MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„An entropic regularity approach to the Caffarelli's contraction theorem"

verfasst von / submitted by

Petros Boufounos, Pt. Metapt.

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2019 / Vienna, 2019

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 066 821
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Mathematik
Betreut von / Supervisor:	Univ.-Prof. Dr. DI Mathias Beiglböck

AN ENTROPIC REGULARITY APPROACH TO THE CAFFARELLI CONTRACTION THEOREM

Abstract

The purpose of this thesis is to give a different proof of the Caffarelli contraction theorem, which states that the Brenier map pushing forward the Standard Gaussian measure onto a logarithmically concave probability measure is Lipschitz with constant 1. Caffarelli's original proof was mostly based on PDE theory and the fact that Brenier maps appear as solutions to a Monge-Ampère equation and was not directed towards optimal transport theory. In the current proof we mainly follow the research work of Fathi, Gozlan and Prodhomme [14] who exploited a recent characterization of Lipschitz transport map given by Gozlan and Juillet together with a convexity property of optimizers in the entropic transport cost minimization problem.

Zusammenfassung

In dieser Arbeit präsentieren wir einen alternativen Beweis des Kontraktionssatzes von Caffarelli, welcher besagt, dass die Brenier-Abbildung zwischen einer Standardnormalverteilung und einem logarithmisch konkaven Wahrscheinlichkeitsmass Lipschitz-stetig mit Konstante 1 ist. Der unsprüngliche Beweis von Caffarelli ist nicht direkt mit optimalem Transport verknüpft, sondern basiert auf der Theorie der partiellen Differentialgleichungen und der Tatsache, dass die Brenier-Abbildung einer Lösung der Monge-Ampèreschen Gleichung entspricht. Der Beweis dieser Arbeit folgt hauptsächlich der Forschungsarbeit von Fathi, Gozlan und Prodhomme [14]. Diese Autoren benutzen eine neue Charakterisierung von Lipschitz-stetigen Transport-Abbildungen nach Gozlan und Juillet, sowie eine Konvexitätseigenschaft der Optimierer des Minimierungsproblems, welches durch die entropischen Transportkosten gegeben ist.

ACKNOWLEDGEMENTS

I would like to thank, first of all, my family, relatives and friends for their unconditional emotional and mental support during all these years of my stay in Vienna. Without their encouragement it would be really difficult to finish my studies properly. Moreover, I would like to express my gratitude to my professsor Mathias Beiglböck for his willingness to supervise my master thesis and to guide me with patience until its completion. He was very cooperative and always open whenever I was in trouble or had a question. He allowed me to work independently but also showed the right direction when this was necessary. Many thanks, in addition, to Bertram Tschiderer and Alexander Mikosch for their crucial help with the translation of the "Abstract" part into German.

Contents

1. INTRODUCTION

Before we proceed, let us give a definition. We denote by γ_d the standard Gaussian measure on $\mathbb{R}^d$, that is $\gamma_d : \mathcal{B}(\mathbb{R}^d) \to [0,1]$ and

$$\gamma_d(A) = \frac{1}{(2\pi)^{\frac{d}{2}}} \int_A e^{-\frac{\|x\|^2}{2}} d\lambda^d \quad , \quad A \in \mathcal{B}(\mathbb{R}^d)$$

where λ^d is the Lebesgue measure on $\mathbb{R}^d$ and $\|x\|$ is the usual Euclidean norm on $\mathbb{R}^d$. Sometimes we will use Radon-Nikodym derivative notation, so the above takes the form:

$$\frac{d\gamma_d}{d\lambda^d} = \frac{1}{(2\pi)^{\frac{d}{2}}} e^{-\frac{\|x\|^2}{2}}$$

or the form:

$$\gamma_d(dx) = \frac{1}{(2\pi)^{\frac{d}{2}}} e^{-\frac{\|x\|^2}{2}} \lambda^d(dx) \ .$$

We can now state the generalized version of Caffarelli's theorem:

Theorem 1 (Caffarelli)**.** *Suppose μ, ν are two probability measures on $\mathbb{R}^d$ of the form $\mu(dx) = e^{V(x)}\gamma_d(dx)$ and $\nu(dx) = e^{-W(x)}\gamma_d(dx)$ for some V and W convex functions. Assume also that μ has finite second moment, i.e. $\int_{\mathbb{R}^d} \|x\|^2 d\mu(x) < \infty$ and that ν is compactly supported. Then there exists a continuously differentiable and convex function $\phi : \mathbb{R}^d \to \mathbb{R}$ such that $\nabla\phi$ is 1-Lipschitz and $\nu = \nabla\phi_{\#}\mu := \mu \circ (\nabla\phi)^{-1}$.*

The original result was only stated for the case where $V = 0$, but the proof can also be generalized (see [28]). Also the assumption that ν has compact support can be removed via approximation (See [37], Corollary 5.21). Throughout the essay we always assume that convex functions are also taking the value $+\infty$ and that they are lower semicontinuous.

