# 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 $\xi_k$ 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 $\xi_n$ 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. By definition of convexity, for any $x \in [0,1]$,

$$\varphi(x) \leq (1-x)\varphi(0) + x\varphi(1).$$

Substituting $X$ for $x$ and taking expectations yields the desired inequality. □

This argument applies also to conditional distributions, and by translation and rescaling we may replace the unit interval by any other compact interval. We shall wish to use the inequality in the particular case where $\varphi(x) = e^{\theta x}$. In this case the lemma can be re-formulated in the following way.

**Lemma 1.3.** *Let $X$ be a random variable satisfying $|X| \le \sigma$ and $E(X \mid \mathcal{G}) = 0$. Then*

$$E(e^{\theta X} \mid \mathcal{G}) \le \cosh(\theta\sigma).$$

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

*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

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

□

*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 $\alpha > 0$,

$$P\{S_n \ge \alpha\} \le \frac{Ee^{\theta S_n}}{e^{\theta\alpha}}. \tag{3}$$

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

$$\begin{aligned} Ee^{\theta S_n} &= EE(e^{\theta S_n} \mid \mathcal{F}_{n-1}) \\ &\le Ee^{\theta S_{n-1}} E(e^{\theta \xi_n} \mid \mathcal{F}_{n-1}) \\ &\le Ee^{\theta S_{n-1}} \cosh(\theta\sigma_n) \\ &\le Ee^{\theta S_{n-1}} \exp\{\theta^2\sigma_n^2/2\} \end{aligned} \tag{4}$$

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

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

Thus,

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

for *every* value $\theta > 0$. A sharp inequality can now be obtained by choosing the value of $\theta$ that minimizes the right side, or at least a value of $\theta$ 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 $\theta$, the bound becomes

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

□

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 1.5.** *(McDiarmid) Let $X_1, X_2, \ldots, X_n$ 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 $k \le n$ and any two sequences $x, x' \in \prod_{i=1}^n \mathscr{X}_i$ that differ only in the $k$th coordinate,*

$$|f(x) - f(x')| \le \sigma_k. \tag{5}$$

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

$$P\{|Y - EY| \ge \alpha\} \le 2\exp\left\{-2\alpha^2 / \sum_{k=1}^n \sigma_k^2\right\}. \tag{6}$$

*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

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

Clearly, the sequence $X_k$ is a martingale (by the "tower property" of conditional expectations). Moreover, the successive differences satisfy $|Y_k - Y_{k-1}| \le \sigma_k$. To see this, we use the following general rule for calculating conditional expectations:

**Fact:** *If $U, V$ are random variables taking values in $\mathscr{U}, \mathscr{V}$, respectively, such that $U$ is $\mathscr{G}-$measurable and $V$ is independent of $\mathscr{G}$, then for any nonnegative (measurable) function $h : \mathscr{U} \times \mathscr{V} \to \mathbb{R}_+$,*

$$E(h(U,V) \mid \mathscr{G}) = H(U) \quad \textit{where} \quad H(x) = Eh(x,V).$$

Thus, if $Y' = f(X')$, where $X'$ is obtained from $X$ by replacing the $k$th coordinate $X_k$ with an independent copy $X_k'$ and leaving all of the other coordinates alone, then

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

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

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

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

In many applications the constants $\sigma_k$ in (5) will all be the same. In this case the hypothesis (5) 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 $x_i \ne y_i$. 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 $d_{\mathscr{Z}}(f(y), f(y')) \le C d_{\mathscr{Y}}(y, y')$ 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 $\le 1$. Hence, McDiarmid's inequality implies that if $X_i$ are independent $\mathscr{X}_i$–valued random variables then for any set $A \subset \prod_{i=1}^n \mathscr{X}_i$,

$$P\{|d_H(\mathbf{X}, A) - Ed_H(\mathbf{X}, A)| \ge t\} \le 2\exp\{-2t^2/n\}, \tag{7}$$

