1 Concentration in terms of metric spaces

In this course, a metric probability or m.p. space is a triple \((X, d, \mu)\) in which \((X, d)\) is a compact metric space and \(\mu \in \text{Prob}(X)\).

Our concentration results from Lecture 15 have the following consequence. If \(A\) is finite, \(p \in \text{Prob}(A)\), and \(d_n\) denotes normalized Hamming metric on \(A^n\), then the m.p. spaces \((A^n, d_n, p^{\times n})\) satisfy

\[
p^{\times n}\left\{ f \geq \int f \, dp^{\times n} + \varepsilon \right\}, \quad p^{\times n}\left\{ f \leq \int f \, dp^{\times n} - \varepsilon \right\} \leq e^{-\varepsilon^2 n/2}
\]

for all \(f \in \text{Lip}_1(A^n)\).

This concentration phenomenon is a distinctive feature of the sequence of m.p. spaces \((A^n, d_n, p^{\times n})\). Various other natural sequences of m.p. spaces exhibit the same phenomenon, and it often has other important consequences for these spaces. The most classical examples are high-dimensional spheres:

**Theorem 1** (Lévy’s concentration theorem). *Let \(S^n\) be the unit sphere in \(\mathbb{R}^{n+1}\), let \(d_n\) be the restriction of the Euclidean distance to \(S^n\), and let \(\mu_n\) be surface measure on \(S^n\). Then the sequence* \((S^n, d_n, \mu_n)\)

*exhibits exponential concentration.*

(Lévy’s actual theorem gives a precise solution to the isoperimetric problem on the sphere: for any value in \((0, 1)\), the optimal set of that measure is a spherical cap. This results in a precise rate of concentration on the sphere.)
At the end of Lecture 15 we turned (1) into an ‘isoperimetric’ result: if \( U \subseteq A^n \) has measure greater than \( e^{-\varepsilon^2 n/2} \) according to \( p^n \), then the \( (2\varepsilon) \)-neighbourhood \( B_{2\varepsilon}(U) \) fills up all but at most \( e^{-\varepsilon^2 n/2} \) of the measure. (In Lecture 15 we proved this only for the uniform measure on \( A^n \), but the argument is the same for any other product measure.) This isoperimetric point of view gives another roughly equivalent way to characterize concentration.

**Proposition 2.** Let \( (X_n, d_n, \mu_n) \) be a sequence of m.p. spaces all of diameter at most 1. Then the following are equivalent to each other:

1. For any \( \alpha \in (0, 1) \) and \( \delta > 0 \) we have
   \[
   \sup \left\{ \mu_n(B_\delta(A)^c) : A \in \mathcal{B}_{X_n} \text{ and } \mu_n(A) \geq \alpha \right\} \rightarrow 0,
   \]
   where \( \mathcal{B}_{X_n} \) is the Borel \( \sigma \)-algebra of \( X_n \) and \( B_\delta(A) \) is the \( \delta \)-neighbourhood of \( A \);
2. For any \( \varepsilon > 0 \) we have
   \[
   \sup \left\{ \mu_n \left\{ \left| f - \int f \, d\mu_n \right| > \varepsilon \right\} : f \in \operatorname{Lip}_1(X) \right\} \rightarrow 0.
   \]

Moreover, the following are also equivalent to each other:

1. For any \( \alpha \) and \( \delta \) the convergence is exponentially fast in \( n \) in part 1 above;
2. For any \( \varepsilon > 0 \) the convergence is exponentially fast in \( n \) in part 2 above.

A sequence of m.p. spaces satisfying either of the first two conditions above is said to **exhibit concentration** or to be a **Lévy family**. If either of the second two conditions is satisfied then they exhibit **exponential concentration**. For a given m.p. space, the supremum appearing in the first definition above is sometimes called the **concentration function** of the space (considered as a function of the parameters \( \alpha \) and \( \delta \)).