Caffarelli's result enables to transfer geometric inequalities (such as the Sobolev or Isoperimetric inequalities) from the Gaussian measure onto probability measures having a logarithmically concave density. This is really important for applications in statistics for example, where many common probability densities are logarithmically concave functions (e.g. Student's distribution, Normal distribution, Exponential distribution etc). The crucial point is the dimension-free nature of the above bound which allows to preserve the dimension-independent estimates that arise from these inequalities. See [7,22,23,33] and [27,28] for more applications of Theorem 1.

As already mentioned, in this essay we will try to give a proof based on ideas from optimal transport theory. First of all, we have to recall a result obtained by Gozlan and Juillet in [20], which gives a variational characterization of Lipschitz regularity of optimal transport maps. Let us introduce some notation. Denote by $\mathcal{P}(\mathbb{R}^d)$ the set of probability measures on $\mathbb{R}^d$ and by $\mathcal{P}_k(\mathbb{R}^d)$ the subset of $\mathcal{P}(\mathbb{R}^d)$ consisting of those probability measures having finite moment of order k, $k \geq 1$, i.e.

$$\mathcal{P}_k(\mathbb{R}^d) = \left\{ P \in \mathcal{P}(\mathbb{R}^d) : \int_{\mathbb{R}^d} \|x\|^k dP(x) < \infty \right\} .$$

Next we define the quadratic Wasserstein distance for $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ as follows:

$$W_2^2(\mu, \nu) := \inf_{\pi \in C(\mu,\nu)} \left\{ \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \pi(dx, dy) \right\}$$

where $C(\mu, \nu)$ is the set of all couplings between μ and ν, i.e. the probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ satisfying $\pi(A \times \mathbb{R}^d) = \mu(A)$ and $\pi(\mathbb{R}^d \times B) = \nu(B)$ for all Borel sets $A, B \in \mathbb{R}^d$. Furthermore, for $\eta_1, \eta_2 \in \mathcal{P}(\mathbb{R}^d)$, we say η_1 is dominated by η_2 in the convex order if $\int f d\eta_1 \leq \int f d\eta_2$ for all convex functions $f : \mathbb{R}^d \to \mathbb{R}$. In that case we simply write $\eta_1 \leq_c \eta_2$. We can now state the variational characterization (see [20]):

Theorem 2. *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$. Then the following are equivalent:*

(i) There exists a continuously differentiable and convex function $\phi : \mathbb{R}^d \to \mathbb{R}$ such that $\nabla\phi$ is 1-Lipschitz and $\nu = \nabla\phi_{\#}\mu$

(ii) For all $\eta \in \mathcal{P}_2(\mathbb{R}^d)$ such that $\eta \leq_c \nu$ it holds

$$W_2(\mu, \nu) \leq W_2(\mu, \eta)$$

This means that the Brenier map between μ and ν is 1-Lipschitz iff the measure ν is the closest point to μ among all probability measures which are dominated by ν in the convex order. Thus, in order to prove Theorem 1, it suffices to show that whenever μ and ν satisfy the assumptions of Theorem 1, it holds that:

$$W_2(\mu, \nu) \leq W_2(\mu, \eta), \quad \forall \eta \leq_c \nu \ . \tag{1}$$

We will prove a similar statement replacing the Wasserstein distance by the entropic transport cost, $\mathcal{T}_H^\varepsilon$, which is defined by means of minimization of the relative entropy between π and a "reference measure" R^ε (details in the next section). So, our purpose is to prove that:

$$\mathcal{T}_H^\varepsilon(\mu,\nu) \leq \mathcal{T}_H^\varepsilon(\eta,\mu) \tag{2}$$

for all $\eta \leq_c \nu$ with finite "Shannon information". It has already been proved by Mikami (see [31]) and Léonard (see [29,30]) that passing to the limit we get: $\lim_{\varepsilon\to 0} \varepsilon\mathcal{T}_H^\varepsilon = \frac{1}{2}W_2^2$, so by letting $\varepsilon \to 0$ in (2) we immediately get (1).

In the next section we introduce the entropic transport costs, the Ornstein-Uhlenbeck process which gives us the reference measure R^ε, the concept of relative entropy as well as some results and representation formulas of the relative entropy.

