OPTIMAL TRANSPORT AND CONCENTRATION OF MEASURE

NATHAEL GOZLAN[*],
TRANSCRIBED BY BARTŁOMIEJ POLACZYK

Abstract. This is a preliminary version of the notes of the lectures on the optimal transport and concentration of measure that were given during the 2nd Warsaw Summer School in Probability on July 2017. We introduce here the notion of concentration of measure and develop its connections with some functional inequalities via the methods of optimal transport theory. At the end we turn to the convex concentration property and weak transport costs.

Any comments or corrections are welcome. In case of such, please contact notes' author at polaczyk.b@gmail.com.

1. Concentration of measure phenomenon

1.1. **Preliminaries.**

Concentration of measure phenomenon states in general that a random variable that depends on many i.i.d. variables via some Lipschitz function is essentially concentrated around its mean. This phenomenon has many consequences in various fields of probability theory or statistics.

Throughout the notes we will denote by $(\mathcal{X}, d, \mu)$ an abstract metric space equipped with a distance d and a probability measure μ on some sigma algebra Σ. Moreover, we set $\mathcal{P}(\mathcal{X})$ to be the set of all probability measures on $\mathcal{X}$. We will start with the core definition of this lecture.

Definition. *Let $\alpha : [0, +\infty) \to [0, 1]$ be non increasing. We say that a probability measure $\mu \in \mathcal{P}(\mathcal{X})$ satisfies concentration of measure property with the concentration profile α if*

$$\mu(A) \geq \frac{1}{2} \Rightarrow \mu(A_t) \geq 1 - \alpha(t) \quad \forall_{A \subseteq \mathcal{X}} \forall_{t \geq 0},$$

where $A_t := \{x \in \mathcal{X} : d(x, A) \leq t\}$. We will denote this fact by $\mu \in C_\alpha(\mathcal{X}, d)$.

Example. *Some popular profiles are of the form*

$$\alpha_k(t) = a \exp\left(-b(t - t_0)_+^k\right),$$

for $k \in \mathbb{N}, a, b, t_0 \in \mathbb{R}$. Setting $k = 1, 2$ results in exponential and Gaussian profile respectively.

Proposition 1.1. *The following statements are equivalent:*

1. *$\mu \in C_\alpha(\mathcal{X}, d)$,*
2. *$\mu(f > m_\mu(f) + t) \leq \alpha(t) \quad \forall_{f \in Lip_1}(\mathcal{X}, d)$.*

Here

$$m_\mu(f) := \arg\inf_{t \in \mathbb{R}} \left\{ \mu(f \geq t) \leq \frac{1}{2} \right\}$$

[*] Universit Paris Descartes; nathael.gozlan@univ-mlv.fr.

is a median of f w.r.t. μ and $Lip_k(\mathcal{X}, d)$ is a space of all k-Lipschitz functions on $\mathcal{X}$ w.r.t. the metric d. In general will write Lip_k if there is no confusion about the underlying space and the metric.

Proof. (1. $\Rightarrow$ 2.) :
Set $A := \{f \leq m_\mu(f)\}$ for some arbitrary function $f \in Lip_1$. Clearly $\mu(A) \geq \frac{1}{2}$ and using the fact that A is closed

$$A_t = \{x : d(x, A) \leq t\} = \{x : \exists_{y\in\mathcal{X}} : \quad d(x,y) \leq t \wedge f(y) \leq m_\mu(f)\} \subseteq \{f \leq m_\mu(f) + t\},$$

whence

$$\mu(f > m_\mu(f) + t) \leq 1 - \mu(A_t) \leq \alpha(t).$$

(2. $\Rightarrow$ 1.) :
Let $A \subseteq \mathcal{X}$ such that $\mu(A) \geq \frac{1}{2}$ and set $f(x) = d(x, A)$. Whence, $f \in Lip_1$ by the triangle inequality and $m_\mu(f) = 0$, thus

$$\mu(A_t) = \mu(f \leq t) = 1 - \mu(f > m_\mu(f) + t) \geq 1 - \alpha(t).$$

□

Remark 1.2. *From the proof of Proposition 1.1 it follows that f has to be Lipschitz only in the support of μ.*

Remark 1.3. *If $\mu \in C_\alpha(\mathcal{X}, d)$ then applying the above proposition for $-f$ results in the following inequality*

$$\mu(|f - m_\mu(f)| > t) \leq 2\alpha(t), \tag{1}$$

which holds for any $f \in Lip_1$.

Depending on the situation, it can sometimes be easier to work with a mean instead of a median. Fortunately, concentration inequalities involving the median are, up to the constant, equivalent to the ones with the mean which illustrates the following remark.

Remark 1.4 (Mean vs. median)**.** *Let $\mu \in C_\alpha(\mathcal{X}, d)$. Integrating (1) with respect to t results in*

$$|\mu(f) - m_\mu(f)| \leq \int |f - m_\mu(f)|\, d\mu \overset{(1)}{\leq} 2 \int \alpha(t)\, dt =: t_0,$$

from which it follows that

$$\mu(f > \mu(f) + t) \leq \alpha(t - t_0),$$

where $\mu(f) := \int_{\mathcal{X}} f\, d\mu$ is a mean of f w.r.t. the measure μ.

1.2. **Concentration inequalities.** Let us focus now on some important results in the concentration theory, beginning with the result due to Hoeffding [16].

Proposition 1.5 (Hoeffding's inequality)**.** *Let X_i be an i.i.d sequence with a compact support enclosed in the interval $[a, b]$. It follows that*

$$\mathbb{P}\left(\bar{X} \geq \mathbb{E}[X_1] + t\right) \leq \exp\left(-2n \frac{t^2}{(b-a)^2}\right),$$

were $\bar{X} = \frac{1}{n} \sum X_i$.

The above result can be seen as an easy consequence of the following proposition.

Proposition 1.6 (Hoeffding's concentration inequality)**.** *Let $\mu \in \mathcal{P}(\mathcal{X})$ be a probability measure on the space $\mathcal{X}$ and let $\mu^{\otimes n} := \bigotimes_{j=1}^{n} \mu$ be the product measure. Then $\mu^{\otimes n} \in C_{\alpha_n}(\mathcal{X}^n, d_H)$, where $\alpha_n = \exp(-2t^2/n)$ and d_H is the Hamming distance on $\mathcal{X}^n$ given by the formula*

$$d_H(x,y) = \frac{1}{n} \left| \{ i : x_i \neq y_i \} \right| .$$

Remark 1.7. *Proposition 1.6 implies Proposition 1.5.*

Proof. Set $\mu = \mathcal{L}(X_1)$ be the law of X_1. Then $f(x) = \bar{x}$ is $\frac{b-a}{n}$ Lipschitz on the support of $\mu^{\otimes n}$ (recall Remark 1.2) w.r.t. d_H. Applying the assumption that $\mu^{\otimes n} \in C_{\alpha_n}(\mathcal{X}, d_H)$ yields the proof. □

We now turn to the proof of Proposition 1.6.

Proof of Proposition 1.6. The idea is to make use of exponential Chebyshev's inequality for $F(X_1, \ldots, X_n)$ where X_i are an i.i.d. sample, $X_1 \sim \mu$ and F is some 1-Lipschitz function w.r.t. the metric d_H. To that end, we need an estimate on the Laplace transform $\mathbb{E}e^{tF(X_1,\ldots,X_n)}$, which can be obtained using conditioning and some basic properties of the moment generating function.

Let us split the proof into separate parts.