We have seen in (1) that that the sequence \( (A^n, d_n, p^n) \) exhibits exponential concentration. Lévy’s theorem gives the same for the sequence of spheres. Lévy’s name is attached to this phenomenon because his result is the first of this kind.

Other examples that exhibit exponential concentration include the symmetric groups \( S_n \) under an appropriate metric and the orthogonal and unitary groups under suitable Riemannian metrics. The concentration of the latter examples plays
a key role in modern random matrix theory. In this lecture we stick to the setting of product spaces, but explore some further formulations and consequence of the concentration phenomenon. The theory in this setting is the key to some results in ergodic theory later in the course.

Just as in probability theory, we can often prove concentration for a Lipschitz function on \((X, d, \mu)\) by bounding its moment generating function and applying Markov’s inequality. Since our concern here is with uniform estimates for all 1-Lipschitz functions, the relevant quantity is the so-called \textbf{Laplace transform} of the space \((X, d, \mu)\):

\[
\mathbb{E}_{(X, d, \mu)}(\lambda) := \sup \left\{ \int \exp \left[ \lambda \left( f - \int f \, d\mu \right) \right] \, d\mu : \ f \in \text{Lip}_1(X) \right\}
= \sup \left\{ \int e^{\lambda f} \, d\mu : \ f \in \text{Lip}_1(X) \text{ and } \int f \, d\mu = 0 \right\}.
\]

In these terms, we saw in the proof of Theorem 1 from Lecture 15 that

\[
\mathbb{E}_{(A^n, d_n, \hat{\nu}^n)}(\lambda n) \leq e^{\lambda^2 n/2}.
\]

\section{Transportation inequalities}

Our next topic is another formulation of measure concentration, this time in terms of transportation metrics.

To state the relevant inequalities in general m.p. spaces, we need a generalization of KL-divergence to that non-discrete probability spaces. If \((X, \mathcal{B}, \mu)\) is a general probability space and \(\nu\) is another probability measure on \((X, \mathcal{B})\) such that \(\nu \ll \mu\), then we define

\[
D(\nu \parallel \mu) := \int \frac{d\nu}{d\mu} \log \frac{d\nu}{d\mu} \, d\mu = \int \log \frac{d\nu}{d\mu} \, d\nu.
\]

This may be \(+\infty\) in case \(d\nu/d\mu\) fails to be in \(L^1(\mu)\). If \(\nu \ll \mu\), then we always have \(D(\nu \parallel \mu) = +\infty\) by convention. As in the case of finite sets, the quantity \(D(\nu \parallel \mu)\) is a way to quantify how singular is \(\nu\) relative to \(\mu\). It is non-negative, and zero if and only if \(\nu = \mu\), by essentially the same proofs as in the finite case.

**Definition 3.** An m.p. space \((X, d, \mu)\) satisfies the \textbf{entropy-transportation inequality with constant} \(C\), or \(T(C)\), if

\[
\overline{d}(\mu, \nu) \leq \sqrt{CD(\nu \parallel \mu)} \quad \forall \nu \in \text{Prob}(X).
\]
There are many possible variants here: for instance, one can try replacing the square root with another kind of dependence. But the square root turns out to be the correct choice in many cases of interest.

Intuitively, if \((X, d, \mu)\) satisfies \(T(C)\) for a small value of \(C\), then any other measure \(\nu\) which is not too singular to \(\mu\), as quantified by KL-divergence, can be transported to \(\mu\) itself for very little cost. The next lemma gives the simplest connection from this phenomenon to concentration.