2. ENTROPIC TRANSPORT COSTS AND RELATIVE ENTROPY

2.1. **Entropic costs, their zero noise limit and other results.**

Suppose we have the following stochastic differential equation:

$$dZ_t = -\frac{1}{2}Z_t dt + dW_t, \qquad t \geq 0$$

where $(W_t)_{t\geq 0}$ is the standard d-dimensional Brownian motion and Z_0 has distribution γ_d, $Z_0 \sim \gamma_d$. We will apply Itô's formula to find an explicit representation for its solution. To this end, let

$$f(t,x) = xe^{\frac{t}{2}}$$

Then:

$$df(t,Z_t) = \frac{\partial f}{\partial t}(t,Z_t)dt + \frac{\partial f}{\partial x}(t,Z_t)dZ_t + \frac{1}{2}\frac{\partial^2 f}{\partial x^2}(t,Z_t)dt =$$

$$= \frac{1}{2}Z_t e^{\frac{t}{2}}dt + e^{\frac{t}{2}}(-\frac{1}{2}Z_t dt + dW_t) = e^{\frac{t}{2}}dW_t$$

which is equivalent to:

$$Z_t e^{\frac{t}{2}} = Z_0 + \int_0^t e^{\frac{s}{2}} dW_s \ .$$

Finally, the solution is given by:

$$Z_t = Z_0 e^{-\frac{t}{2}} + e^{-\frac{t}{2}}\int_0^t e^{\frac{s}{2}} dW_s, \qquad t \geq 0 \ .$$

In order to define the entropic transport cost, we denote by R^ε the law of the random vector (Z_0, Z_ε). If we set

$$I_\varepsilon = \int_0^\varepsilon e^{\frac{s-\varepsilon}{2}} dW_s$$

it is known that I_ε is distributed as a normal random vector with zero mean and variance

$$\mathbb{E}\big(\int_0^\varepsilon (e^{\frac{s-\varepsilon}{2}})^2 ds\big)$$

hence

$$I_\varepsilon \sim \mathcal{N}(0, 1-e^{-\varepsilon}) \sim \sqrt{1-e^{-\varepsilon}}\mathcal{N}(0,1) \ .$$

From all these we deduce that

$$R^\varepsilon = \text{Law}(Z_0, Z_\varepsilon) = \text{Law}(X, Xe^{-\frac{\varepsilon}{2}} + \sqrt{1-e^{-\varepsilon}}Y)$$

where X, Y are independent standard Gaussian random vectors on $\mathbb{R}^d$. We would like to argue that the measure R^ε is in fact equal to:

$$R^\varepsilon(dx, dy) = \gamma_d(dx) r_x^\varepsilon(dy)$$

or equivalently,

$$R^\varepsilon(A \times B) = \int\limits_A r_x^\varepsilon(B) d\gamma_d(x)$$

where, for fixed x, the measure $r_x^\varepsilon(\cdot)$ is defined as:

$$r_x^\varepsilon(A) = \int\limits_A \frac{1}{(2\pi(1-e^{-\varepsilon}))^{\frac{d}{2}}} e^{-\frac{\|y - xe^{-\frac{\varepsilon}{2}}\|^2}{2(1-e^{-\varepsilon})}} d\lambda^d(y) \ .$$

For simplicity, we will restrict ourselves to $d = 1$, but of course the argument is the same for higher dimensions. For convenience, let $c = e^{-\frac{\varepsilon}{2}}, k = \sqrt{1 - e^{-\varepsilon}}$ and also

$$g : \mathbb{R}^2 \to \mathbb{R}^2$$

$$(x, y) \mapsto (x, cx + yk)$$

Then we get:

$$R^\varepsilon((-\infty, a] \times (-\infty, b]) = \mathbb{P}\left\{\omega : (X(\omega), Y(\omega)) \in g^{-1}((-\infty, a] \times (-\infty, b])\right\} =$$

$$= \iint\limits_{g^{-1}((-\infty,a]\times(-\infty,b])} f_{(X,Y)} \mathrm{d}s\mathrm{d}t = \int\limits_{-\infty}^{a} f_X(s) \left(\int\limits_{-\infty}^{\frac{b-cs}{k}} f_Y(t)\mathrm{d}t \right) \mathrm{d}s =$$

$$= \int\limits_{-\infty}^{a} f_X(s) \left(\int\limits_{-\infty}^{\frac{b-cs}{k}} \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}} \mathrm{d}t \right) \mathrm{d}s = \int\limits_{-\infty}^{a} f_X(s) \left(\int\limits_{-\infty}^{b} \frac{e^{-\frac{(t-sc)^2}{2k^2}}}{\sqrt{2\pi k^2}} \mathrm{d}t \right) \mathrm{d}s =$$

$$= \int\limits_{-\infty}^{a} \int\limits_{-\infty}^{b} \frac{e^{-\frac{s^2}{2}}}{\sqrt{2\pi}} \frac{e^{-\frac{(t-sc)^2}{2k^2}}}{\sqrt{2\pi k^2}} \mathrm{d}t\mathrm{d}s$$

which is the desired equality.