Part 1. Firstly, let Z be a centered random variable on the interval $[-1,1]$. And let $\phi(t)$ denote its moment generating function, i.e. $\phi(t) = \log \mathbb{E}e^{tZ}$. Since

$$\phi'(t) = \frac{\mathbb{E}Ze^{tZ}}{\mathbb{E}e^{tZ}}, \qquad \phi''(t) = \frac{\mathbb{E}Z^2e^{tZ}}{\mathbb{E}e^{tZ}} - \left(\frac{\mathbb{E}Ze^{tZ}}{\mathbb{E}e^{tZ}} \right)^2 \leq \frac{\mathbb{E}Z^2e^{tZ}}{\mathbb{E}e^{tZ}},$$

then $\phi'(0) = 0$ and $\phi''(t) \leq 1$ from which it follows that

$$\mathbb{E}e^{tZ} \leq e^{t^2/2}. \tag{2}$$

Part 2. Let $\mathbb{E}_{X_k}$ denote integration w.r.t. the variable X_k, that is

$$\mathbb{E}_{X_k}[f(X_1, \ldots, X_n)] = \mathbb{E}[f(X_1, \ldots, X_n) | X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_n]$$

for any measurable function f. Note that since $F \in Lip_1(\mathcal{X}^n, d_H)$, we have $|F(x) - F(y)| \leq 1$ and thus $F(X_1, \ldots, X_n) - \mathbb{E}F$ is a centered random variable on the interval $[-1,1]$, where $\mathbb{E}F := \mathbb{E}F(X_1, \ldots, X_n)$. We can now estimate

$$\begin{aligned}
\mathbb{E}\left[e^{tF(X_1,\ldots,X_n)} \right] &= \mathbb{E}_{X_1,\ldots,X_{n-1}} \mathbb{E}_{X_n} \left[e^{tF(X_1,\ldots,X_n)} \right] \\
&\overset{(2)}{\leq} \mathbb{E}_{X_1,\ldots,X_{n-1}} \left[\exp\left(t^2/2 + \mathbb{E}_{X_n} tF(X_1, \ldots, X_n) \right) \right] \\
&\ \ \vdots \\
&\leq \exp\left(nt^2/2 + t\mathbb{E}F. \right)
\end{aligned}$$

Part 3. Finally, it suffices to apply Chebyschev's exponential inequality

$$\mathbb{P}\left(F(X_1, \ldots, X_n) \geq \mathbb{E}F + t \right) \leq \mathbb{E}e^{s(F(X_1,\ldots,X_n) - \mathbb{E}F) - st} \leq \exp\left(ns^2/2 - st \right).$$

Optimizing with respect to s yields the proof. □

While Proposition 1.6 is much stronger than Proposition 1.5, it still suffers weak dimension scaling which demonstrates the following example.

Example. *Take any $f \in Lip_1(\mathcal{X}, d_H)$. By Proposition 1.6*

$$\mu(f > m_\mu(f) + t) \leq \exp(-2t^2).$$

But on the other hand, setting $f_n : \mathcal{X}^n \to \mathbb{R}$, $f_n(x) = f(x_1)$ results in

$$\mu^{\otimes n}(f_n > m_{\mu^{\otimes n}}(f) + t) \leq \exp(-2t^2/n),$$

which is clearly not optimal since

$$\mu^{\otimes n}(f_n > m_{\mu^{\otimes n}}(f) + t) = \mu(f > m_\mu(f) + t).$$

The example above gives a motivation to introduce a notion of a concentration that is not dependent on the dimension.

Definition (dimension-free concentration property)**.** *A measure μ on $(\mathcal{X}, d)$ satisfies the dimension-free concentration property with a concentration profile α, denoted as $\mu \in C_\alpha^\infty(\mathcal{X}, d)$, if $\mu^{\otimes n} \in C_\alpha(\mathcal{X}^n, d_2)$ for every $n \geq 1$, where*

$$d_2(x,y) = \left(\sum d(x,y)^2\right)^{1/2}.$$

Obviously $C_\alpha^\infty(\mathcal{X}, d) \subseteq C_\alpha(\mathcal{X}, d)$. Converse is false which illustrates the following observation left as an exercise.

Exercise 1.1. *Show that if μ satisfies the dimension-free concentration property, then* $\operatorname{supp}\mu$ *is connected.*

A prominent example of a measure that satisfies the dimension-free concentration property is Gaussian measure on $\mathbb{R}^d$.

Theorem 1.8 (Borell[4], Sudakov – Tsirelson[31])**.** *Setting γ to be standard Gaussian measure on $\mathbb{R}$, ϕ – its c.d.f. and $\bar{\phi}(t) = 1 - \phi(t) = \gamma\left(]t, +\infty[\right)$ we have $\gamma \in C_{\bar{\alpha}}^\infty(\mathbb{R}, |\cdot|)$, where $\alpha = \bar{\phi}$.*

Remark 1.9. *The concentration profile $\bar{\phi}$ is optimal, which can be seen by plugging $f(x) = x$ into the equivalent definition of the concentration property (see Proposition 1.1).*

The following exercise proves that obtained concentration profile is in fact subgaussian.

Exercise 1.2. *Show that for $t > 0$*

1. $\bar{\phi}(t) \leq \frac{1}{2}\exp(-t^2/2)$,
2. $\phi(t) \leq \sqrt{\frac{1}{2\pi}}\frac{1}{t}\exp(-t^2/2)$.

The proof of Theorem 1.8 we are about to show is due to G. Piser [26]. It is much less technical than the original proof but provides a weaker concentration profile.

Proof of Theorem 1.8 for $\alpha(t) = e^{-2t^2/\pi^2}$. In the same manner as in the previous proof we will try to arrive at some estimate on the Laplace transform of a pushforward of γ^n by some Lipschitz function and then make use of Chebyshev's inequality. The main idea behind the proof is to exploit Jensen's inequality

$$\mathbb{E}e^{t(F(X) - \mathbb{E}F)} \leq \mathbb{E}e^{t(F(X) - F(Y))} = (\star),$$

where $X, Y \sim \mathcal{N}(0, Id)$ are independent standard Gaussian vectors in $\mathbb{R}^n$ and $F \in Lip(\mathbb{R}^n, |\cdot|)$, where $|\cdot|$ stands for the standard Euclidean metric. Set now

$$X_\theta = X\cos\theta + Y\sin\theta, \quad Y_\theta = \tfrac{d}{d\theta}X_\theta = Y\cos\theta - X\sin\theta,$$

for $\theta \in [0, \frac{\pi}{2}]$. Both X_θ and Y_θ are Gaussian vectors and by checking the covariance, one can see that they are independent as well. Moreover, $X_0 = X$ and $X_{\pi/2} = Y$, thus again exploiting Jensen's inequality we arrive at

$$(\star) = \mathbb{E}\exp\left(-t\int_0^{\pi/2} \nabla F(X_\theta)Y_\theta\, d\theta\right) \leq \mathbb{E}\fint_0^{\pi/2} \exp\left(-\frac{\pi t}{2}\nabla F(X_\theta)Y_\theta\right) d\theta = (\clubsuit).$$

Now, using independence of X_θ and Y_θ and conditioning, we can recognize in the above expression the Laplace transform of Y_θ, whence

$$(\clubsuit) = \fint_0^{\pi/2} \mathbb{E}\exp\left(\frac{\pi^2 t^2}{8}\left|\nabla F(X_\theta)\right|^2\right) d\theta \leq \exp\left(\frac{\pi^2 t^2}{8}\right),$$

where in the last inequality we have used the fact, that F is 1-Lipschitz. To conclude the proof it suffices to apply and optimize Chebyshev's inequality in the same manner as in the proof of Proposition 1.6. □

We will now present other examples of the concentration phenomenon.

Theorem 1.10 (see [34], Theorem 22.14)**.** *Let (M, d) be a (complete, connected) Riemannian manifold equipped with a measure $\nu(dx) = e^{-V(x)}dx$ for some $V \in C^2(M)$, such that*

$$\nabla^2 V + \mathrm{Ric} \geq cId$$

on M for some positive c. Then $\mu \in C_\alpha^\infty(M, d)$ for $\alpha = e^{-ct^2/2}$.

Exercise 1.3. *Consider uniform measure μ on the sphere (S^{n-1}, d) equipped with Euclidean metric. Then $\mu \in C_{\alpha_n}(S^{n-1}, d)$, where $\alpha_n(t) = \exp(-(n-1)t^2/2)$.*
Hint: *Consider $\mu = \mathcal{L}(X/\left|X\right|)$ for $X \sim \mathcal{N}(0, Id)$.*

Example (Non-Gaussian concentration)**.** *Set*

$$\mu(dx) \sim \exp\left(-x^p/p\right) dx \in \mathcal{P}(\mathbb{R})$$

and

$$\alpha_p(t) = a_p \exp(-b_p t)$$

for some $a_p, b_p > 0$. Then

1. *$\mu_p \in C_{\alpha_p}^\infty(\mathbb{R}, |\cdot|)$ for $p \in [1, 2]$,*
2. *$\mu_p \in C_{\alpha_p}^\infty(\mathbb{R}, \|\cdot\|_p)$ for $p > 0$.*

1.3. **Typical applications of concentration.** One of the main areas of studying in probability theory is the behavior of sequences $(X_n)_{n\in\mathbb{N}}$ of random variables taking values in some Polish space $\mathcal{X}$. In that area, there are two approaches (extreme to each other): (1) exact study, in which one tries to calculate the exact distribution of X_n for any n, and (2) asymptotic approach, in which one tries to establish some convergence results of (normalized) sequence X_n. From this perspective, the concentration of measure approach can be seen as in between the previous two, trying to estimate the rate of deviation of some Lipschitz statistics of the sequence sample from its mean.

