^c, since F has Lipschitz constant C, so (26) follows from the definition of the concentration function. □

Corollary 3.3. *For any two nonempty Borel sets $A, B \subset \mathscr{X}$,*

$$\mu(A)\mu(B) \leq 4\alpha_\mu(d(A,B)/2). \tag{28}$$

Proof. Let $F(x) = d(x, A)$ and $2r = d(A, B)$; then F is 1−Lipschitz, takes the value 0 on A, and satisfies $F \geq 2r$ on B. Let X, X' be independent $\mathscr{X}$−valued random variables each with distribution μ. Then

$$\begin{aligned}\mu(A)\mu(B) &= P\{X \in A;\, X' \in B\} \\ &\leq P\{F(X') - F(X) \geq 2r\} \\ &\leq 2P\{|F(X) - m_F| \geq r\} \\ &\leq 4\alpha_\mu(r).\end{aligned}$$

□

3.2. **Isoperimetric constants and concentration.** For any connected graph $G = (V, \mathscr{E})$ there is a natural metric d on the set V of vertices: for any two vertices x, y define $d(x, y)$ to be the length of the shortest path connecting x and y. A function $F : V \to \mathbb{R}$ is 1−Lipschitz relative to this metric if and only if $|F(x) - F(y)| \leq 1$ for any two nearest neighbors $x, y \in V$. The *Cheeger constant* (sometimes called the *isoperimetric ratio*) of G is defined to be

$$h_G = \inf\left\{\frac{|\partial A|}{|A|} : A \subset V, |A| \leq |V|/2, |A| < \infty\right\}, \tag{29}$$

where ∂A is the set of all vertices $x \notin A$ at distance 1 from A.

Proposition 3.4. *Let $G = (V, \mathscr{E})$ be a finite graph with Cheeger constant $h_G > 0$ and let ν be the uniform distribution on V. Then for every integer $m \geq 1$*

$$\alpha_\nu(m) \leq \frac{1}{2}(1 + h_G)^{-m}. \tag{30}$$

Proof. Let $A \subset V$ be a subset of cardinality $\geq |V|/2$, and let $B \subset V$ be such that $d(A, B) > m$, equivalently, $A \cap B_m = \emptyset$ where B_m is the set of all vertices at distance $\leq m$ from B. Thus, in particular $|B_m| \leq |V|/2$. By definition of the Cheeger constant, $|B_m| \geq (1 + h_G)^m |B|$, and hence

$$|B| \leq 2|V|(1 + h_G)^{-m}.$$

□

A family $G_n = (V_n, \mathscr{E}_n)$ of graphs is said to be an *expander family* with expansion constant ε if (a) there exists $d < \infty$ such that in each graph G_n every vertex has no more than d edges, and (b) the Cheeger constant of each G_n is at lest ε. Expanders are extremely important, both in computer science and in mathematics. For us, as probabilists, the most important feature of an expanders family is that the simple random walks on the graphs G_n in the family are rapidly mixing, in the sense that the number of steps necessary for the TV distance to uniformity to drop below e^{-1} is of order $O(|V_n| \log |V_n|)$.

Example: For each $n \geq d + 1$ let $\mathscr{G}_{n,d}$ be the set of all connected, d−regular graphs with vertex set $[n]$. Let G_n be chosen randomly from the uniform distribution on $\mathscr{G}_{n,d}$. Then with probability one, the family G_n is an ε−expander family for some $\varepsilon > 0$. □

The concentration inequality (30) implies that for any expander family, the concentration function of the uniform distribution is bounded above by a geometric distribution that does not depend on the size of the graph. (Recall that the standard Gaussian distributions on $\mathbb{R}^n$ also have this property.) This has some interesting consequences about the geometric structure of an expander graph. Suppose, for instance, that $G_n = (V_n, \mathscr{E}_n)$ is an expander family with $|V_n| \to \infty$, and for each G_n fix a distinguished vertex v_n^*. For each n define f_n on V_n by setting $f_n(x) = d(x, v_n^*)$; these functions are clearly 1−Lipschitz. Thus, if m_n is the median of f_n relative to the uniform distribution ν_n on V_n, then

$$\nu_n\{|f_n - m_n| \geq t\} \leq 2\alpha_{\nu_n}(t) \leq 4(1+\varepsilon)^{-t}.$$

In particular, nearly 100% of the vertices in V_n are at approximately the same distance from v_n^*.

3.3. **Reversible Markov chains and concentration.** The Cheeger constant “controls” the mixing rate of the simple random walk on a finite graph: the larger the Cheeger constant, the faster the simple random walk tends to equilibrium. (There are explicit bounds, but we won’t use these here.) This suggests that more generally the mixing rate of a Markov chain (or at any rate a reversible Markov chain) might be tied up with concentration properties of its stationary distribution. This is in fact the case, as we will show in section **??**. In this section we will show that the concentration function of the stationary distribution μ of a reversible Markov chain is also controlled by the *spectral gap* of the transition probability matrix. You may recall that the spectral gap is closely related to the the rate of mixing of a reversible Markov chain.