Another very important concept is that of the relative entropy. For a probability measure α which is absolutely continuous with respect to another probability measure β on some probability space $(\mathcal{X}, \mathcal{A})$, the relative entropy is defined by

$$H(\alpha|\beta) := \int \log\left(\frac{d\alpha}{d\beta}\right) d\alpha$$

If α is not absolutely continuous with respect to β, then we set $H(\alpha|\beta) = +\infty$. Using the convention $0 \cdot \infty = 0$, one sees that in fact we integrate over the set where the measure α is non-vanishing. Also, it holds that the function $s(\log(s))^-$ is bounded on $(0, \infty)$ (because $\lim_{x\to 0^+} x \log x = 0$), so

$$\int \left(\log\left(\frac{d\alpha}{d\beta}\right)\right)^- \frac{d\alpha}{d\beta} d\beta < \infty$$

thus H is well defined. The fact that $x\log(x) \geq x - 1$ with equality iff $x = 1$ implies that $H(\alpha|\beta) \geq 0$ and $H(\alpha|\beta) = 0 \iff \alpha = \beta$. We are ready to give the following:

Definition 3. *(Entropic transport cost) For $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, the entropic transport cost associated to R^ε is defined by*

$$\mathcal{T}_H^\varepsilon = \inf_{\pi \in C(\mu,\nu)} H(\pi|R^\varepsilon) \ .$$

It has already been shown by Mikami, Léonard and others (see [29,31]) that $\lim_{\varepsilon\to 0} \varepsilon \mathcal{T}_H^\varepsilon = \frac{1}{2}W_2^2$. By using intuitive arguments, we expect this to be true since:

$$\varepsilon \mathcal{T}_H^\varepsilon = \varepsilon \int \log\left(\frac{d\pi}{dx}\frac{dx}{dR^\varepsilon}\right) d\pi = \varepsilon \int \log\left(\frac{d\pi}{dx}\right) d\pi - \varepsilon \int \log\left(\frac{dR^\varepsilon}{dx}\right) d\pi$$

Using the formula of the previous page for the density of R^ε with respect to the Lebesgue measure, this equals:

$$\varepsilon \int \log\left(\frac{d\pi}{dx}\right) d\pi - \varepsilon \int -\frac{\|x\|^2}{2} + \log(\sqrt{2\pi}) d\pi + \frac{\varepsilon}{2(1-e^{-\varepsilon})} \int \|y - xe^{-\frac{\varepsilon}{2}}\|^2 d\pi +$$

$$+\varepsilon(\log(\sqrt{2\pi}) + \log(\sqrt{1-e^{-\varepsilon}}))$$

The last term is independent of μ, ν, π and goes to zero, when $\varepsilon \to 0$. Also $\int \|x\|^2 d\pi(x,y) = \int \|x\|^2 d\mu(x) < \infty$, when $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, so, for small ε, minimizing $\pi \mapsto H(\pi|R^\varepsilon)$ amounts to minimizing $\pi \mapsto \frac{1}{2}\int \|x-y\|^2 d\pi(x,y)$. In the sequel we will say that a probability measure η is of finite Shannon entropy if it is absolutely continuous with respect to the Lebesgue measure and if $\int \log(\frac{d\eta}{d\lambda^d}) d\eta < +\infty$. By splitting the derivative as before, we get that $\log(\frac{d\eta}{d\lambda^d}) = \log(\frac{d\eta}{d\gamma_d}) + \log(\frac{d\gamma_d}{d\lambda^d})$, which means that if $\eta \in \mathcal{P}_2(\mathbb{R}^d)$, then it is of finite Shannon entropy iff $H(\eta|\gamma_d) < +\infty$. We will also use the following result (see Carlier, Duval, Peyré, Schmitzer [4]):

Theorem 4. *(Carlier et al.) Assume $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ are of finite Shannon entropy. Then it holds*

$$\varepsilon \mathcal{T}_H^\varepsilon(\mu,\nu) \xrightarrow{\varepsilon\to 0} \frac{1}{2}W_2^2(\mu,\nu)$$

In order to use the latter, we state a Lemma which will be proved in Section 3:

Lemma 5. *If μ, ν satisfy the assumptions of Theorem 1, then they are of finite Shannon entropy.*