**Lemma 4.** If \((X, d, \mu)\) satisfies \(T(C)\), and \(A \in \mathcal{B}(X)\) has \(\mu(A) \geq \alpha\), then

\[
\mu(B_\delta(A)^c) \leq \frac{1}{\delta} \sqrt{C \log \frac{1}{\alpha}}.
\]

**Proof.** Let \(\nu\) be the conditioned measure \(\mu(\cdot | A)\). Then \(D(\nu \| \mu) = \log \frac{1}{\mu(A)}\), and so \(T(C)\) provides some \(\lambda \in \text{Prob}(\mu, \nu)\) such that

\[
\int d \lambda \leq \sqrt{C \log \frac{1}{\mu(A)}} \leq \sqrt{C \log \frac{1}{\alpha}}.
\]

Now the fact that \(\nu(A) = \lambda(X \times A) = 1\) and Markov’s inequality give

\[
\mu(B_\delta(A)^c) = \lambda(B_\delta(A)^c \times X) = \lambda(B_\delta(A)^c \times A)
\]

\[
\leq \lambda\{(x, y) : d(x, y) \geq \delta\} \leq \frac{1}{\delta} \sqrt{C \log \frac{1}{\alpha}}.
\]

\[\square\]

Lemma 4 converts an entropy-transportation inequality into an estimate on the concentration function of an m.p. space. If one has a sequence of m.p. spaces \((X_n, d_n, \mu_n)\), and if they satisfy \(T(C_n)\) for some constants \(C_n \to 0\), then it follows that they also exhibit concentration. The same proof also gives a less quantitative but more general result along these lines:

**Proposition 5.** Consider a sequence \((X_n, d_n, \mu_n)\) of m.p. spaces. Suppose that for any other sequence \(\nu_n \in \text{Prob}(X_n)\) we have

\[
D(\nu_n \| \mu_n) = O(1) \quad \implies \quad d_n(\nu_n, \mu_n) \to 0.
\]

Then \((X_n, d_n, \mu_n)\) exhibits concentration.

\[\square\]
We have seen that the inequality $T(C)$ implies a certain amount of concentration, but a much more complete result is available in terms of the metric-space Laplace transform of $(X, d, \mu)$.

**Theorem 6** (Bobkov–Götze equivalence). An m.p. space $(X, d, \mu)$ satisfies $T(4C)$ if and only if
\[
\text{Ex}_{(X,d,\mu)}(\lambda) \leq e^{C\lambda^2} \quad \forall \lambda \geq 0.
\]

(3)

This proof combines two major results from earlier in the course. One is Monge–Kantorovich–Rubinstein duality, proved in Lecture 16. The other is the divergence-minimizing property of exponential families of measures. For measures on finite sets this was Theorem 1 in Lecture 11. The extension to other probability spaces requires no new ideas, but we also need the statement in a slightly different form, so we present it again here.

**Proposition 7.** Let $(X, \mathcal{B}, \mu)$ be a probability space and let $f \in L^\infty(\mu)$. Among other probability measures $\nu$ on $(X, \mathcal{B})$, the difference
\[
D(\nu \| \mu) = \int f \, d\nu
\]
is minimized uniquely by the measure
\[
\nu(dx) := \frac{1}{M_\mu(f)} e^{f(x)} \mu(dx), \quad M_\mu(f) := \int e^f \, d\mu.
\]
For this measure the difference equals $-\log M_\mu(f)$. ☐

**Proof of Theorem 6.** By Monge–Kantorovich–Rubinstein duality, $T(4C)$ is equivalent to
\[
\int f \, d\nu - \int f \, d\mu \leq 2\sqrt{C\text{D}(\nu \| \mu)} \quad \forall \nu \in \text{Prob}(X) \forall f \in \text{Lip}_1(X).
\]

Let us linearize the square-root here using the inequality of arithmetic and geometric means. This means we switch to the equivalent family of inequalities
\[
\int f \, d\nu - \int f \, d\mu \leq \frac{CD(\nu \| \mu)}{t} + t \quad \forall \nu \in \text{Prob}(X) \forall f \in \text{Lip}_1(X) \forall t > 0.
\]