To be more precise, let $(X_i)_{i\in\mathbb{N}}$ be an i.i.d. sequence in $\mathcal{X}$. Consider for each n the random empirical measure

$$L_n^X = \frac{1}{n}\sum_{i=1}^n \delta_{X_i},$$

where δ_a is Dirac mass at point $a \in \mathbb{R}$. Strong law of large numbers imply that

$$\mathbb{P}\left(L_n^X \rightharpoonup \mu\right) = 1, \tag{3}$$

where $\rightharpoonup$ denotes weak convergence and $\mu = \mathcal{L}(X_1)$. There are various ways to metricize the weak convergence of measures. The one for which the concentration theory proves useful is by the so-called bounded Lipschitz metric

$$d_{BL}(\mu,\nu) = \sup_{f \in Lip_1, \|f\|_\infty \le 1} \left\{ \int f\, d\mu - \int f\, d\nu \right\},$$

which is a well known result due to Varadarajan (1958, [32]). For a proof see e.g. [2], Theorem 8.3. Since $d_{BL}(\cdot,\mu)$, treated as a function from $\mathcal{X}^n \to \mathbb{R}$ is $\frac{2}{n}$-Lipschitz with respect to Hamming distance d_H, applying Hoeffding inequality (Proposition 1.5) one arrives at

$$\mathbb{P}\left(d_{BL}(L_n^X,\mu) \ge \mathbb{E} d_{BL}(L_n^X \mu) + t\right) \le \exp\left(-nt^2/2\right),$$

which holds for all $t \ge 0$. Thus, if we knew the bound on $\mathbb{E}d_{BL}$, then we would know the speed of convergence of (3). Some bounds on $\mathbb{E}d_{BL}$ are known (see [3]).

There is also another metric that metricizes the weak convergence of measures that we will deal with in the subsequent chapters, namely the Kantorovich power metric[1]

$$W_p(\mu,\nu) := \left(\inf_{X\sim\mu, Y\sim\nu} \mathbb{E}\,|X-Y|^p \right)^{1/p}.$$

From the definition it is clear that W_p deals with measures of finite p-th norm. It is a hypothesis of Proposition 3.1 that if $\mu \in C_\alpha^\infty(\mathcal{X}, \|\cdot\|_p)$ for $\alpha(t) = ae^{-bt^2}$ (for some $a, b > 0$), then

$$\mathbb{P}\left(W_p(L_n^X,\mu) \ge \mathbb{E} W_p(L_n^X \mu) + t\right) \le a\exp\left(-nbt^2\right),$$

for all $t \ge 0$. For some bounds on $\mathbb{E}W_p$ see e.g. [3] or [9].

Remark 1.11. *There's a following duality*

$$d_{BL}(\nu,\mu) = W_1(\nu,\mu)$$

which is a statement of Theorem 2.7.

Remark 1.12. *There are many more fields in which concentration theory proves useful.*

1. *Some type of results are true for $L_n = \frac{1}{n}\sum \delta_{d_i}$ where $d_1 \le \ldots \le d_n$ is a spectrum of a Wigner Matrix (see* [15]*).*
2. *There are also some new results on Cuolomb gases (see* [5]*).*

For more examples see [11]*, Section 1.3.*

2. Functional inequalities and optimal transport

2.1. **Functional inequalities and concentration of measure.** As we will see, the concentration of measure phenomenon turns out to be equivalent to a wide class of functional inequalities, some of which are presented here. Let us start with a definition.

[1] In the literature, the distance W_p is also called the Monge-Kantorovich, or Kantorovich-Rubinshtein, or Wasserstein transport distance, as well as the Frechet distance (in case $p = 2$), or a minimal distance. We will stick to the name Kantorovich metric due to the reasoning presented by Bobkov and Ledoux (see [3] footnote on p.4) that traces back to the article [33] by Vershick.

Definition (functional inequalities)**.** *We say that a measure μ on a metric space $(\mathcal{X}, d)$ satisfies*

$$\begin{aligned} &\textit{Logarithmic Sobolev inequality, } LSI(C), \textit{ if} && \operatorname{Ent}_\mu(f^2) \le C \|\nabla f\|_{L_2(\mu)} \ \ \forall_{f \in C^1(\mathcal{X})}, \\ &\textit{Poincaré inequality, } PI(C), \textit{ if} && \operatorname{Var}_\mu(f) \le C \|f\|_{L_2(\mu)} \ \ \forall_{f \in C^1(\mathcal{X})}, \\ &\textit{Talagrand's inequality, } T_2(C), \textit{ if} && W_2^2(\mu, \nu) \le C H(\nu|\mu) \ \ \forall_{\nu \in \mathcal{P}(\mathcal{X})}, \end{aligned}$$

where

$$\operatorname{Ent}_\nu(f) = \int f \log f \, d\nu - \left(\int f \, d\nu\right) \log \left(\int f \, d\nu\right)$$

and

$$H(\nu|\mu) = \begin{cases} \operatorname{Ent}_\mu\left(\frac{d\nu}{d\mu}\right) = \int_{\mathcal{X}} \frac{d\nu}{d\mu} \log\left(\frac{d\nu}{d\mu}\right) d\mu = \int_{\mathcal{X}} \log\left(\frac{d\nu}{d\mu}\right) d\nu & \textit{if } \nu << \mu, \\ \infty \ \ \textit{else} \end{cases}$$

is the usual Kullback-Leibler divergence also known as the relative entropy.

Remark 2.1. *Talagrand's inequality $T_2(C)$ is a special case of more general L^p-transportation cost inequality T_p. We say that μ satisfies $T_p(C)$ for $p \ge 1$ if*

$$W_p(\mu, \nu) \le \sqrt{C H(\nu|\mu)}$$

for all $\nu << \mu$.

It turns out that these inequalities can be linearly ordered in terms of their strength which is a famous result due to Otto and Villani (2000).

Theorem 2.2 (Otto – Villani, [24])**.** *Let μ be an absolutely continuous probability measure on a complete connected Riemannian manifold.*

1. *If μ satisfies $LSI(C)$, then μ satisfies $T_2(C)$.*
2. *If μ satisfies $T_2(C)$, then μ satisfies $PI(C/2)$.*

Moreover, if one restricts himself to the Euclidean situation, all three functional inequalities can be effectively characterized in terms of dimension free Gaussian concentration and one can provide the proof of Theorem 2.2 relying only on this characterization. A sketch of that argumentation is presented below and in Section 3 (c.f. Remark 3.4).

Let us start with the basic but very important tensorization property of LSI and T_2.

Proposition 2.3. *Let μ_1, μ_2 be probability measures on $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$ and set $\mu = \mu_1 \otimes \mu_2$. If μ_1, μ_2 satisfy $LSI(C)$, $PI(C)$ or $T_2(C)$ then μ does so as well.*

The proof in the case of LSI and $PI(C)$ is a simple calculation, left as an exercise (c.f. [19]).

Proof in case of T_2. Let $\nu \in \mathcal{P}(\mathbb{R}^{d_1+d_2})$ such that $d\nu(x, y) = d\nu_1(x) d\nu_2(y|x)$ and $\nu << \mu$. Observe that the Kullback-Leibler divergence can be decomposed as

$$\begin{aligned} H(\nu|\mu) &= \int_{\mathbb{R}^{d_1}} \int_{\mathbb{R}^{d_2}} \log\left(\frac{d\nu_1(x) d\nu_2(y|x)}{d\mu_1(x) d\mu_2(y)}\right) d\nu_1(x) d\nu_2(y|x) \\ &= \int_{\mathbb{R}^{d_1}} \int_{\mathbb{R}^{d_2}} \log\left(\frac{d\nu_1(x)}{d\mu_1(x)}\right) + \log\left(\frac{d\nu_2(y|x)}{d\mu_2(y)}\right) d\nu_1(x) d\nu_2(y|x) \\ &= H(\nu_1|\mu_1) + \int_{\mathbb{R}^{d_1}} H(\nu_2(\cdot|x)|\mu_2) \, d\nu_1(x). \end{aligned}$$

It therefore suffices to prove the following estimate on the Kantorovich metric

$$W_2^2(\nu,\mu) \le W_2^2(\nu_1,\mu_1) + \int_{\mathbb{R}^{d_1}} W_2^2(\nu_2(\cdot|x),\mu_2)\, d\nu_1(x).$$

Since the right hand side is expressed only in terms of marginal distributions ν and μ, it is enough to find one coupling that would satisfy the above inequality. It can be easily achieved by plugging the so-called Knoth-Rosenblatt coupling (c.f. [34], p. 26), that is

$$d\pi(x_1x_2y_1y_2) = d\pi_1(x_1y_1)\, d\pi_2^{y_1}(x_2y_2),$$

where π_1 is the optimal [2] coupling for the pair μ_1 and ν_1, which means that

$$W_2^2(\nu_1,\mu_1) = \int_{\mathbb{R}^{d1}} \int_{\mathbb{R}^{d_1}} |x-y|^2\, d\pi_1(x,y)$$

and $\pi_2^{y_1}$ is the optimal coupling between $\nu_2(\cdot|y_1)$ and μ_2. □

We will now show that the functional inequalities always imply some form of dimension-free concentration which is the statement of the following theorem.