2.2. **Relation of entropic cost to Caffarelli's theorem.**

Theorem 6. *Let μ and ν satisfy the assumptions of Theorem 1 and additionally assume that the function V is bounded from below. If η is such that $\eta \leq_c \nu$, then for all $\varepsilon > 0$ it holds*

$$\mathcal{T}_H^{\varepsilon}(\mu, \nu) \leq \mathcal{T}_H^{\varepsilon}(\mu, \eta) \ .$$

Having Theorem 6 at our hands we can complete the proof of Theorem 1:

Proof of Theorem 1. Step 1: V is bounded from below. By Lemma 5, μ, ν have finite Shannon entropy, hence if we take an η which also has finite Shannon entropy and is $\eta \leq_c \nu$, we can combine Theorems 4 and 6 to deduce that

$$W_2(\mu, \nu) \leq W_2(\mu, \eta) \ .$$

We would like the above inequality to hold not only for those $\eta \leq_c \nu$ which have finite Shannon entropy but for *every* η with $\eta \leq_c \nu$. To this end, fix a compactly supported probability measure ν_0 of the form $\nu_0(dx) = e^{-W_0(x)}\gamma_d(dx)$, where $W_0 : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is convex and take an arbitrary $\eta \leq_c \nu_0$. For $\theta \in (0, \pi/2)$ define

$$\nu_\theta = \mathrm{Law}((\cos\theta)X + (\sin\theta)Z), \quad \eta_\theta = \mathrm{Law}((\cos\theta)Y + (\sin\theta)Z)$$

where $X \sim \nu_0$, $Y \sim \eta$ and Z is independent of X and Y having density $\frac{1}{C}\mathbb{I}_B(x)e^{-\frac{\|x\|^2}{2}}$, where B is the Euclidean unit ball and C is some normalizing constant. In Lemma 14 we will prove that ν_θ has compact support and is of the form $\nu_\theta(dx) = e^{-W_\theta(x)}\gamma_d(dx)$, where W_θ is convex and also that η_θ has finite Shannon entropy and satisfies $\eta_\theta \leq_c \nu_\theta$. This means that

$$W_2(\mu, \nu_\theta) \leq W_2(\mu, \eta_\theta) \ .$$

From this, we would like to pass to the limit as $\theta \to 0$ and deduce that $W_2(\mu, \nu_0) \leq W_2(\mu, \eta)$ and then we are done with Step 1. In Lemma 19 we also prove that $\nu_\theta \xrightarrow[\theta\to 0]{w} \nu_0$ and $\int \|y\|^2 d\nu_\theta(y) \xrightarrow{\theta\to 0} \int \|y\|^2 d\nu_0(y)$ which are of course equivalent to $W_2(\mu,\nu_\theta) \xrightarrow{\theta\to 0} W_2(\mu,\nu_0)$. Similar facts hold for the measures η_θ and η as well.

Step 2: V is not necessarily bounded from below. Since any convex function can be written as a supremum of affine functions, we can find an $\alpha \in \mathbb{R}^d$ such that $x \mapsto V(x) + <\alpha, x>$ is bounded from below. Let $T:\mathbb{R}^d \to \mathbb{R}^d$ with $T(x) = x+\alpha$ and set $\tilde{\mu} = \mu\circ T^{-1}$. Then

$$\tilde{\mu}(A) = \mu(T^{-1}(A)) = \int\limits_{T^{-1}(A)} e^{V(x)} d\gamma_d(x) = \int\limits_A e^{V(y-\alpha)} d\gamma_d(T^{-1}(y)) =$$

$$= \int\limits_A e^{V(y-\alpha)} d\gamma_d(y-\alpha) = \int\limits_A e^{V(y-\alpha)} \frac{e^{-\frac{\|y-\alpha\|^2}{2}}}{(\sqrt{2\pi})^d} d\lambda^d(y) =$$

$$= \int\limits_A e^{V(y-\alpha)} e^{\langle\alpha,(y-\frac{\alpha}{2})\rangle} d\gamma_d(y) = \int\limits_A e^{V(y-\alpha)} e^{\langle\alpha,(y-\alpha)\rangle} e^{\frac{\|\alpha\|^2}{2}} d\gamma_d(y)$$

so we see that $\tilde{\mu}$ satisfies our assumptions. Hence there exists a C^1 convex function $\tilde{\phi}$ such that $\nabla\tilde{\phi}$ is 1-Lipschitz and $\nu = \nabla\tilde{\phi}_{\#}\tilde{\mu}$. By setting $\phi(x) = \tilde{\phi}(x+\alpha)$, we get that $\nu = \nabla\phi_{\#}\mu$. □

In order to prove Theorem 6, we have to first discuss some important properties of the relative entropy, which of course have their own independent interest. These properties are the variational formula, the factorization property, convexity and lower semicontinuity and the Donsker-Varadhan dual representation of the relative entropy.