Assume now that μ is the stationary distribution of an ergodic, reversible Markov chain on a finite set $\mathscr{X}$. Denote by $\pi(x,y)$ or $\pi_{x,y}$ the one-step transition probabilities, and let $a(x,y) = a_{x,y} = \mu_x \pi_{x,y}$ be the *conductances* associated with the transition probabilities. The *Dirichlet form* associated with the Markov chain is the quadratic form on functions $f : \mathscr{X} \to \mathbb{R}$ defined by

$$\mathscr{D}(f,f) = \frac{1}{2}\sum\sum_{x,y} a_{x,y}(f(x) - f(y))^2, \tag{31}$$

that is, the quadratic form associated with the symmetric $\mathscr{X} \times \mathscr{X}$−matrix with entries $a_{x,y}$. Because the Markov chain is assumed to be ergodic, this matrix has a simple eigenvalue 0 (the associated eigenfunction is $f \equiv 1$), and all other eigenvalues are strictly positive. The smallest nonzero eigenvalue β_1 is called the *spectral gap*; it is determined by the variational formula

$$\beta_1 = \min_f \frac{\mathscr{D}(f,f)}{\mathrm{Var}_\mu(f)} \tag{32}$$

where the minimum is over all nonconstant functions $f : \mathscr{X} \to \mathbb{R}$ and $\mathrm{Var}_\mu(f)$ denotes the variance of the random variable f with respect to μ.

For any function $f : \mathscr{X} \to \mathbb{R}$, define the pseudo-Lipschitz norm

$$|||f|||_\infty = \max_x \sum_y (f(y) - f(x))^2 \pi_{x,y}.$$

Theorem 3.5. *For every $0 < \lambda < 2\sqrt{\beta_1}$ there exists $C_\lambda < \infty$ such that for any $f : \mathscr{X} \to \mathbb{R}$ with mean $E_\mu f = 0$ and pseudo-Lipschitz norm $|||f|||_\infty \leq 1$,*

$$\mu\{f \geq t\} \leq C_\lambda \exp\{-\lambda t\} \quad \forall\ t > 0. \tag{33}$$

NOTE. The constants C_λ depend only on the spectral gap β_1, so for any two reversible Markov chains with the same spectral gap the same bounds hold. It can be shown that $C_{\sqrt{\beta_1}/2} \leq 3$.

Proof. This argument is due to Aida and Stroock. The objective is to bound the moment generating function $\varphi(\lambda) = E_\mu e^{\lambda f}$; the concentration inequality will then follow by the Chebyshev-Markov inequality. The key to obtaining a bound here is that $\text{Var}_\mu(e^{\lambda f/2}) = \varphi(\lambda) - \varphi(\lambda/2)^2$, so the variational formula (32) will provide an upper bound for $\varphi(\lambda) - \varphi(\lambda/2)^2$ in terms of the Dirichlet form $\mathscr{D}(e^{\lambda f/2}, e^{\lambda f/2})$. The terms in the Dirichlet form can be partitioned into those for which $f(x) > f(y)$ and those for which $f(y) \geq f(x)$; by symmetry, and the convexity of the exponential function,

$$\begin{aligned}\mathscr{D}(e^{\lambda f/2}, e^{\lambda f/2}) &= \sum\sum\nolimits_{f(x)>f(y)} a_{x,y} e^{\lambda f(x)} (1 - \exp\{-\lambda(f(x) - f(y)/2)\})^2 \\ &\leq \sum\sum\nolimits_{f(x)>f(y)} \mu_x \pi_{x,y} e^{\lambda f(x)} \lambda^2 (f(x) - f(y))^2/4 \\ &\leq \sum\nolimits_x \mu_x e^{\lambda f(x)} |||f|||_\infty \lambda^2/4 \\ &= \lambda^2 \varphi(\lambda)/4\end{aligned}$$

Hence,

$$\begin{aligned}\varphi(\lambda) - \varphi(\lambda/2)^2 &= \text{Var}_\mu(e^{\lambda f/2}) \\ &\leq \beta_1^{-1} \mathscr{D}(e^{\lambda f/2}, e^{\lambda f/2}) \\ &\leq \frac{\lambda^2}{4\beta_1} \varphi(\lambda).\end{aligned}$$

Consequently, for $0 < \lambda < 2\sqrt{\beta_1}$,

$$\begin{aligned}\varphi(\lambda) &\leq (1 - \lambda^2/4\beta_1)\varphi(\lambda/2)^2 \quad \Longrightarrow \\ \varphi(\lambda) &\leq \left\{ \prod_{k=1}^{n} (1 - \lambda^2/4^k \beta_1)^{2^k} \right\} \varphi(\lambda/2^n)^{2^n}.\end{aligned}$$

Since $E_\mu f = 0$, by hypothesis, the derivative of φ at $\lambda = 0$ is 0. Thus, $\lim_{n\to\infty} \varphi(\lambda/2^n)^{2^n} = 1$, and so we obtain the bound

$$\varphi(\lambda) \leq \prod_{k=1}^{\infty} (1 - \lambda^2/4^k \beta_1)^{2^k} < \infty.$$