Theorem 2.4. *Let $\mu \in \mathcal{P}(\mathbb{R}^d)$.*

1. *If μ satisfies $LSI(C)$ or $T_2(C)$, then $\mu \in C^\infty_{\alpha_2}(\mathbb{R}^d, |\cdot|)$, where $\alpha_2(t) = \exp((t-t_0)_+^2/C)$ and $t_0 = \sqrt{C\log 2}$.*
2. *If μ satisfies $PI(C)$, then $\mu \in C^\infty_{\alpha_1}(\mathbb{R}^d, \|\cdot\|)$, where $\alpha_1(t) = a\exp(-t/\sqrt{C})$ and a is some universal constant.*

Proof of the part 1. We will split the proof into two parts. The proof of the first part is based upon the famous Herbst's argument.

Assume that μ satisfies $LSI(C)$, that is

$$\int f^2 \log(f^2)\, d\mu - \left(\int f^2\, d\mu\right) \log\left(\int f^2\, d\mu\right) \le C\, \|\nabla f\|_{L_2(\mu)}$$

Then, just as in the proof of Proposition 1.6, we will try to arrive to some estimate of

$$Z(t) := \int e^{tg}\, d\mu$$

for a 1-Lipschitz function g and then apply Chebyshev's exponential inequality. Setting in the above f to be $\exp(tg/2)$ and using $|\nabla g| \le 1$ results in

$$Z'(t)t - Z(t)\log(Z(t)) \le C\frac{t^2}{4}Z(t). \tag{4}$$

Noting that $Z(t)/t \to \mu(g)$ as $t \to 0$ and rewriting (4) as

$$\frac{d}{dt}\left(\frac{\log Z(t)}{t}\right) \le \frac{C}{4},$$

whence

$$\frac{\log Z(t)}{t} \le \mu(g) + \frac{Ct}{4}.$$

Applying now Chebyshev's exponential inequality, optimizing it (c.f. the proof of Proposition 1.6), substituting $\mu(g)$ with $m_\mu(g)$ (c.f. Remark 1.4) and finally using the tensorization property 2.3 yields the result.

Assume now that μ satisfies $T_2(C)$.

[2] Such coupling always exists, c.f. Remark 2.6

Let $A \subseteq \mathbb{R}^d$, s.t. $\mu(A) \in [1/2, 1[$. Set $B = A_t^c$ and $\mu_A = \mathbb{1}_{\{A\}}/\mu(A)$, $\mu_B = \mathbb{1}_{\{B\}}/\mu(B)$. Using the fact that W_2 is a norm and exploiting T_2 property we obtain

$$\begin{aligned}W_2(\mu_A, \mu_B) &\leq W_2(\mu_A, \mu) + W_2(\mu_B, \mu)\\ &\leq \sqrt{CH(\mu_A|\mu)} + \sqrt{CH(\mu_B|\mu)}\\ &\leq \sqrt{C \log\left(1/\mu(A)\right)} + \sqrt{C \log\left(1/\mu(B)\right)}\\ &\leq t_0 + \sqrt{C \log\left(1/\mu(B)\right)}.\end{aligned}$$

On the other hand $d(x, y) \geq t$ if $x \in A$ and $y \in B$, thus $W_2(\mu_A, \mu_B) \geq t$. We arrive therefore at

$$\log(\mu(B)) \leq -(t - t_0)^2/C.$$

Again, tensorization property 2.3 yields the result. □

For an elementary proof of the second part of Theorem 2.4 we refer to [19] Corrolary 3.2.

As mentioned before, it turns out that the opposite implications also hold which, among other results, will be presented in Section 3.

2.2. **Optimal transport.** We now focus our attention on the optimal transport theory and then develop its connections to the concentration of measure theory. We will only sketch here the main ideas and theorems related to that field. For more comprehensive study, we refer to the great monograph [34].

The main objective of optimal transport is to minimize the cost of transporting one measure μ onto another measure ν subject to some cost function c.

Definition. *Let $(\mathcal{X}, d)$ be a metric space and $\mu, \nu \in \mathcal{P}(\mathcal{X})$. Given a continuous cost function $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$, we define the optimal transport cost between μ and ν to be*

$$\mathcal{T}(\mu, \nu) = \inf_{\pi \in C(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} c(x, y)\, d\pi(x, y),$$

where $C = \{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{X}) : \pi_1 = \mu, \pi_2 = \nu\}$ is a set of all coupling between μ and ν.

Remark 2.5. *Setting $c = d_p$ yields $\mathcal{T} = W_p^p$.*

Remark 2.6. *There always exists a coupling π, called optimal, which minimizes $\mathcal{T}$. For an elementary proof see* [34] *Theorem 4.1.*

While, informally, the definition of the optimal transport was focused on the cost of transporting one unit of goods from x to y, described in the parlance of cost functions, the next theorem shows different approach - through prices. It says that, given the market is in equilibrium (whatever that means), transporting one unit of goods from x to y by a cost function c should be economically equal to selling one unit of goods from x for a price $\varphi(x)$ and then buying the same unit at y for a price $\psi(y)$.

Theorem 2.7 (Kantorovich duality)**.** *Let $(\mathcal{X}, d)$ be a metric space and $\mu, \nu \in \mathcal{P}(\mathcal{X})$. We have the following duality property*

$$\mathcal{T}(\mu, \nu) = \sup_{(\varphi,\psi)\in\phi_c} \left\{ \int_{\mathcal{X}} \varphi\, d\mu - \int_{\mathcal{X}} \psi\, d\nu \right\},$$

where

$$\begin{aligned}\phi_c = \{(\varphi, \psi) :\ & \varphi \in L^1(\mu), \psi \in L^1(\nu),\ \varphi(x) - \psi(y) \leq c(x, y)\\ & \textit{for } \mu\textit{-almost every } x \textit{ and } \nu\textit{-almost every } y\}.\end{aligned}$$

Remark 2.8. *Since L^1 functions can be approximated by continuous bounded functions in the L^1 norm, the set ϕ_c can be replaced by*

$$\tilde{\phi_c} = \{(\varphi, \psi) : \varphi(x) - \psi(y) \leq c(x,y) \quad \text{for } \varphi, \psi \in C_b(\mathcal{X})\},$$

where $C_b(\mathcal{X})$ denotes the set of all continuous bounded functions on $\mathcal{X}$.

For the proof and a more detailed interpretation of the above result see [34], Chapter 5. We will admit also the following theorem.

Theorem 2.9 (Brenier, [1])**.** *Given $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ with finite second moments, such that μ is absolutely continuous w.r.t. the Lebesgue measure, then there exists a unique optimal coupling π^* for the cost function*

$$c(x,y) = \frac{|x-y|^2}{2},$$

which is deterministic in the sense that $\pi^* = \mathcal{L}(X, T(X))$ for some function T, where $X \sim \mu$. Moreover, there exists a convex function $\phi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ such that $\mu(\phi < \infty) = 1$ and $T = \nabla\phi$.

Remark 2.10. *Since ϕ is convex, then it is differentiable almost everywhere (c.f. [28], Theorem 25.5) and thus it makes sense of writing $\nabla\phi$.*

2.3. **Displacement convexity of entropy and consequences.** We shall now discuss a time-dependent version of optimal transportation and its consequence.

Theorem 2.11 (McCann [23])**.** *Let $\mu \in \mathcal{P}(\mathbb{R}^d)$ be of the form $d\mu(x) = e^{-V(x)}dx$ for some $V : \mathbb{R}^d \to \mathbb{R}$ twice differentiable s.t. $\nabla^2 V \geq cId$ for some $c \in \mathbb{R}$. Take $\nu_0, \nu_1 \in \mathcal{P}(\mathbb{R}^d)$ to be absolutely continuous w.r.t. the Lebesgue measure. Then*

$$H(\nu_t|\mu) \leq (1-t)H(\nu_0|\mu) + tH(\nu_1|\mu) - \frac{ct(1-t)}{2} W_2^2(\nu_0, \nu_1)$$

with $\nu_t = \mathcal{L}((1-t)X + t\nabla\phi(X))$, $t \in [0,1]$, where $X \sim \nu_0$ and $\nabla\phi$ is the optimal map from Theorem 2.9 transporting ν_0 onto ν_1.