Proposition 7. *Let $(\mathcal{N},\mathcal{A})$ be a measurable space, θ a probability measure on $\mathcal{N}$ and $k:\mathcal{N}\to\mathbb{R}$ a bounded measurable function. Then the following hold:*

(α) *We have the variational formula*

$$\inf_{\gamma\in\mathcal{P}(\mathcal{N})} \left\{ H(\gamma|\theta) + \int_{\mathcal{N}} k d\gamma \right\} = -\log(\int_{\mathcal{N}} e^{-k} d\theta)$$

(b) *The above infimum is uniquely attained at γ_0, which is absolutely continuous with respect to θ and satisfies*

$$\frac{d\gamma_0}{d\theta}(x) := e^{-k(x)} \frac{1}{\int_{\mathcal{N}} e^{-k} d\theta}$$

Proof. For part (α) it suffices to restrict ourselves to the

$$\inf\left\{H(\gamma|\theta)+\int_{\mathcal{N}} k d\gamma : \gamma\in\mathcal{P}(\mathcal{N}), H(\gamma|\theta)<+\infty\right\}$$

If $H(\gamma|\theta)<+\infty$ then $\gamma\ll\theta$ and since also $\theta\ll\gamma_0$ (the density $\frac{d\gamma_0}{d\theta}$ is strictly posiitive, so the two measures are mutually absolutely continuous) we get that $\gamma\ll\gamma_0$. Hence:

$$H(\gamma|\theta)+\int_{\mathcal{N}} k d\gamma=\int_{\mathcal{N}}\left(\log\frac{d\gamma}{d\gamma_0}\right)d\gamma+\int_{\mathcal{N}}\left(\log\frac{d\gamma_0}{d\theta}\right)d\gamma+\int_{\mathcal{N}} k d\gamma=$$

$$=H(\gamma|\gamma_0)-\log\left(\int_{\mathcal{N}} e^{-k}d\theta\right)$$

Now, since $H(\gamma|\gamma_0)\geq 0$ and is zero iff $\gamma=\gamma_0$, the proof of both parts (α) and (b) is complete. □

In order to prove the other properties of the relative entropy, we have to speak about the concept of a **stochastic kernel**. Let $(\mathcal{N},\mathcal{A})$ be a measurable space, $\mathcal{Y}$ a Polish space and let $\{\tau(dy|x)\}_{x\in\mathcal{N}}$ be a family of probability measures in $\mathcal{Y}$ parametrized by $x\in\mathcal{N}$. We call $\tau(dy|x)$ a **stochastic kernel on $\mathcal{Y}$ given $\mathcal{N}$** if for every Borel subset $E\in\mathcal{B}(\mathcal{Y})$, the function $x\in\mathcal{N}\mapsto\tau(E|x)\in[0,1]$ is measurable. In order to establish an equivalent condition that characterizes a stochastic kernel, we will first state a technical lemma without proof:

Lemma 8. *For $E\in\mathcal{B}(\mathcal{Y})$ define $f_E:\mathcal{P}(\mathcal{Y})\to[0,1]$ by $f_E(\theta)=\theta(E)$. Then*

$$\mathcal{B}_{\mathcal{P}(\mathcal{Y})}=\sigma\left[\bigcup_{E\in\mathcal{B}(\mathcal{Y})} f_E^{-1}(\mathcal{B}(\mathbb{R})\right]$$

In other words, $\mathcal{B}_{\mathcal{P}(\mathcal{Y})}$ is the smallest σ-algebra with respect to which f_E is measurable for every E.

The following theorem gives a useful equivalent characterization of a stochastic kernel:

Theorem 9. *Let $\{\tau(dy|x)\}_{x\in\mathcal{N}}$ be a family of probability measures in $\mathcal{Y}$. Then $\tau(dy|x)$ is a stochastic kernel if and only if the mapping $x\in\mathcal{N}\to\tau(\cdot|x)\in\mathcal{P}(\mathcal{Y})$ is $\mathcal{A}/\mathcal{B}_{\mathcal{P}(\mathcal{Y})}$ measurable.*

Proof. Let $g:\mathcal{N}\to\mathcal{P}(\mathcal{Y})$ with $g(x)=\tau(\cdot|x)$ and define, for $E\in\mathcal{B}(\mathcal{Y})$, $h_E:\mathcal{N}\to[0,1]$ by $h_E(x)=\tau(E|x)$. By Lemma 8, it holds that $h_E=f_E\circ g$. Then, the Theorem just states that g is $\mathcal{A}/\mathcal{B}_{\mathcal{P}(\mathcal{Y})}$ measurable iff h_E is $\mathcal{A}$ measurable for every E.