Set $C_\lambda = \varphi(\lambda)$ and use the Chebyshev-Markov inequality to obtain (33). □

3.4. **Mixing rates and concentration.** Let $\pi_n(x, y)$ be the $n-$step transition probabilities of an ergodic Markov chain on a finite or countable state space $\mathscr{X}$ with stationary probability distribution μ. If the state space is finite, then for any $0 < \alpha < 1$ there exists a (smallest) positive integer $m = m(\alpha)$, the $\alpha-$*mixing time*, such that for any two states $x, y \in \mathscr{X}$,

$$\pi_m(x, y) \geq \alpha\mu(y). \tag{34}$$

Say that a function $F : \mathscr{X} \to \mathbb{R}$ is $1-$Lipschitz (relative to the natural digraph structure of the Markov chain) if $|F(y) - F(x)| \leq 1$ for any two states x, y such that $p_1(x, y) > 0$.

Theorem 3.6. *Let F be $1-$Lipschitz with mean $E_\mu F = 0$, and let $m = m(\alpha)$ be an $\alpha-$mixing time (that is, an integer m for which the inequalities* (34) *hold). Then for any $\lambda > 0$ such that $\lambda m < -\log(1 - \alpha)$,*

$$E_\mu e^{\lambda F} \leq C e^{\lambda m} \quad \textit{where } C = \alpha/(1 - e^{\lambda m}(1 - \alpha)). \tag{35}$$

Consequently, for any $\lambda > 0$ such that $\lambda m < -\log(1-\alpha)$ there exists $C_\lambda < \infty$ such that

$$\mu\{F \geq t\} \leq C_\lambda e^{-\lambda t} \tag{36}$$

REMARK. The final inequality (36) indicates that fluctuations of F are of size $O(m)$ or smaller. The $\alpha-$mixing time will in general be larger than the total variation mixing time, which is more closely related to the geometric quantities (Cheeger constant and spectral gap) discussed earlier, so the concentration inequality (36) need not be as sharp as the earlier bounds. The main point of Theorem 3.6 is that there is a *direct* connection between the mixing properties of an ergodic Markov chain and the concentration function of its stationary distribution, and that this is so even for irreversible Markov chains.

Example 3.7. Consider the Ehrenfest random walk on the hypercube $\mathbb{Z}_2^n$. This is defined to be the Markov chain whose one-step moves are made by first choosing a coordinate $j \in [n]$ at random and then replacing the jth coordinate of the current configuration by a Bernoulli-1/2 random variable. The stationary distribution is the uniform distribution on $\mathbb{Z}_2^n$, and $\alpha-$mixing time is $C_\alpha + n\log n$ (Exercise: Check this.) Thus, in this case the concentration inequality (36) is weaker, by an extraneous factor of $\log n$ in the exponential, than McDiarmid's concentration inequality.

Proof of Theorem 3.6. By the duplication trick, to prove (35) it suffices to prove that if X, X' are independent, both with distribution μ, then

$$Ee^{\lambda F(X)-\lambda F(X')} \leq Ce^{\lambda m}.$$

Suppose that the Markov chain reached equilibrium in precisely m steps, that is that the inequalities (34) held with $\alpha = 1$. Then one could build independent X, X' with distribution μ by first choosing $X = X_0 \sim \mu$, then running the Markov chain X_j from the initial state X_0, and finally setting $X' = X_m$. It would then follow from the hypothesis that F is $1-$Lipschitz that for any $\lambda > 0$,

$$Ee^{\lambda F(X)-\lambda F(X')} = E\exp\{\lambda\sum_{j=1}^m (F(X_j) - F(X_{j-1}))\} \leq e^{\lambda m}.$$

Unfortunately, very few Markov chains reach equilibrium exactly in finitely many steps. Nevertheless, the foregoing argument suggests a way to proceed when there is an $\alpha > 0$ for which the $\alpha-$mixing time $m = m(\alpha)$ is finite. Assume that the underlying probability space is large enough to support independent X, X' and countably many independent $U[0,1]$ random variables for auxiliary randomization. Let ξ be Bernoulli-α, independent of X, X'. Set $X_0 = X$, and construct X_m by setting $X_m = X'$ when $\xi = 1$ and otherwise, on $\xi = 0$, choosing from the conditional distribution

$$P(X_m = y \,|\, X_0 = x, \xi = 0) = \frac{\pi_m(x,y) - \alpha\mu(y)}{1-\alpha}.$$

Complete the construction by choosing $X_1, X_2, \dots, X_{m-1}$ from the conditional distribution of the first $m-1$ steps of the Markov chain given the values of X_0 and X_m.