Re-arranging and letting $\lambda := t/C$, this becomes
\[
D(\nu \| \mu) - \int (\lambda f) \, d\nu \geq - \int (\lambda f) \, d\mu - C\lambda^2 \quad \forall \nu \in \text{Prob}(X) \forall f \in \text{Lip}_1(X) \forall \lambda > 0.
\]
However, for fixed $f$ and $\lambda$, Proposition 7 shows that this holds for all $\nu$ if and only if it holds for the particular choice

$$\nu(dx) = \frac{1}{M_\mu(\lambda f)} e^{\lambda f(x)} \mu(dx).$$

Substituting this measure, our family of inequalities becomes

$$- \log M_\mu(\lambda f) \geq - \int (\lambda f) \, d\mu - C\lambda^2 \quad \forall f \in \text{Lip}_1(X) \, \forall \lambda > 0,$$

and hence

$$\log M_\mu(\lambda f) \leq \int (\lambda f) \, d\mu + C\lambda^2 \quad \forall f \in \text{Lip}_1(X) \, \forall \lambda > 0.$$

Exponentiating, we arrive at (3).

In view of (2), we immediately deduce the following.

**Theorem 8** (Transportation inequality in product spaces). For any finite set $A$, any $p \in \text{Prob}(A)$, and any $n \in \mathbb{N}$, the space $(A^n, d_n, p^n)$ satisfies $T(2/n)$.

This result is not quite optimal. As discussed in Lecture 16, the right-hand side of (2) can be improved to $e^{\lambda^2 n/8}$, and from this we can improve the conclusion of Theorem 8 to $T(1/2n)$.

### 3 A stability result for product measures

We finish this lecture with a special consequence of Theorem 8 in the setting of product spaces.

Let $A$ be a finite set, and let $\mu \in \text{Prob}(A)$. Let $\mu_1, \ldots, \mu_n \in \text{Prob}(A)$ be the marginals of $\mu$. Then the strict subadditivity of Shannon entropy gives

$$H(\mu) \leq \sum_{i=1}^n H(\mu_i),$$

and this is an equality if and only if $\mu = \times_{i=1}^n \mu_i$.

Our next result shows that if (4) is almost satisfied then $\mu$ is close to the product of its marginals. This result is sometimes referred to as the ‘stability’ of the inequality (4).

In fact, for the sake of an application later in the course, it is worth stating and proving a slightly more general result.
Theorem 9. Fix a finite set $A$. For every $\varepsilon > 0$ there is a $\delta > 0$, depending on $\varepsilon$ and $|A|$, such that the following holds. Let $p_1, \ldots, p_n \in \text{Prob}(A)$ and $\mu \in \text{Prob}(A^n)$, and suppose that

i) the marginals $\mu_1, \ldots, \mu_n$ of $\mu$ on $A$ satisfy

$$\sum_{i=1}^{n} \|\mu_i - p_i\| < \delta,$$

and

ii) we have

$$H(\mu) > H(p_1) + \cdots + H(p_n) - \delta.$$

Then

$$d_n(\mu, p_1 \times \cdots \times p_n) < \varepsilon.$$

Heuristically:

Product measures such as $p_1 \times \cdots \times p_n$ are approximately determined by their one-dimensional marginals and their total entropy.

Proof. Step 1. First we treat the special case when $p_i = \mu_i$ for every $i$. This case reduces to Theorem 8 by virtue of the following calculation:

$$D(\mu \parallel p_1 \times \cdots \times p_n) = \sum_{a=(a_1, \ldots, a_n) \in A^n} \mu(a) \log \frac{\mu(a)}{\mu_1(a_1) \cdots \mu_n(a_n)}$$

$$= \sum_{a} \mu(a) \log \mu(a) - \sum_{i=1}^{n} \sum_{a} \mu(a) \log \mu_i(a_i)$$

$$= \sum_{a} \mu(a) \log \mu(a) - \sum_{i=1}^{n} \sum_{a} \mu_i(a_i) \log \mu_i(a_i)$$

$$= \sum_{i=1}^{n} H(\mu_i) - H(\mu),$$

where we use the fact that $\mu_i$ is the $i^{th}$ marginal of $\mu$ in going from the second line to the third. (In case $n = 2$, this calculation is just the familiar identity $I(X ; Y) = D(p_{X,Y} \parallel p_X \times p_Y).$)
Thus, if $p_i = \mu_i$ for each $i$, then assumption (ii) gives $D(\mu \parallel \mu_1 \times \cdots \times \mu_n) < \delta n$, and Theorem 8 turns this into
\[
\overline{d}_n(\mu \parallel \mu_1 \times \cdots \times \mu_n) \leq \sqrt{\delta/2}. \tag{5}
\]