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 $Ed_H(\mathbf{X}, A)$ cannot be much larger than $\sqrt{n}$. In particular, if $P\{\mathbf{X} \in A\} = \varepsilon > 0$ then (7) implies that

$$Ed_H(\mathbf{X}, A) \le \sqrt{-(n/2)\log(\varepsilon/2)}. \tag{8}$$

Substituting in (7) gives

$$P\{d_H(\mathbf{X}, A) \ge \sqrt{n}(t + \alpha)\} \le 2\exp\{-2t^2\} \quad \text{where} \quad \alpha = \sqrt{-\frac{1}{2}\log(P\{\mathbf{X} \in A\}/2)}. \tag{9}$$

## 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,

$$|g(x_1, x_2, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n) - g(x_1, x_2, \dots, x_{i-1}, x_i', x_{i+1}, \dots, x_n)| \le |x_i - x_i'|. \tag{10}$$

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,

$$x(y)_k = \frac{1}{\sqrt{m}} \sum_{i=(k-1)m+1}^{km} y_i. \tag{11}$$

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., $P\{Y_i = \pm 1\} = 1/2$). Now as $m \to \infty$ the random variables in (11) 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*

$$P\{|Y - EY| \ge t\} \le 2e^{-2t^2/n}. \tag{12}$$

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 (12) holds.

**Theorem 2.2.** *(Gaussian concentration) Let $\gamma$ 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 $t>0$,*

$$\gamma\{F - E_\gamma F \geq t\} \leq \exp\{-t^2/\pi^2\} \tag{13}$$

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 (13) yields the two-sided bound

$$\gamma\{|F - E_\gamma F| \geq t\} \leq 2\exp\{-t^2/\pi^2\}. \tag{14}$$

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*

$$P\{Y\geq t\} \leq C\exp\left\{\frac{-t^2}{4A}\right\}. \tag{15}$$

*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 $\geq 1$ for all values of $\lambda\in\mathbb{R}$. Consequently, if $X'$ is an independent copy of $X$ then for any $\lambda\in\mathbb{R}$,

$$E\exp\{\lambda F(X) - \lambda F(X')\} \leq Ce^{A\lambda^2} \quad\Longrightarrow\quad E\exp\{\lambda F(X)\} \leq Ce^{A\lambda^2}; \tag{16}$$

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 $\operatorname{dist}(X,X')$, 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 (13) 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 (16), 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 better 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 (i) each $X_t$ along this path is a standard normal random vector; and (ii) the derivative $Y_t$ is also standard normal and *independent* of $X_t$. (The independence follows because $X_t$ and $Y_t$ are *uncorrelated*.) By Jensen's inequality (using the fact that the path integral is an average),

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

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 $|\nabla F(X_t)|^2 \le 1$. Consequently,

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

This proves that inequality (16) holds with $C = 1$ and $A = \pi^2/4$. Hence, for any $t > 0$ and $\lambda$

$$P\{F(X) \ge t\} \le e^{-\lambda t} E e^{\lambda F(X)} \le e^{-\lambda t} e^{\lambda^2\pi^2/4}.$$

By Lemma 2.3, the concentration bound (13) 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 $(n-1)-$sphere

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

**Theorem 2.4.** *Let $\nu = \nu_{n-1}$ 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 $\mathbb{S}^n$,*

$$\nu\{F - E_\nu F \ge t\} \le Ce^{-nt^2/A} \quad \forall\ t > 0. \tag{17}$$

The proof will require some preliminary results concerning the uniform distribution $\nu_m$ on the sphere $\mathbb{S}^m$. This is the unique probability measure supported by $\mathbb{S}^m$ that is invariant by orthogonal transformations of the ambient space $\mathbb{R}^{m+1}$; thus, if $Y$ is a random element of $\mathbb{S}^m$ with distribution $\nu_m$ and $U$ is an $(m+1)\times(m+1)$ orthogonal matrix then $UY$ also has distribution $\nu_m$. One way of sampling from the distribution $\nu_m$ exploits the rotational invariance of the standard normal distribution in $\mathbb{R}^{m+1}$: if $X$ is a standard normal random vector in $\mathbb{R}^{m+1}$, then for any orthogonal matrix $U$ the random vector $UX$ is also standard normal. Consequently, the

random *unit* vector $X/|X|$ must have distribution $\nu_m$, because its distribution is the same as that of $U(X/|X|)$.