By construction, the joint distribution of X_0, X_m is the same as if the Markov chain were run for m steps beginning at the initial state X_0, that is,

$$P(X_0 = x, X_m = y) = \mu(x)\pi_m(x,y).$$

Hence, the *marginal* distribution of X_m is μ, because $X_0 \sim \mu$ and μ is the stationary distribution of the Markov chain. This implies that the *conditional* distribution of X_m given that $\xi = 0$ is also μ, because by construction the conditional distribution of X_m given $\xi = 1$ is μ. Therefore, setting

$$\varphi(\lambda) = Ee^{\lambda F(X) - \lambda F(X')},$$

we have

$$\begin{aligned}\varphi(\lambda) &= Ee^{\lambda F(X_0) - \lambda F(X_m)} e^{\lambda F(X_m) - \lambda F(X')} \\ &= Ee^{\lambda F(X_0) - \lambda F(X_m)} \mathbf{1}\{\xi = 1\} + Ee^{\lambda F(X_0) - \lambda F(X_m)} e^{\lambda F(X_m) - \lambda F(X')} \mathbf{1}\{\xi = 0\} \\ &\leq e^{\lambda m}(\alpha + (1-\alpha)\varphi(\lambda)).\end{aligned}$$

If $e^{\lambda m}(1-\alpha) < 1$ then this inequality implies that

$$\varphi(\lambda) \leq \alpha e^{\lambda m} / (1 - e^{\lambda m}(1-\alpha)).$$

□

4. Log-Sobolev Inequalities

4.1. **The Herbst argument.** Proving an exponential concentration inequality for a random variable Y is tantamount to obtaining a bound on the moment generating function $\varphi(\lambda) = Ee^{\lambda Y}$. One strategy for doing this is to bound the derivative $\varphi'(\lambda) = EYe^{\lambda Y}$ for each value of λ in a neighborhood of 0. This derivative is an entropy-like quantity: except for normalization, $\varphi'(\lambda)$ is just the usual Shannon entropy of the probability density proportional to $e^{\lambda Y}$.

Definition 4.1. Let $(\mathscr{X}, \mathscr{F}, \mu)$ be a probability space and let $f : \mathscr{X} \to [0, \infty]$ be a nonnegative, integrable function. Define the *entropy* of f relative to μ by

$$\operatorname{Ent}_\mu(f) = E_\mu(f \log f) - (E_\mu f)(\log E_\mu f). \tag{37}$$

Definition 4.2. A Borel probability measure μ on $\mathbb{R}^m$ is said to satisfy a *log-Sobolev inequality* with log-Sobolev constant C if for every smooth, bounded, compactly supported function $f : \mathbb{R}^m \to \mathbb{R}$,

$$\operatorname{Ent}_\mu(f^2) \leq C E_\mu |\nabla f|^2. \tag{38}$$

We will prove a bit later, for instance, that the standard Gaussian probability measure $\mu = \gamma$ on $\mathbb{R}^n$ satisfies a log-Sobolev inequality with log-Sobolev constant independent of dimension n (in fact, $C = 2$, but we will not prove this). It is also true that for any probability density on $\mathbb{R}^n$ of the form $\exp\{-U(x)\}$, where U is a smooth, convex function whose *Hessian matrix* (i.e., the matrix H of second partial derivatives) is such that $H - cI$ is positive definite, there is a log-Sobolev inequality with log-Sobolev constant c.

Theorem 4.3. *If μ satisfies a log-Sobolev inequality with log-Sobolev constant C then every $1-$Lipschitz function f is integrable with respect to μ and satisfies the concentration inequality*

$$\mu\{f \geq E_\mu f + t\} \leq \exp\{-t^2/4C\}. \tag{39}$$

Proof. This is due to Herbst. First, by Lemma 2.3, it suffices to prove that for any 1−Lipschitz function f with expectation $E_\gamma f = 0$,

$$Ee^{\lambda f} \le e^{C\lambda^2}.$$

To prove this, it suffices, by a routine truncation and smoothing argument, to prove it for bounded, smooth, compactly supported functions f such that $|\nabla f| \le 1$. Assume that f is such a function. Then for every real $\lambda \ge 0$, the log-Sobolev inequality

$$\mathrm{Ent}_\mu(e^{\lambda f}) \le CE_\mu|\nabla e^{\lambda f/2}|^2$$

can be written explicitly as

$$E_\mu \lambda f e^{\lambda f} - E_\mu e^{\lambda f} \log E_\mu e^{\lambda f} \le \frac{C\lambda^2}{4} E_\mu |\nabla f|^2 e^{\lambda f},$$

or alternatively, using the notation $\varphi(\lambda) = E^{\lambda f}$ and $\psi(\lambda) = \log\varphi(\lambda)$ for the moment generating function and cumulative generating function of f,

$$\begin{aligned}\lambda\varphi'(\lambda) &\le \varphi(\lambda)\log\varphi(\lambda) + \frac{C\lambda^2}{2} E_\mu|\nabla f|^2 e^{\lambda f} \\ &\le \varphi(\lambda)\log\varphi(\lambda) + \frac{C\lambda^2}{4}\varphi(\lambda).\end{aligned}$$