Lemma 8 implies that f_E is $\mathcal{B}_{\mathcal{P}(\mathcal{Y})}$ measurable for every E. Since $h_E = f_E \circ g$, it follows that if g is $\mathcal{A}/\mathcal{B}_{\mathcal{P}(\mathcal{Y})}$ measurable then h_E is $\mathcal{A}$ measurable for every E. Conversely, if h_E is $\mathcal{A}$ measurable for every E, then using Lemma 8 we get:

$$g^{-1}\left(\mathcal{B}_{\mathcal{P}(\mathcal{Y})}\right) = g^{-1}\left(\sigma\left[\bigcup_{E\in\mathcal{B}(\mathcal{Y})} f_E^{-1}(\mathcal{B}(\mathbb{R})\right]\right) = \sigma\left[\bigcup_{E\in\mathcal{B}(\mathcal{Y})} g^{-1}\left(f_E^{-1}(\mathcal{B}(\mathbb{R})\right)\right] =$$

$$= \sigma\left[\bigcup_{E\in\mathcal{B}(\mathcal{Y})} h_E^{-1}(\mathcal{B}(\mathbb{R})\right] \subset \mathcal{A}$$

hence g is $\mathcal{A}/\mathcal{B}_{\mathcal{P}(\mathcal{Y})}$ measurable. □

Our next theorem shows that a probability measure defined on some product of spaces can be decomposed into its first marginal and a stochastic kernel. Of course, an analogous decomposition holds in terms of its second marginal.

Theorem 10. *Let $\tau = \tau(dx \times dy)$ be a probability measure on $\mathcal{N} \times \mathcal{Y}$ with the product σ-algebra $\mathcal{A} \bigotimes \mathcal{B}_{\mathcal{Y}}$. Denote by τ_1 the first marginal of τ, i.e. $\tau_1(A) = \tau(A \times \mathcal{Y})$. Then there exists a stochastic kernel $\tau(dy|x)$ on $\mathcal{Y}$ given $\mathcal{N}$ such that:*

$$\tau(A \times B) = \int_A \tau(B|x)\tau_1(dx)$$

for all $A \in \mathcal{A}$ and $B \in \mathcal{B}_{\mathcal{Y}}$. We denote this decomposition by $\tau(dx \times dy) = \tau_1(dx) \otimes \tau(dy|x)$.

Proof. (Sketch) On the product space $(\mathcal{N} \times \mathcal{Y}, \mathcal{A} \bigotimes \mathcal{B}_{\mathcal{Y}}, \tau)$ take the coordinate functions $\tilde{X}(x, y) = x, \tilde{Y}(x, y) = y$. The stochastic kernel is just the regular conditional distribution of $\tilde{Y}$ given $\tilde{X} = x$. □

The following theorem will be used in the proof of the chain rule. It deals with the existence of a Radon-Nikodym derivative between two stochastic kernels which is a jointly measurable function on the product space. More precisely we have:

Theorem 11. *Let $\mathcal{N}$ be a Polish space and $\mathcal{A}$ be its Borel σ-algebra. Let $A \in \mathcal{A}$ and $\sigma(dy|x), \tau(dy|x)$ be two stochastic kernels on $\mathcal{Y}$ given $\mathcal{N}$ with the property that for every $x \in A$, $\sigma(\cdot|x)$ is absolutely continuous with respect to $\tau(\cdot|x)$. Then there exists a version of the Radon-Nikodym derivative*

$$f(x, y) = \frac{d\sigma(\cdot|x)}{d\tau(\cdot|x)}(y)$$

which is a nonnegative measurable function of $(x, y) \in A \times \mathcal{Y}$.

We are now ready to prove the most crucial Theorem of this section.

Theorem 12. *Let X and Y be Polish spaces. The relative entropy $H(\cdot|\cdot)$ has the following properties:*

(α) (Donsker-Varadhan formula) For each γ and θ in $\mathcal{P}(X)$ it holds

$$H(\gamma|\theta) = \sup_{g \in C_b(X)} \left\{ \int_X g d\gamma - \log \left(\int_X e^g d\theta \right) \right\} =$$

$$= \sup_{\psi \in \Psi_b(X)} \left\{ \int_X \psi d\gamma - \log \left(\int_X e^\psi d\theta \right) \right\}$$

where $C_b(X)$ and $\Psi_b(X)$ are the spaces of continuous bounded and measurable bounded functions on X respectively.