**Lemma 2.5.** *Let $Y$ be a random vector in $\mathbb{R}^{m+1}$ with distribution $\nu_m$. Then the distribution of the* first coordinate *$Y_1$ has density $C_m(1-t^2)^{(m-2)/2}$ for $t \in [-1,1]$, where $C_m$ is a normalizing constant that satisfies*

$$C_m \sim \kappa\sqrt{m}$$

*as $m \to \infty$, where $\kappa > 0$ is a constant that does not depend on $m$.*

*Proof.* The first assertion, that the density of $Y_1$ is proportional to $(1-t^2)^{(m-2)/2}$, is a vector calculus exercise. The idea, roughly, is this. The probability that $Y_1$ lies in the interval $[t, t+dt]$ is the relative surface area (actually, $m-$dimensional volume) of the band

$$\{y \in \mathbb{S}^m : y_1 \in [t, t+dt]\}.$$

This band lies at angle $\arcsin(\sqrt{1-t^2})$ to the direction of the first coordinate axis; as $dt \to 0$ it approaches an $(m-1)-$dimensional sphere

$$\{y \in \mathbb{S}^m : y_1 = t\}.$$

This sphere has radius $\sqrt{1-t^2}$, and consequently $(m-1)-$dimensional surface volume $\alpha_{m-1}(1-t^2)^{(m-1)/2}$, where $\alpha_{m-1}$ is the volume (surface area?) of $\mathbb{S}^{m-1}$. Consequently,

$$P\{Y_1 \in [t, t+dt]\} = C_m \frac{(1-t^2)^{(m-1)/2}}{(1-t^2)^{1/2}},$$

as claimed.

The asymptotic formula for the normalizing constant $C_m$ follows by *Laplace's method* of asymptotic expansion. Here is a sketch. Clearly,

$$C_m^{-1} = \int_{-1}^{1} (1-t^2)^{(m-1)/2}\, dt = \int_{-1}^{1} \exp\{(m-1)\log(1-t^2)^{1/2}\}\, dt.$$

The integrand is the $(m-1)$th power of a nonnegative function that has its unique maximum at $t=0$; as $m \to \infty$, it decays to 0 exponentially fast uniformly over the region $|t| \geq \delta$, for any fixed $\delta > 0$. This implies that the integral over this region is exponentially small in $m$, and so is asymptotically negligible compared to $1/\sqrt{m}$. Thus, the main contribution comes from the region $|t| < \delta$, where the log can be arbitrarily well-approximated by its Taylor expansion. This leads to

$$C_m^{-1} \sim \int_{-1}^{1} \exp\{-(m-1)t^2/2\}\, dt$$

Now make the substitution $s = \sqrt{m-1}\,t$, and recognize that what's left is a standard normal density. □

*Proof of Theorem 2.4.* 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

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

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 $A = 2\pi$ diameter of $\mathbb{S}^n$ will work, I think.)

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 distribution of the inner product $N_t^T V_t$ 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 $\lambda > 0$,

$$\begin{aligned} E\exp\{A\lambda N_t^T V_t\} &= C_{n-2}\int_{-1}^1 e^{A\lambda t}(1-t^2)^{(n-4)/2}\, dt \\ &= C_{n-2}\int_{-\pi/2}^{\pi/2} e^{A\lambda \sin\theta}\cos^{n-3}\theta\, d\theta \\ &\le C_{n-2}\int_0^{\pi/2} e^{A\lambda \sin\theta}\cos^{n-3}\theta\, d\theta \end{aligned}$$

Now use the crude bounds

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

for a suitable constant $B > 0$ (the choice $B = 1/\pi$ should do) to conclude that

$$\begin{aligned} \int_0^{\pi/2} e^{A\lambda \sin\theta}\cos^{n-3}\theta\, d\theta &\le \int_0^{\pi/2} e^{A\lambda\theta}(1-B\theta^2)^{n-3}\, d\theta \\ &\le \int_0^{\pi/2} e^{A\lambda\theta} e^{-B(n-3)\theta^2}\, d\theta \\ &\le \int_{-\infty}^{\infty} e^{A\lambda\theta} e^{-B(n-3)\theta^2}\, d\theta \\ &= \frac{\sqrt{2\pi}}{\sqrt{2B(n-3)}}\exp\{A\lambda^2/(4B(n-3))\}. \end{aligned}$$