The last step uses the hypothesis that $|\nabla f| \le 1$. Dividing both sides by $\lambda\varphi(\lambda)$ gives

$$\lambda\psi'(\lambda) \le \psi(\lambda) + C\lambda^2/4.$$

This differential *inequality* implies that for $\lambda > 0$ the function $\psi(\lambda)$ remains below the solution (with the same initial value) of the differential *equation* obtained by replacing the ≤ by = (see the following lemma). But the solution of the differential equation $\lambda u'(\lambda) = u(\lambda) + C\lambda^2/4$ with initial condition $u(0) = 0$ is $u(\lambda) = C\lambda^2/4$; thus,

$$\psi(\lambda) \le \frac{C\lambda^2}{4} \quad \Longrightarrow \quad \varphi(\lambda) \le \exp\{C\lambda^2/4\}.$$

The concentration inequality (39) now follows by the usual argument. □

Lemma 4.4. *Let $g(s)$ and $h(s)$ be smooth, nonnegative functions of $s \ge 0$. Then any function $f(s)$ that satisfies $f(0) \ge 0$ and the differential inequality*

$$f'(s) \le g(s)f(s) + h(s) \tag{40}$$

for $s \ge 0$ is bounded above by the unique function $F(s)$ that satisfies $F(0) = f(0)$ and the differential equation

$$F'(s) = g(s)F(s) + h(s). \tag{41}$$

This remains true if $g(s)$ has a singularity at $s = 0$ such that $\lim_{s\to 0} sg(s) = \gamma > 0$ *provided $f(0) = F(0) = 0$.*

Proof. (Sketch) I will only sketch the proof of the first statement; the extension to singular cases is left as an exercise. If f satisfies (40) then by integration

$$f(t) - f(0) \le \int_0^t g(s)f(s)\,ds + \int_0^t h(s)\,ds.$$

This inequality can be iterated, and since the functions f, g, h are all nonnegative,

$$f(t) - f(0) \leq \int_0^t h(s)\,ds + \int_0^t g(s) \int_0^s h(r)\,dr\,ds + \cdots.$$

This series converges (exercise – see, e.g., the proof of Gronwall's inequality in my Stochastic Differential Equations notes). Furthermore, none of the terms on the right involves the function f, so the series converges to a function $F(x) - f(0)$ such that F satisfies (41). □

4.2. **Log-Sobolev inequalities on graphs.** Let's now return to the setting of reversible Markov chains on finite graphs. As in section 3.3, assume that μ is the stationary distribution of an ergodic, reversible Markov chain on a finite state space $\mathscr{X}$ with one-step transition probabilities $\pi(x, y)$, and let $\mathscr{D}$ be the corresponding Dirichlet form, defined by (31); thus, $\mathscr{D}$ is the quadratic form associated with the matrix $a_{x,y} = \mu_x \pi_{x,y}$ of conductances for the Markov chain.

Definition 4.5. The log-Sobolev constant $C > 0$ of the Markov chain with transition probabilities $\pi_{x,y}$ (or equivalently, the Dirichlet form with conductances $a_{x,y}$) is defined to be

$$C = \min \frac{\mathscr{D}(f, f)}{\mathrm{Ent}_\mu(f^2)} \tag{42}$$

where the minimum is over the set of all non-constant functions $f : \mathscr{X} \to \mathbb{R}$ and

$$\mathrm{Ent}_\mu(f^2) = \sum_{x \in \mathscr{X}} \mu(x) f(x)^2 \log(f(x)^2 / \|f\|_2^2) \quad \text{with } \|f\|_2^2 = \sum_{x \in \mathscr{X}} \mu(x) f(x)^2. \tag{43}$$

Theorem 4.6. *If C is the log-Sobolev constant of an ergodic, reversible Markov chain with Dirichlet form $\mathscr{D}$ and stationary distribution μ then for every function $f : \mathscr{X} \to \mathbb{R}$ with pseudo-Lipschitz norm $|||f|||_\infty \leq 1$,*

$$\mu\{f \geq E_\mu f + t\} \leq \exp\{-t^2/C\}. \tag{44}$$

Proof. This is nearly identical to the proof of Theorem 4.3, but using the Aida-Stroock bound on the Dirichlet form obtained in the proof of Theorem 3.6 in place of the simpler bound on the gradient used in Herbst's argument. Assume without loss of generality that $E_\mu f = 0$, and for $\lambda \in \mathbb{R}$ set $\varphi(\lambda) = E_\mu e^{\lambda f}$. Then by the Aida-Stroock argument (see the proof of Theorem 3.6),

$$\mathscr{D}(e^{\lambda f/2}, e^{\lambda f 2/}) \leq \lambda^2 \varphi(\lambda)/4.$$

Hence, for $\lambda > 0$ the log-Sobolev inequality for the function $e^{\lambda f/2}$ reads

$$\mathrm{Ent}_\mu(e^{\lambda f}) \leq C\mathscr{D}(e^{\lambda f/2}, e^{\lambda f 2/}) \leq C\lambda^2 \varphi(\lambda)/4.$$

The rest of the proof is identical to that of Theorem 4.3. □

Example 4.7. (Gross) Consider the special case of the 2−point space $\{-1, +1\}$ with the uniform distribution μ and the Dirichlet form

$$\mathscr{D}(f, f) = \frac{1}{4}(f(+1) - f(-1))^2 \tag{45}$$

corresponding to the trivial Markov chain on $\mathscr{X}$ with transition probabilities $\pi(x, y) = 1/2$ for all $x, y \in \mathscr{X}$. Let C be the log-Sobolev constant of μ with respect to $\mathscr{D}$. EXERCISE: Show that this is finite and if possible evaluate it. Hint: It is enough to consider functions of the form $f(-1) = a$

and $f(+1) = a+1$, by homogeneity. The answer is in Gross' original paper (Amer. J. Math. v. 97), cf. Theorem 3.