**Step 2.** Since $H : \text{Prob}(A) \rightarrow [0, \infty)$ is a continuous function on a compact set, it is uniformly continuous, and hence
\[
\|p - q\| \leq \delta \implies |H(p) - H(q)| \leq \Delta_A(\delta) \quad \forall p, q \in \text{Prob}(A). \tag{6}
\]
(A simple, explicit choice for $\Delta$ can be deduced from Fano’s inequality: exercise!)

By Markov’s inequality, the general form of assumption (i) implies that
\[
\left| \left\{ i = 1, 2, \ldots, n : \|\mu_i - p_i\| > \sqrt{\delta} \right\} \right| < \sqrt{\delta n}.
\]
Therefore (6) provides a function $\Delta_A(\delta)$ so that assumption (i) implies
\[
\left| \left\{ i = 1, 2, \ldots, n : |H(\mu_i) - H(p_i)| > \Delta_A(\delta) \right\} \right| < \Delta_A(\delta)n,
\]
and hence
\[
\sum_{i=1}^{n} |H(\mu_i) - H(p_i)| \leq \Delta_A(\delta)(n + 2 \log |A|) < \Delta_A(\delta)n
\]
(recalling that each appearance of $\Delta$ may stand for a different function).

Combined with condition (ii), this now implies that
\[
H(\mu) > \sum_{i=1}^{n} H(\mu_i) - \Delta_A(\delta)n,
\]
and hence
\[
\overline{d}_n(\mu, \mu_1 \times \cdots \times \mu_n) < \Delta_A(\delta), \tag{7}
\]
by Step 1.

On the other hand, a simple induction on $n$ gives
\[
\overline{d}_n(p_1 \times \cdots \times p_n, \mu_1 \times \cdots \times \mu_n) \leq \frac{1}{n} \sum_{i=1}^{n} \overline{d}_1(\mu_i, p_i) = \frac{1}{n} \sum_{i=1}^{n} \|\mu_i - p_i\| < \delta.
\]

Adding this to (7) and letting $\delta \downarrow 0$ completes the proof of Theorem 9. \qed
4 Notes and remarks

Many aspects of concentration as a metric-space phenomenon are covered in [Gro01, 
Chapter 3 1/2]. This lecture is influenced by the spirit of that treatment, although our 
goals here are more modest. Lévi’s theorem is in [Gro01, Section 3 1/2,19]. A gen-
eral discussion of concentration functions can be found in [Tal96]. The Laplace 
transform of a metric space is discussed in [Gro01, Section 3 1/2,62, part (3)].

The Bobkov–Götze equivalence is from [BG99]. That paper includes a nice 
introduction to some important related topics, particularly logarithmic Sobolev 
inequalities.

With the improved conclusion T(1/2n), Theorem 8 is due to Marton [Mar86, 
Mar96], but her proof is quite different. She does not use any result like Theo-
rem 6, but explicitly constructs a coupling that witnesses the desired bound on $\bar{d}$. 
The approach I have taken here is somewhat closer to Ledoux [Led97, Section 4], 
although Ledoux’s approach also explores the relation to to logarithmic Sobolev 
inequalities.

See the later sections of [Tal95] for other, more complicated relations between 
entropy and notions of transportation.

References


[Gro01] Mikhail Gromov. *Metric Structures for Riemannian and Non-

[Led97] Michel Ledoux. On Talagrand’s deviation inequalities for product mea-


**TIM AUSTIN**

**Email:** tim@math.ucla.edu  
**URL:** math.ucla.edu/~tim