Remark 2.12. *The above theorem holds true in the case of general Rimennian manifold under the uniform condition $\nabla^2 V + \mathrm{Ric} \geq cId$. The extension is due to Candero-McCann-Schmuckenschlâger [6] and was further developed by Stumm-Van Renese in [27] and Lott-Villani in [20].*

To prove Theorem 2.11 we will need and admit the following lemma.

Lemma 2.13. *Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be a convex function.*

1. *For a.e. $x \in \mathbb{R}^d$ there exist a unique symmetric matrix $H(x)$ s.t.*

$$\phi(x+u) = \phi(x) + u\nabla\phi(x) + \frac{1}{2}u(H(x)u) + o(|u|^2).$$

   We will further write $H(x) = \nabla^2\phi(x)$.

2. *If $\nu = \nabla\phi_{\#}\mu := \mathcal{L}(\nabla\phi(X))$ with $X \sim \mu$ and $d\mu(x) = f(x)dx$, $d\nu(y) = g(y)dy$, then*

$$f(x) = g(\nabla\phi(x)) \det(\nabla^2\phi(x))$$

   x-a.e. The above is the so-called Monge-Ampére equation.

3. *If ν_0, ν_1 are absolutely continuous w.r.t. the Lebesgue measure, then ν_t is so for every $t \in [0,1]$.*

The first part of Lemma is a slightly stronger version of Alexandrov's theorem. The proof can be found in [29], Collorary 2.9. For a proof of the second part see e.g [8], Theorem 3.6. The third part follows simply from the definition of ν_t.

We can now turn to the proof of the theorem.

Proof of Theorem 2.11. By the part three of Lemma 2.13 ν_t has a p.d.f.. Let us call it f_t. We can therefore write

$$H(\nu_t|\mu) = \mathrm{Ent}_\mu\left(\frac{d\nu_t}{d\mu}\right) = \mathrm{Ent}_\mu(f_t e^V) = \underbrace{\int \log(f_t)\, d\nu_t}_{I(t)} + \underbrace{\int V\, d\nu_t}_{J(t)}.$$

In the I term we recognize Shannon entropy of ν_t. The general idea for the proof is to show and exploit convexity of I and J. To that end, set $T_t(x) = (1-t)x + t\nabla\phi(x)$.

Convexity of I:

The idea is to change measure and write I_t in terms of integral w.r.t. ν_0. Then show pointwise estimate on the integrated functions. By Monge-Ampére equation

$$f_0(x) = f_t(T_t(x)) \det((1-t)I + t\nabla^2\phi(x)),$$

whence

$$I(t) = \int \log(f_t(T_t))\, d\nu_0.$$

As mentioned, we will try to prove a pointwise estimate on

$$\log(f_t(T_t)) = \log f_0 - \log\det((1-t)I + t\nabla^2\phi). \tag{5}$$

Since $\nabla^2\phi$ is symmetric(Lemma 2.13), it has nonnegative eigenvalues $(d_i)_{1\le i\le n}$. By the spectral decomposition theorem, it follows that

$$\begin{aligned}\log\det((1-t)I + t\nabla^2\phi) &= \log\prod_{i=1}^{d}((1-t)+td_i)\\ &= \sum \log((1-t)+td_i) \ge \sum t\log d_i = \log\det\nabla^2\phi\end{aligned}$$

using $\log(1+t\alpha) \ge t\log(\alpha+1)$. It follows then from (5) that

$$\begin{aligned}\log(f_t(T_t)) &\le \log f_0 - t\log\det\nabla^2\phi\\ &= (1-t)\log f_0 + t\log f_1(T_1).\end{aligned}$$

By integrating w.r.t. ν_0 we arrive at

$$I(t) \le (1-t)I(0) + tI(1). \tag{6}$$

Convexity of J:

In this part we will exploit the assumption that $\nabla^2\phi \ge cId$. Again, by the change of measure principle

$$J(t) = \int V(T_t)\, d\nu_0.$$

By the assumption, the function $x \mapsto V(x) - \frac{c}{2}|x|^2$ is convex, thus

$$\begin{aligned}V((1-t)x+ty) &\le (1-t)V(x) + tV(y) - \frac{c}{2}\left[(1-t)x^2 + ty^2 - ((1-t)x+ty)^2\right]\\ &= (1-t)V(x) + tV(y) - \frac{ct(1-t)}{2}|x-y|^2.\end{aligned}$$

Setting $y = \nabla\phi(x)$ and integrating w.r.t. ν_0 yields

$$J(t) \leq tJ(0) + (1-t)J(1) - \frac{ct(1-t)}{2}W_2^2(\nu_0, \nu_1) \tag{7}$$

Combining together (6) and (7) yields the result. □

We will now present some consequences of Theorem 2.11.

Proposition 2.14 (Talagrand's identity). *If ν_0 and μ satisfy conditions of Theorem 2.11 and $c > 0$, then*

$$W_2^2(\nu_0, \mu) \leq \frac{2}{c}H(\nu_0|\mu).$$

Proof. Apply Theorem 2.11 for $\nu_1 = \mu$, $t = 1$. □

Proposition 2.15 (HWI inequality). *Let $c \in \mathbb{R}$. For ν_0, μ satisfying assumptions of Theorem 2.11 we have*

$$I(\nu_0|\mu) \leq W_2(\nu_0|\mu)\sqrt{I(\nu_0|\mu)} - \frac{c}{2}W_2^2(\nu_0, \mu),$$

where

$$I(\nu_0|\mu) = \int \frac{|\nabla h_0|^2}{h_0}d\mu, \quad \textit{with } h_0 = \tfrac{d\nu_0}{d\mu}.$$

In the above, HWI stands for the usual symbols for entropy H, Kantorovich distance W and Fisher information I.

Sketch of a proof. Setting $\nu_1 = \mu$ in McCann's theorem we arrive at

$$H(\nu_t|\mu) \leq (1-1)H(\nu_0|\mu) - \frac{ct(1-t)}{2}W_2^2(\nu_0, \mu),$$

which is equivalent to

$$\Delta(t) := \frac{H(\nu_t|\mu) - H(\nu_0|\mu)}{t} \leq -H(\nu_0|\mu) - \frac{c(1-t)}{2}W_2^2(\nu_0, \mu). \tag{8}$$

It can be easily shown that (c.f. [34] Theorem 20.1)

$$\liminf_{t\to 0} \Delta(t) \geq \int \frac{\nabla h_0(x)}{h_0(x)}(x - \nabla\phi(x))\, d\nu_0(x).$$

Now, using Cauchy-Schwarz inequality

$$\begin{aligned}\liminf_{t\to 0} \Delta(t) &\geq -\sqrt{\int \frac{|\nabla h_0|^2}{|h_0|^2}\, d\nu_0 \int |x - \nabla\phi(x)|^2\, d\nu_0(x)} \\ &= -\sqrt{\int \frac{|\nabla h_0|^2}{h_0}\, d\mu}\sqrt{\int |x - \nabla\phi(x)|^2\, d\nu_0(x)} \\ &= -\sqrt{I(\nu_0|\mu)}W_2(\nu_0, \mu).\end{aligned}$$

Plugging the above into (8) yields the result. □

Remark 2.16. *As an extra result, from the proof of Theorem 2.11 one can deduce Brunn-Minkowski inequality. Recall that*

$$I(t) \leq tI(0) + (1-t)I(1),$$

where $I(t) = \int \log(f_t)\, d\nu_t$, $\nu_t = [(1-t)I + t\nabla\phi]_{\#}\nu_0$. Setting

$$d\nu_0 = \frac{\mathbb{1}_{\{A\}}}{|A|}dx, \quad d\nu_1 = \frac{\mathbb{1}_{\{B\}}}{|B|}dx$$

for A, B - compact sets, one gets

$$I(t) \le t\log\left(\frac{1}{|A|}\right) + (1-t)\log\left(\frac{1}{|B|}\right).$$

Using the well known result in information theory that Schannon entropy is maximized for uniform measure one gets

$$I(t) \ge \log\left(\frac{1}{|\operatorname{supp}\nu_t|}\right) \ge \log\left(\frac{1}{|tA + (1-t)B|}\right)$$

since $\operatorname{supp}\nu_t \subseteq tA + (1-t)B$. Combining the above observations we arrive at classical Brunn-Minkowski inequality

$$|tA + (1-t)B| \ge |A|^t\, |B|^{1-t}$$

for all $t \in [0,1]$ and all compact sets A, B.