4.3. **Entropy.** By Theorems 4.3-4.6, log-Sobolev inequalities imply concentration inequalities. To make use of this principle we must have a way of proving log-Sobolev inequalities, and for this, some basic properties of entropy will be needed. To keep the discussion elementary we restrict attention to the case of discrete probability spaces; however, many of the results carry over without much change to continuous settings.

Definition 4.8. Let μ and ν be probability measures on a finite set $\mathscr{X}$. The *relative entropy* of ν relative to μ (also called the *information divergence* or the *Kullback-Leibler distance*, although it is not a metric) is defined by

$$\text{Ent}_\mu(\nu) = \sum_{x\in\mathscr{X}} \nu_x \log\frac{\nu_x}{\mu_x}. \tag{46}$$

More generally, if μ and ν are probability measures on a common measurable space $(\mathscr{X},\mathscr{F})$ such that ν is absolutely continuous with respect to μ, then

$$\text{Ent}_\mu(\nu) = E_\nu \log\frac{d\nu}{d\mu}. \tag{47}$$

If ν is not absolutely continuous with respect to μ then define $\text{Ent}_\mu(\nu) = \infty$.

Although the notation Ent_μ conflicts with the earlier use of Ent_μ, the meaning is the same: if f is a nonnegative function such that $E_\mu f = 1$ then f is the likelihood ratio (i.e., Radon-Nikodym derivative) of a probability measure ν with respect to μ, and $\text{Ent}_\mu(\nu) = \text{Ent}_\mu(f)$. The earlier definition (37) extends to all nonnegative, integrable functions f by homogeneity, that is, $\text{Ent}_\mu(f) = \|f\|_1 \text{Ent}_\mu(f/\|f\|_1)$.

In general, the relative entropy of ν relative to μ measures the rate of exponential decay of the likelihood ratio of ν to μ for a sequence of i.i.d. random variables $X_1, X_2, \ldots$ with common distribution ν: in particular, by the strong law of large numbers,

$$\exp\{\text{Ent}_\mu(\nu)\} = \lim_{n\to\infty}\left\{\prod_{i=1}^n \frac{d\nu}{d\mu}(X_i)\right\}^{1/n} \quad a.s.(\nu). \tag{48}$$

For this reason, it should be clear that $\text{Ent}_\mu(\nu) \geq 0$, with equality if and only if $\mu = \nu$. This can be proved formally using Jensen's inequality (exercise). Another way to characterize relative entropy is by the *Gibbs Variational Principle*, whose proof is left as another *exercise*:

$$\text{Ent}_\mu(\nu) = \max\{E_\nu g : E_\mu e^g = 1\}. \tag{49}$$

Consider now a product measure $\mu\times\nu$ on $\mathscr{X}\times\mathscr{Y}$, where μ,ν are probability measures on $\mathscr{X},\mathscr{Y}$ respectively. For simplicity assume that $\mathscr{X}$ and $\mathscr{Y}$ are finite sets. Let α,β be probability measures on $\mathscr{X}$ and $\mathscr{Y}$, respectively, and consider the set $\mathscr{P}(\alpha,\beta)$ of all probability measures $\lambda = (\lambda_{x,y})_{x,y\in\mathscr{X}\times\mathscr{Y}}$ with $\mathscr{X}-$ and $\mathscr{Y}-$marginals α and β.

Lemma 4.9. *The minimum relative entropy* $\text{Ent}_{\mu\times\nu}(\lambda)$ *over* $\lambda\in\mathscr{P}(\alpha,\beta)$ *is attained at the product measure* $\lambda = \alpha\times\beta$. *Thus, for any* $\lambda\in\mathscr{P}(\alpha,\beta)$,

$$\text{Ent}_{\mu\times\nu}(\lambda) \geq \text{Ent}_\mu(\alpha) + \text{Ent}_\nu(\beta), \tag{50}$$

and equality holds if and only if $\lambda = \alpha\times\beta$.