Together with the asymptotic formula for the normalizing constant $C_m$ obtained in Lemma 2.5, this proves that

$$E\exp\{\lambda F(U) - \lambda F(U')\} \le C'\exp\{A^2\lambda^2/(4B(n-3))\}$$

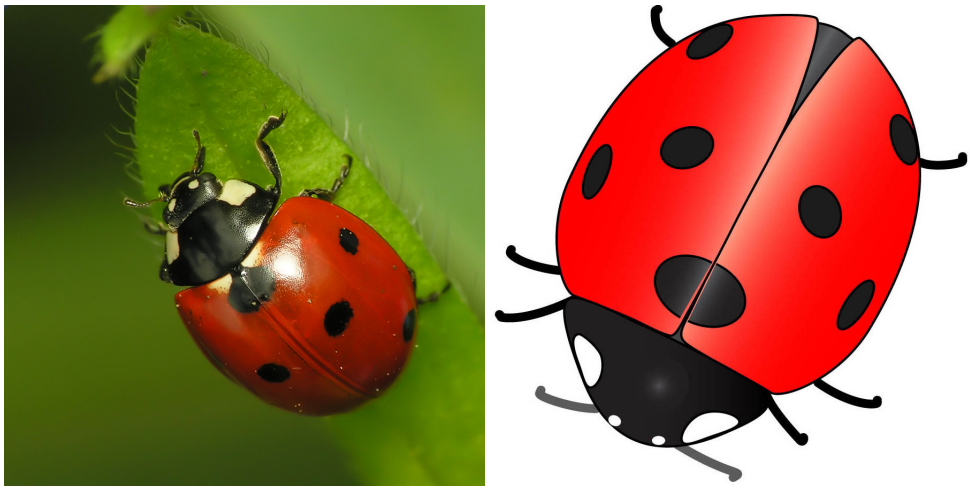

FIGURE 1. Data Compression by Orthogonal Projection

for constants $A, C'$ not depending on $m$. By Lemma 2.3 the concentration inequality (17) 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 2.6.** *(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,*

$$(1+\varepsilon)^{-1}|x_i - x_j| \leq \sqrt{n/k}|Ax_i - Ax_j| \leq (1+\varepsilon)|x_i - x_j|. \tag{18}$$

*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 $Y_1, Y_2, \ldots, Y_k$ 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 2.7.** *Let $A$ be the orthogonal projection of $\mathbb{R}^n$ onto a random $k-$dimensional subspace. Then for every fixed $x \neq 0 \in \mathbb{R}^n$,*

$$P\left\{-\varepsilon\sqrt{\frac{k}{n}} \leq \left|\frac{|Ax|}{|x|} - \sqrt{\frac{k}{n}}\right| \leq \varepsilon\sqrt{\frac{k}{n}}\right\} \geq 1 - C\exp\{-B'k\varepsilon^2\} \tag{19}$$

*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 (17) is applicable. Second, by the rotational invariance of the uniform distribution $\nu_n$, 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 (19) when $A$ is a fixed projection and $x$ is chosen randomly from the uniform distribution on the unit sphere. Since $x \mapsto |Ax|$ is 1−Lipschitz, the inequality (17) for the uniform distribution $\nu_n$ on $\mathbb{S}^n$ implies that for suitable constants $B, C < \infty$ independent of dimension, if $Z \sim \nu_n$

$$P\{||AZ| - E|AZ|| \geq t\} \leq Ce^{-Bnt^2}. \tag{20}$$

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 (20) 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 (20) implies that

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

But $\mathrm{var}(|AZ|) = E|AZ|^2 - (E|AZ|)^2$, 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 (20) 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 2.6.* 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 $k \le \varepsilon^{-2} \log m$ 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 2.7, 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 $k \le D\varepsilon^{-2} \log m$ then

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

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