3. Characterization of some concentration of measure phenomena

In this section we will present some results that are complementary to the ones achieved in Section 2. Namely, we will try to prove that some functional inequalities imply concentration of measure inequalities.

Theorem 3.1 (Gozlan, [10])**.** *If μ satisfies the dimension-free Gaussian concentration property with a profile*

$$\alpha_2(t) = \exp\left(-C(t-t_0)_+^2\right),$$

then μ satisfies $T_2(\frac{1}{C})$.

Theorem 3.2 (Gozlan, Roberto, Samson [12])**.** *Let $\mu \in \mathcal{P}(\mathbb{R}^d)$. If μ satisfies the dimension-free concentration property with a profile α such that $\alpha(t) < 1/2$ for some t, then μ satisfies $PI(C)$, where*

$$C = \left(\inf\left\{\frac{r}{\bar{\phi}^{-1}(\alpha(r))} : \quad r > 0, \alpha(r) < 1/2\right\}\right)^2.$$

Remark 3.3. *In view of the above, we get that the concentration profile $\bar{\phi}$ in Theorem 1.8 is optimal. Indeed, plugging $\alpha = \bar{\phi}$ into Theorem 3.2 implies that γ satisfies $PI(1)$, which cannot be improved (i.e. 1 is a minimal constant in PI for any metric space).*

Remark 3.4. *Combining together theorems 2.4, 3.1 and 3.2 gives the proof of Theorem 2.2 in the restricted case of $\mathbb{R}^d$.*

Remark 3.5. *The dimension-free concentration is always et least exponential. Indeed, assume that μ satisfies dimension-free concentration property with some polynomial profile α. Then by Theorem 3.2 μ satisfies $PI(C)$ for some C. Now, Theorem 2.4 says that α is in fact exponential.*

Remark 3.6. *If μ satisfies the dimension-free concentration property, then μ satisfies $PI(C)$, which implies that its support is connected (which is a simple exercise). As a consequence, we get that discrete measures cannot satisfy dimension-free concentration property.*

We will now turn to the proof of Theorem 3.1. First, we have to start with recalling a well known result from large deviation theory.

Theorem 3.7 (Sanov, [30])**.** *Let $X, X_1, \ldots, X_n$ be a sequence of i.i.d. random variables taking values in a Polish space $\mathcal{X}$. Set $\mu = \mathcal{L}(X)$ and $L_n^X = \frac{1}{n}\sum \delta_{X_i}$. By the strong law of large numbers L_n^X converges weakly almost surely to μ. Let $\mathcal{O}, \mathcal{F} \in \mathcal{P}(\mathcal{X})$ be an open and a closed set (in weak topology) respectively.*

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(L_n^X \in \mathcal{O}\right) \geq -H(\mathcal{O}|\mu)$$
$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(L_n^X \in \mathcal{F}\right) \leq -H(\mathcal{F}|\mu),$$

where

$$H(A|\mu) = \inf_{\nu\in A} H(\nu|\mu).$$

We say that $(L_n^X)_{n\in\mathbb{N}}$ satisfies the large deviation property with a decay function $H(\cdot|\mu)$.

Remark 3.8. *Rougly speaking, the above theorem states that*

$$\mathbb{P}\left(L_n^X \in A\right) \simeq e^{-nH(A|\mu)}.$$

Moreover,

$$\mu \in \operatorname{Int} A \Rightarrow \mathbb{P}\left(L_n^X \in A\right) \to 1,$$
$$\mu \notin \operatorname{cl} A \Rightarrow \mathbb{P}\left(L_n^X \in A\right) \to 0 \textit{ exponentially fast,}$$

where $\operatorname{Int} A$ *and* $\operatorname{cl} A$ *denote interior and closure of* A *respectively.*

Equipped with the above theorem, we can begin the proof of Theorem 3.1.

Proof of Theorem 3.1. The idea of the proof is to plug Kantorovich metric into the definition of the dimension-free concentration property and then express the appropriate term in terms of Sanov's theorem.

Let us therefore set

$$(\mathbb{R}^d)^n \ni X = (x_1, \ldots, x_n) \xrightarrow{F_n} W_2(L_n^X, \mu).$$

It straightforward to check that F_n is $\frac{1}{\sqrt{n}}$-Lipschitz and thus by the dimension-free concentration property

$$\mathbb{P}\left(\sqrt{n}F_n(X) > \sqrt{n}\mathbb{E}F_n(X) + t\right) \leq \exp(-C(t-t_0)_+^2),$$

where $X = (X_1, \ldots, X_n)$ with $X_1, \ldots, X_n$ - an i.i.d. sample from the distribution μ. Setting $t = u\sqrt{n}$ and $\epsilon_n = \mathbb{E}F_n(X)$ we arrive at

$$\mathbb{P}\left(W_2(L_n^X, \mu) > \epsilon_n + u\right) \leq \exp\left(-nC\left[u - t_0/\sqrt{n}\right]_+^2\right).$$

It is not that hard to show (c.f. [3], Theorem 2.14) that $\epsilon_n \to 0$, which implies that

$$\limsup_{n\to\infty}\{W_2(L_n^X, \mu) > \epsilon_n + u\} = \limsup_{n\to\infty}\{W_2(L_n^X, \mu) > u\}$$

and, using the inequality $\limsup \mathbb{P}(A_n) \leq \mathbb{P}(\limsup A_n)$, we arrive at

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(W_2(L_n^X, \mu) > u\right) \leq -Cu^2. \tag{9}$$

On the other side, setting

$$\mathcal{O} = \{\nu \in \mathcal{P}(\mathbb{R}^d) : W_2(\nu, \mu) > u\}$$

and applying Sanov's theorem results in

$$-H(\mathcal{O}|\mu) \leq \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(L_n^X \in \mathcal{O}\right). \tag{10}$$

Equations (9) and (10) imply

$$Cu^2 \leq \inf\{H(\nu|\mu) :\, W_2(\nu,\mu) > u\}.$$

Take now ν_0 such that $H(\nu_0|\mu) < \infty$ and $u = W_2(\nu_0,\mu) - \epsilon$ for some $\epsilon > 0$. Then

$$C(W_2(\nu_0,\mu) - \epsilon)^2 \leq H(\nu_0|\mu).$$

Tending with ϵ to 0 yields the result. □

Remark 3.9. *It still remains an open question what functional inequality is equivalent to the dimension-free concentration with a concentration profile of the form*

$$\alpha_p(t) = a\exp\left(-b[t-t_0]_+^p\right)$$

with $p \in]1,2[$.

4. Concentration for convex functions

4.1. **Marton's transport inequality.** Let $\nu_0, \nu_1 \in \mathcal{P}(\mathcal{X})$ for some metric space $(\mathcal{X}, d, \mu)$, $\nu_0, \nu_1 << \mu$ and set $\nu_t = (1-t)\nu_0 + t\nu_1$ and $H(t) = H(\nu_t|\mu)$. It is well known that $H(t)$ is a convex function

$$H(t) \leq (1-t)H(0) + tH(1).$$

The interesting question is if it is better than convex, that is – how big curvature term F can we subtract from the right hand of the above inequality

$$H(t) \leq (1-t)H(0) + tH(1) - \frac{t(1-t)}{2}F(\nu_0,\nu_1)$$

for it to still hold true. Denoting $h_t = \frac{d\nu_t}{d\mu}$ we get

$$\begin{aligned} H(t) &= \int [\log((1-t)h_0 + th_1)]((1-t)h_0 + th_1)\, d\mu, \\ H'(t) &= \underbrace{\int (h_1 - h_0)\, d\mu}_{=0} + \int [\log((1-t)h_0 + th_1)](h_1 - h_0)\, d\mu, \\ H''(t) &= \int \frac{(h_1-h_0)^2}{(1-t)h_0 + th_1}\, d\mu. \end{aligned}$$

We would like to find some upper bound on $H''(t)$.

4.2. **First attempt.** Cauchy-Schwarz inequality gives

$$\begin{aligned} \int \frac{(h_1-h_0)^2}{(1-t)h_0 + th_1}\, d\mu &= \int \frac{(h_1-h_0)^2}{(1-t)h_0 + th_1}\, d\mu \int ((1-t)h_0 + th_1)\, d\mu \\ &\geq \left(\int |h_1 - h_0|\, d\mu\right)^2 = 4\left\|\nu_0, \nu_1\right\|_{TV}^2, \end{aligned}$$

where $\|\cdot,\cdot\|_{TV}$ denotes the total variation distance

$$\|\mu - \nu\|_{TV} = \sup\{\nu(A) - \mu(A) :\, A \in \Sigma\}.$$

Subtracting the curvature term implies thus

$$0 \leq H(\nu_t|\mu) \leq (1-t)H(\nu_0|\mu) + tH(\nu_1|\mu) - 2(t-1)t\,\|\nu_0 - \nu_1\|_{TV}^2 .$$

By setting $\nu_1 = \mu$, we arrive at Pinsker's inequality ([25]) with optimal constant 2 provided independently by Csiszár ([7]), Kempermann ([17]) and Kullback ([18]).