Proof. There are probably at least 5 different ways to do this. A mundane but straightforward approach is by calculus: (i) verify that the mapping $(\lambda_{x,y}) \mapsto \mathrm{Ent}_{\mu\times\nu}(\lambda)$ is a convex function of the vector λ, so the only critical points are minima; and (ii) use Lagrange multipliers to verify that the only critical points are at product measures. A longer but more illuminating argument uses the characterization (48) of relative entropy. For any probability measure $\lambda \in \mathscr{P}(\alpha,\beta)$ the margins are α and β, so if

$$(X_1,Y_1),(X_2,Y_2),\dots$$

are i.i.d. λ, then marginally $X_1,X_2,\dots,X_n$ are i.i.d. α and $Y_1,Y_2,\dots,Y_n$ are i.i.d. β. Consequently,

$$\begin{aligned}\left\{\prod_{i=1}^n \frac{d\lambda}{d\mu\times\nu}(X_i,Y_i)\right\}^{1/n} &= \left\{\prod_{i=1}^n \frac{d\alpha\times\beta}{d\mu\times\nu}(X_i,Y_i)\right\}^{1/n} \left\{\prod_{i=1}^n \frac{d\lambda}{\alpha\times\beta}(X_i,Y_i)\right\}^{1/n} \\ &= \left\{\prod_{i=1}^n \frac{d\alpha}{d\mu}(X_i)\right\}^{1/n} \left\{\prod_{i=1}^n \frac{d\beta}{d\nu}(Y_i)\right\}^{1/n} \left\{\prod_{i=1}^n \frac{d\lambda}{\alpha\times\beta}(X_i,Y_i)\right\}^{1/n} \\ &\to \exp\{\mathrm{Ent}_\mu(\alpha)+\mathrm{Ent}_\nu(\beta)+\mathrm{Ent}_{\alpha\times\beta}(\lambda)\}.\end{aligned}$$

The final exponential is minimized when $\lambda = \alpha\times\beta$, for which choice $\mathrm{Ent}_{\alpha\times\beta}(\lambda) = 0$. But

$$\exp\{\mathrm{Ent}_{\mu\times\nu}(\lambda)\} = \lim_{n\to\infty}\left\{\prod_{i=1}^n \frac{d\lambda}{d\mu\times\nu}(X_i,Y_i)\right\}^{1/n},$$

by (48), so the result (50) follows. □

Given a function $f:\mathscr{X}\times\mathscr{Y}\to\mathbb{R}$, denote by $f_x:\mathscr{Y}\to\mathbb{R}$ and $f_y:\mathscr{X}\to\mathbb{R}$ the functions obtained by "freezing" the first and second arguments of f, respectively, that is, $f_x(y) = f(x,y) = f_y(x)$. Write $f_{x+} = \sum_y f(x,y)$ and $f_{+y} = \sum_x f(x,y)$.

Corollary 4.10. *Let μ and ν be probability measures on $\mathscr{X}$ and $\mathscr{Y}$, respectively, and let f be a probability density on $\mathscr{X}\times\mathscr{Y}$ relative to $\mu\times\nu$. Then*

$$\mathrm{Ent}_{\mu\times\nu}(f) \le \sum_x \mu_x \mathrm{Ent}_\nu(f_x) + \sum_y \nu_y \mathrm{Ent}_\mu(f_y) \tag{51}$$

Proof. The right side of the inequality can be written as

$$\begin{aligned}\sum_x\sum_y \mu_x\nu_y f_x(y)\log(f_x(y)/f_{x+}) + \sum_y\sum_x \nu_y\mu_x f_y(x)\log(f_y(x)/f_{+y}) \\ = \sum_x\sum_y \mu_x\nu_y f_{x,y}\log(f_{x,y}^2/f_{x+}f_{+y}),\end{aligned}$$

and the left side of the inequality (51) is

$$\mathrm{Ent}_{\mu\times\nu}(f) = \sum_x\sum_y \mu_x\nu_y f_{x,y}\log f_{x,y}.$$

Hence, the inequality (51) is equivalent to

$$0 \le \sum_x\sum_y \mu_x\nu_y f_{x,y}\log(f_{x,y}/f_{x+}f_{+y}).$$

But this follows from Lemma 4.9. □

4.4. **Product measures and log-Sobolev inequalities.** Assume now that $\mathscr{X},\mathscr{Y}$ are finite sets endowed with Dirichlet forms $\mathscr{D}_X$ and $\mathscr{D}_Y$, each associated with a reversible, irreducible, aperiodic Markov chain. Let μ and ν be the stationary distributions of these chains. Define the *product* Dirichlet form $\mathscr{D} = \mathscr{D}_X \times \mathscr{D}_Y$ as follows: for any $f : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}$,

$$\mathscr{D}(f,f) = \sum_x \mu_x \mathscr{D}_Y(f_x, f_x) + \sum_y \mu_y \mathscr{D}_X(f_y, f_y). \tag{52}$$

Corollary 4.11. *Suppose that both μ and ν satisfy log-Sobolev inequalities relative to the Dirichlet forms $\mathscr{D}_X$ and $\mathscr{D}_Y$, with log-Sobolev constants C_X and C_Y, respectively. Then $\mu\times\nu$ satisfies a log-Sobolev inequality with respect to the product Dirichlet form $\mathscr{D}$ defined by* (52)*, with log-Sobolev constant* $\leq \max(C_X, C_Y)$*, that is, for every non-constant function $f : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}$,*

$$\mathrm{Ent}_{\mu\times\nu}(f^2) \leq \max(C_X, C_Y)\mathscr{D}(f,f). \tag{53}$$

Proof. This is an immediate consequence of Corollary 4.10 and the definition (52). □