Corollary 4.1 (Pinsker's inequality). *For any probability measures $\mu, \nu \in \mathcal{P}(\mathcal{X})$*

$$\|\mu - \nu\|_{TV} \leq \sqrt{\frac{1}{2}H(\nu|\mu)}.$$

Remark 4.2. *Since the relative entropy $H(\nu|\mu)$ becomes infinity unless $\nu << \mu$, we could have removed in Corollary 4.1 the assumption that $\nu << \mu$.*

Remark 4.3. *Recall that*

$$\|\mu - \nu\|_{TV} = \inf_{X\sim\mu, Y\sim\nu} \mathbb{P}\left(X \neq Y\right)$$

for any $\mu, \nu \in \mathcal{P}(\mathcal{X})$.

Proof. It suffices to apply Kantorovich duality (Theorem 2.7) for the cost function

$$c(x,y) = \mathbb{1}_{\{x\neq y\}}.$$

□

Remark 4.4. *In the view of the above, Pinsker inequality can be seen a special case of the L^1-transportation inequality (c.f. Remark 2.1) which is therefore true for any probability measure.*

Exercise 4.1. *Prove the following tensorization property of Pinsker's inequality*

$$d_1(\nu, \mu^{\otimes n}) \leq \sqrt{\frac{n}{2}H(\nu|\mu^{\otimes n})},$$

where

$$d_1(X,Y) = \sum_{i=1}^{n} \mathbb{1}_{\{X_i \neq Y_i\}}$$

for $X_i, Y_i \in \mathcal{X}^n$ (c.f. [21]*, Proposition 1).*

Exercise 4.2. *Apply the above to deduce Hoeffding concentration inequality in the case of d_1 metric.*

It turns out that we receive bad dimension scaling behavior and in that sense we cannot retrieve dimension-free concentration from the above results. We need to try something else.

4.3. **Second attempt: Marton's argument.** Estimating weighted average from the denominator in the expression for $H''(t)$ by the maximum results in

$$H''(t) = \int \frac{(h_1-h_0)^2}{(1-t)h_0 + th_1}\,d\mu \geq \int \frac{(h_1-h_0)^2}{\max(h_0,h_1)}\,d\mu = \tilde{\mathcal{T}}_2(\nu_0|\nu_1) + \tilde{\mathcal{T}}_2(\nu_1|\nu_0),$$

where

$$\tilde{\mathcal{T}}_2(\nu_0|\nu_1) := \int \left(1 - \frac{d\nu_0}{d\nu_1}\right)_+^2 d\nu_1 = \int \left(1 - \frac{h_0}{h_1}\right)_+^2 h_1\,d\mu$$

is so-called Marton's transport cost which is a special case of a weak transport cost.

Proposition 4.5. *For any pair of measures $\nu_0, \nu_1 \in \mathcal{P}(\mathcal{X})$ one has*

$$\tilde{\mathcal{T}}_2(\nu_0|\nu_1) = \inf_{X\sim\nu_0, Y\sim\nu_1} \left\{ \int \left[\mathbb{P}\left(X \neq y | Y = y\right)\right]^2 \, d\nu_1(y) \right\}$$

The above observation is analogous to the one from Remark 4.3. It can be written also in the form

$$\tilde{\mathcal{T}}_2(\nu_0|\nu_1) = \inf_{X\sim\nu_0, Y\sim\nu_1} \mathbb{E}\left[\mathbb{E}[\mathbb{1}_{\{X\neq Y\}}|Y]^2\right]$$

Proposition 4.6 (Marton, Lemma 3.2 in [22])**.** *For any $\mu, \nu \in \mathcal{P}(\mathcal{X})$*

$$\max\left(\tilde{\mathcal{T}}_2(\nu|\mu), \tilde{\mathcal{T}}_2(\mu|\nu)\right) \leq 2H(\nu|\mu).$$

In that case, we observe good tensorization property, not depending on the dimension.

Proposition 4.7. *For any $\mu \in \mathcal{P}(\mathcal{X})$, $\nu \in \mathcal{P}(\mathcal{X}^n)$*

$$\max\left(\tilde{\mathcal{T}}_2^{(n)}(\nu|\mu^{\otimes n}), \tilde{\mathcal{T}}_2^{(n)}(\mu^{\otimes n}|\nu)\right) \leq 2H(\nu|\mu^{\otimes n}),$$

where

$$\tilde{\mathcal{T}}_2^{(n)}(\nu_0, \nu_1) = \inf_{X\sim\nu_0, Y\sim\nu_1} \left\{ \int \sum_{i=1}^{n} \left[|\mathbb{P}\left(X_i \neq y_i | Y_i = y_i\right)|^2 \, d\nu_1(y)\right] \right\}.$$

For a complete proof we refer to [22]. It is basically the same as in the standard case $n = 1$.

Now, we shall turn to the very interesting consequence of the above result which is a dimension-free concentration for convex functions. Let us start with a definition.

Definition (Convex concentration property)**.** *We say that a measure μ on some metric space $(\mathcal{X}, d)$ satisfies (dimension-free subgaussian) convex concentration property with a constant C ($CCP(C)$) if conditions*

$$\mu^{\otimes n}(f \geq \mu^{\otimes n}(f) + t) \leq \exp(-t^2C),$$
$$\mu^{\otimes n}(f \leq \mu^{\otimes n}(f) - t) \leq \exp(-t^2C)$$

are satisfied for all $n \in \mathbb{N}$, $t \geq 0$ and all $f \in Lip_1(\mathcal{X}, d)$.

Theorem 4.8. *Suppose $\mu \in \mathcal{P}(\mathbb{R}^d)$ with bounded support on a disk of a diameter not greater than D. Then μ satisfies $CCP(1/2D^2)$.*

Sketch of a proof. Setting X, Y to be a random variables on $(R^d)^n$, f - a convex 1-Lipschitz function on $(R^d)^n$ results in

$$f(X) - f(Y) \leq \nabla f(X)(X - Y), \tag{11}$$

where we assume $\nabla f(X)$ exists almost surely. The main idea is to exploit the above inequality, arrive to the point where Proposition 4.7 can be used and then plug appropriate random variables X and Y.

Conditioning w.r.t. X in (11) results in

$$f(X) - \mathbb{E}[f(Y)|X] \leq \nabla f(X)(X - \mathbb{E}[Y|X]).$$

Now, taking expectation an using Cauchy-Schwarz inequality,

$$\mathbb{E}[f(X) - f(Y)] \leq \left(\mathbb{E}\,|\nabla f(X)|^2 \,\mathbb{E}\,|X - \mathbb{E}[Y|X]|^2\right)^{1/2}.$$

Assuming that $X = (X_1, \dots, X_n)$ and $Y = (Y_1, \dots, Y_n)$ and X_i, Y_i are for every index i centered on a disk with a diameter not greater than D and using that f is 1-Lipschitz we arrive at

$$\begin{aligned}\mathbb{E}[f(X) - f(Y)] &\le \left(\mathbb{E}\,|X - \mathbb{E}[Y|X]|^2\right)^{1/2} \\ &= \left(\mathbb{E}\sum_{i=1}^{n} |\mathbb{E}[X-Y|X]|^2\right)^{1/2} \le \left(2D^2\mathbb{E}\sum_{i=1}^{n}\left[\mathbb{E}[\mathbb{1}_{\{X_i \neq Y_i\}}|X]\right]^2\right)^{1/2}.\end{aligned}$$

The left hand side of the above inequality depends only on the marginals of $\mathcal{L}(X,Y)$, thus we can take infimum over all couplings, whence

$$\mathbb{E}[f(X) - f(Y)] \le \sqrt{2D^2\tilde{\mathfrak{T}}_2^{(n)}(\mathcal{L}(Y)|\mathcal{L}(X))}.$$

Now it suffices to choose appropriate distributions of X and Y. Setting $Y \sim \mu^{\otimes n}$, $X \sim \nu$ for $d\nu/d\mu = \mathbb{1}_{\{A\}}/\mu(A)$ with $A = \{f \ge \mu^{\otimes n}(f) + t\}$ results in

$$\mathbb{E}f(Y) = \mu^{\otimes n}(f), \quad \mathbb{E}f(X) \ge \mu^{\otimes n}(f) + t.$$

Therefore, by Proposition 4.7

$$t^2 \le 2D^2\tilde{\mathfrak{T}}(\mu^{\otimes n}|\nu) \le 2D^2 H(\nu|\mu^{\otimes n}) = 2D^2\log\left(\frac{1}{\mu^{\otimes n}(f \ge \mu^{\otimes n} + t)}\right).$$

Similar reasoning will lead to the proof of the second condition from the definition of the convex concentration property. □

Remark 4.9. *From the above, it follows that the convex concentration property can be satisfied by discrete measures, which is not true for the general dimension free concentration (c.f. Exercise 1.1 and Remark 3.6).*

4.4. **Generalizations.** In the previous section we have considered Marton's transport cost of the form

$$\tilde{\mathfrak{T}}(\nu|\mu) = \inf\left\{\int c(x,p^x)\,d\mu(x) : \quad \int dp^x(y) = d\nu(x) \ \text{ for } \mu - \text{a.e. } x\right\} \tag{12}$$

with a cost function

$$c(x,p) = \left(\int \mathbb{1}_{\{x\neq y\}}\,dp(y)\right)^2.$$

On the other side, the usual transport cost W_2^2 corresponds to a cost function

$$c(x,p) = \int d^2(x,y)\,dp(y).$$

There are plenty of natural questions that arise here. Can one generalize some of the above results onto some other cost functions? Are there some classes of cost functions for which the general results hold true? In the rest of this section we will present some of the results from are related to the above questions. For more on the subject we refer to [13].

There are two (among others) possible, natural form of a cost function that turn out to have good properties. Namely we can consider the following

$$c(x,p) = \left(\int d(x,y)\,dp(y)\right)^2, \quad c(x,p) = \left|x - \textstyle\int y\,dp(y)\right|^2.$$

For the above, one can prove similar tensorization results as in Property 4.7 (c.f. [13]).

Definition. *Let $\mu \in \mathcal{P}(\mathbb{R}^d)$. On says that μ satisfies $\bar{\mathcal{T}}_2(C)$ if*

$$\max(\bar{\mathcal{T}}_2(\nu|\mu), \bar{\mathcal{T}}_2(\mu|\nu)) \leq CH(\nu|\mu)$$

for all $\nu \in \mathcal{P}(\mathbb{R}^d)$, where

$$\bar{\mathcal{T}}_2(\nu|\mu) = \inf\left\{ \int \left| x - \int y\, dp^x(y) \right| d\mu(x) : \quad \int dp^x(y) = d\nu(x) \ \ \text{for } \mu - a.e. \ x \right\}.$$

Theorem 4.10. *If μ satisfies $\bar{\mathcal{T}}_2(C)$, then μ satisfies $CCP(1/C)$.*

The proof is identical to the proof of Theorem 4.8. An interesting fact is that CCP is actually equivalent to $\bar{\mathcal{T}}_2$.

Theorem 4.11. *If $\mu \in \mathcal{P}(\mathbb{R}^d)$ satisfies $CCP(C)$, then μ satisfies $\bar{\mathcal{T}}_2(a/C)$ for some universal constant a.*

For the proof, we refer to [13].

References

[1] Y. Brenier (1991). *Polar factorization and monotone rearrangement of vector-valued functions.* Comm. Pure Appl. Math., vol. 44, pp. 1097–0312.

[2] V. Bogachev (2007). *Measure Theory vol. II.* Springer.

[3] S. Bobkov, M. Ledoux (2016). *One-dimensional empirical measures, order statistics, and Kantorovich transport distances.* Preprint. To appear in: Memoirs of the AMS.

[4] Ch. Borell (1975). *The Brunn-Minkowski inequality in Gauss space.* Invent. Math., vol. 30, no. 2, pp. 207 – 216.

[5] D. Chafai, A. Hardy, M. Maida (2017). *Concentration for Coulomb gases and Coulomb transport inequalities. Improvement on an assumption, and minor modifications.* Retrieved September 1, 2017 from `https://hal.archives-ouvertes.fr/hal-01374624v3`.

[6] D. Cordero-Erausquin, R. McCann, M. Schmuckenschlâger (2001). *A Riemannian interpolation inequality á la Borell, Brascamp and Lieb.* Invent. Math., vol. 146, pp. 219 – 257.

[7] I. Csiszár (1967). *Information-type measures of difference of probability distributions and indirect observations.* Studia Sci. Math. Hungarica, vol. 2, pp. 299 – 318.

[8] G. De Philippis, A. Figalli (2013). *The Monge-Ampére equation and its link to optimal transportation.* Bull. Amer. Math. Soc., vol. 51, no. 4, pp. 527–580.

[9] N. Fournier, A. Guillin (2013). *On the rate of convergence in Wasserstein distance of the empirical measure.* Springer, Vol. 162, No. 34, pp 707 – 738.

[10] N. Gozlan (2009) *A characterization of dimension free concentration in terms of transportation inequalities.* Ann. Probab., vol. 37, no. 6., pp. 2480–2498.

[11] N. Gozlan (2016) *Transport inequalities and concentration of measure.* ESAIM: Proc., vol. 51., pp. 1–23.

[12] N. Gozlan, C. Roberto, P. Samson (2013). *Characterization of Talagrands transport-entropy inequalities in metric spaces* Ann. Probab., vol. 41, no. 5, pp. 3112–3139.

[13] N. Gozlan, C. Roberto, P. Samson, P. Tetali (2014). *Kantorovich duality for general transport costs and applications.* to appear in J. Funct. Anal., preprint, arXiv:1412.7480v4.

[14] L. Gross (1975) *Logarithmic Sobolev inequalities.* Amer. J. Math., vol. 97, pp. 1061–1083.

[15] A. Guionnet, O. Zeitouni (2000) *Concentration of the Spectral Measure for Large Matrices* Electron. Commun. Probab., Vol. 5, no. 14, pp. 119–136.

[16] W. Hoeffding (1963). *Probability inequalities for sums of bounded random variables.* J. of the Amer. Stat. Assn., vol. 58, no. 301, pp. 13–30.

[17] J. Kempermann (1967). *On the optimum rate of transmitting information.* Probab. Inf. Theory, Lecture Notes in Mathematics, vol. 89, pp. 126–169.

[18] S. Kullback (1967). *A lower bound for discrimination information in terms of variation.* IEEE Trans. on Inf. Theory, vol. 13, pp. 126–127.

[19] M. Ledoux (2001). *The concentration of measure phenomenon.* Mathematical Surveys and Monographs, vol. 89, Amer. Math. Soc.

[20] J. Lott, C. Villani (2009). *Ricci curvature for metric-measure spaces via optimal transport.* Ann. Math., vol. 169, pp. 903–991.

[21] K. Marton (1996). *Bounding $\bar{d}$-distance by informational divergence: a method to prove measure concentration.* Ann. Prob., vol 24., no. 2, pp. 857–866.

[22] K. Marton (1996). *A measure concentration inequality for contracting Markov chains.* Geom. Funct. Anal., vol. 6, no. 3.

[23] R. McCann. (1997). *A convexity principle for interacting gases.* Adv. Math., vol. 128, pp. 153–179.

[24] F. Otto, C. Villani (2000). *Generalization of an Inequality by Talagrand and Links with the Logarithmic Sobolev Inequality.* J. Funct. Anal., vol. 137, pp. 361–400.

[25] M. Pinsker (1960). *Information and information stability of random variables and processes.* Izd. Akad. Nauk. SSSR.

[26] G. Pisen (1989). *The volume of convex bodies and Banach space geometry.* Tracts in Math 94, Cambridge University Press, pp. 44–47.

[27] M. von Renesse, K. Sturm (2009). *Entropic measure and Wasserstein diffusion.* Ann. Probab., vol. 37, no. 3, pp. 1114–1191.

[28] R. Rockafellar (1970). *Convex Analysis.* Princeton University Press.

[29] R. Rockagellar (1999). *Second-order convex analysis.* J. Nonlinear and Convex Anal., vol.1, pp. 1–16.

[30] I. Sanov (1957). *On the probability of large deviations of random variables.* Mat. Sbornik, vol. 42, pp. 11–44.

[31] V. Sudakov, B. Tsirelson (1974). *Extremal properties of half-spaces for spherically invariant measures.* Zap. LOMI, vol. 41 , pp. 14–24.

[32] V. Varadarajan (1958). *On the Convergence of Sample Probability Distributions.* The Indian J. of Stat. (1933-1960), vol. 19, no. 1/2, pp. 23–26.

[33] A. Vershik (2013). *Long history of the Monge-Kantorovich transportation problem.* Math. Intelligencer, vol. 35, pp. 1–9.

[34] C. Villani (2009). *Optimal Transport: Old and New.* Springer.