NOTE. This clearly extends to products with more than two factors, by induction. Thus, in particular, if $\mathscr{D}^N$ is the product of N copies of a Dirichlet form $\mathscr{D}$ on $\mathscr{X}$, then the log-Sobolev constant C of $\mathscr{D}^N$ is *independent of the number N of factors.*

Example 4.12. (Ehrenfest Urn) Let $\mathscr{D}$ be the Dirichlet form on the two-point space $\mathscr{X} = \{-1, 1\}$ defined in Example 4.7, and consider the product Dirichlet form $\mathscr{D}^N$ on $\mathscr{X}^N$, as defined by (52), but with N factors instead of only 2. Then $N^{-1}\mathscr{D}^N$ is the Dirichlet form associated with the Ehrenfest random walk on the hypercube $\mathscr{X}^N$. (Note: The reason for the factor $1/N$ is that in the Ehrenfest random walk, at each time only one of the coordinate variables is reset.) Therefore, by Corollary 4.11, the log-Sobolev constant of the Ehrenfest chain is bounded above by CN, where C is the log-Sobolev constant of the two-point chain (Example 4.7). By Theorem 4.6, it follows that the uniform probability distribution $\mu = \mu_N$ on the hypercube $\{-1, 1\}^N$ satisfies a concentration inequality (44), with C replaced by CN. In particular, for every $f : \{-1, 1\}^N \to \mathbb{R}$ with pseudo-Lipschitz norm ≤ 1,

$$\mu_N\{f \geq E_{\mu_N} f + t\} \leq \exp\{-t^2/CN\}. \tag{54}$$

This should be compared with McDiarmid's inequality: McDiarmid gives a better constant C, but requires that the function f be $1-$Lipschitz, whereas (54) requires only that f be pseudo-Lipschitz.

4.5. **Log-Sobolev inequalities for Gaussian measures.**

Theorem 4.13. *The standard Gaussian probability measure γ on $\mathbb{R}^n$ satisfies a log-Sobolev inequality with log-Sobolev constant independent of dimension n.*

A number of different proofs are known. Perhaps the most elegant uses the fact that the Ornstein-Uhlenbeck process has a spectral gap independent of dimension, and uses this to deduce log-Sobolev via integration by parts. See the book by LEDOUX, ch. 5 for details. The proof to follow, due to Gross, avoids use of either stochastic calculus or semigroup theory, instead relying on the Ehrenfest random walk (Example 4.12) together with the central limit theorem.

Proof. We will deduce this from Example 4.12 and the central limit theorem. Consider first the case $n = 1$. Let $f : \mathbb{R} \to \mathbb{R}$ be a smooth function with compact support. For each $N \geq 1$ define a function $F : \{-1, 1\}^N \to \mathbb{R}$ on the $N-$dimensional hypercube by setting

$$F_N(x_1, x_2, \dots, x_N) = f\left(\frac{1}{\sqrt{N}} \sum_{j=1}^{N} x_j\right).$$

By Corollary 4.11 (see also Examples 4.7–4.12), there is a constant $C < \infty$ independent of N such that

$$\text{Ent}_{\mu^N}(F_N^2) \leq C\mathscr{D}^N(F_N, F_N)$$

where $\mathscr{D}^N$ is the product Dirichlet form for $\mathscr{D}$ = the Dirichlet form on the two point space $\{-1, 1\}$ and μ_N is the uniform distribution on $\{-1, 1\}^N$. By the central limit theorem, the distribution of $\sum_{i=1}^N x_i / N^{1/2}$ converges to the standard normal distribution γ, so

$$\lim_{N\to\infty} \text{Ent}_{\mu^N}(F_N^2) = \text{Ent}_\gamma f^2$$

(because $f^2 \log f^2$ is a bounded, continuous function). Thus, to establish the log-Sobolev inequality (38), it suffices to show that

$$\lim_{N\to\infty} \mathscr{D}^N(F_N, F_N) = \frac{1}{\sqrt{2\pi}} \int f'(y)^2 \exp\{-y^2/2\}\, dy.$$

For any sequence $x = (x_1, x_2, \dots, x_N) \in \{-1, 1\}^N$ denote by x^{i+} and x^{i-} the elements of $\{-1, 1\}^N$ obtained from x by replacing the entry x_i with $+1$ (for x^{i+}) or with -1 (for x^{i-}). Then

$$\begin{aligned}
D^N(F_N, F_N) &= 2^{-N} \sum_{x\in\{-1,1\}^N} \sum_{i=1}^{N} \frac{1}{4}(F_N(x^{i+}) - F_N(x^{i-}))^2 \\
&= 2^{-N} \sum_{x\in\{-1,1\}^N} \sum_{i=1}^{N} (N^{-1} f'(\sum_{i=1}^{n} x_i/\sqrt{N})^2 + o(N^{-1})) \\
&= 2^{-N} \sum_{x\in\{-1,1\}^N} f'(\sum_{i=1}^{n} x_i/\sqrt{N})^2 + o(1) \\
&\longrightarrow \frac{1}{\sqrt{2\pi}} \int f'(y)^2 \exp\{-y^2/2\}\, dy,
\end{aligned}$$

by the central limit theorem. □