(b) $H(\gamma|\theta)$ is a convex, lower semicontinuous function of $(\gamma, \theta) \in \mathcal{P}(X) \times \mathcal{P}(X)$. In particular, it is convex and lower semicontinuous on each variable separately. In addition, for each fixed θ, $H(\cdot|\theta)$ is a strictly convex function on the set $\{\gamma \in \mathcal{P}(X) : H(\gamma|\theta) < +\infty\}$.

(c) For each $\theta \in \mathcal{P}(X)$ and for each $M < +\infty$, the set $\{\gamma \in \mathcal{P}(X) : H(\gamma|\theta) \leq M\}$ is compact.

(d) Let α and β two probability measures on $X \times Y$ and denote by α_1 and β_1 their first marginals and $\alpha(dy|x), \beta(dy|x)$ the stochastic kernels on Y given X for which we have the decompositions

$$\alpha(dx \times dy) = \alpha_1(dx) \otimes \alpha(dy|x) \quad \textit{and} \quad \beta(dx \times dy) = \beta_1(dx) \otimes \beta(dy|x)$$

Then the mapping $x \in X \mapsto H(\alpha(\cdot|x)|\beta(\cdot|x))$ is measurable and

$$H(\alpha|\beta) = H(\alpha_1|\beta_1) + \int_X H(\alpha(\cdot|x)|\beta(\cdot|x))\alpha_1(dx) \ .$$

In particular, let $\sigma(dy|x), \tau(dy|x)$ be stochastic kernels on Y given X and θ a probability measure on X. Then the mapping $x \in X \mapsto H(\sigma(\cdot|x)|\tau(\cdot|x))$ is measurable and it holds

$$H(\theta \otimes \sigma|\theta \otimes \tau) = \int_X H(\sigma(\cdot|x)|\tau(\cdot|x))d\theta(x) \ .$$

Proof. (b) We will use the first part. For fixed $g \in C_b(X)$, the mapping

$$(\gamma,\theta) \in \mathcal{P}(X) \times \mathcal{P}(X) \mapsto \int_X g d\gamma - \log\left(\int_X e^g d\theta\right)$$

is convex and continuous, hence $H(\gamma|\theta)$ is convex and lower semicontinuous, as a supremum over $g \in C_b(X)$ of such functions. The strict convexity of $H(\gamma|\theta)$ on the set $\{\gamma \in \mathcal{P}(X) : H(\gamma|\theta) < +\infty\}$ comes from the strict convexity of $s \log s$ (for $s > 0$) and the fact that on the above set we have

$$H(\gamma|\theta) = \int_X \frac{d\gamma}{d\theta} \log\left(\frac{d\gamma}{d\theta}\right) d\theta$$

(c) Let $\{\gamma_n\}_n$ be a sequence in $\mathcal{P}(X)$ satisfying $\sup_n H(\gamma_n|\theta) \leq M < +\infty$. Again, using the variational formula of Varadhan-Donsker we get that for any bounded measurable function $\psi : X \to \mathbb{R}$ it holds

$$\int_X \psi d\gamma_n - \log\left(\int_X e^{\psi} d\theta\right) \leq H(\gamma_n\|\theta) \leq M$$

We will prove that the sequence $\{\gamma_n\}$ is tight. To this end, let $\delta > 0$. Choose $\varepsilon > 0$ such that $\frac{M+\log 2}{\log(1+1/\varepsilon)} < \delta$. Since the single measure θ is tight, there exists a compact set K such that $\theta(K^c) \leq \varepsilon$. In the last display, choose ψ to be 0 on K and $\log(1+1/\varepsilon)$ on K^c. What we get is:

$$\begin{aligned}
\gamma_n(K^c) &\leq \left[\frac{1}{\log(1+1/\varepsilon)}\right]\left(M + \log\left(\theta(K) + (1+\frac{1}{\varepsilon})\theta(K^c)\right)\right) = \\
&= \left[\frac{1}{\log(1+1/\varepsilon)}\right]\left(M + \log\left(1 + \frac{1}{\varepsilon}\theta(K^c)\right)\right) \leq \\
&\leq \left[\frac{1}{\log(1+1/\varepsilon)}\right](M + \log 2) < \delta
\end{aligned}$$

and this holds for all n. By Prohorov's Theorem, there exists $\gamma \in \mathcal{P}(X)$ and a subsequence γ_{n_k} such that $\gamma_{n_k} \Longrightarrow \gamma$. By lower semicontinuity of $H(\cdot|\theta)$ we deduce that $H(\gamma|\theta) \leq \liminf_k H(\gamma_{n_k}|\theta) \leq M$, which yields that the set $\{\gamma \in \mathcal{P}(X) : H(\gamma|\theta) \leq M\}$ is compact.
(a) We will first show that

$$\sup_{g \in C_b(X)} \left\{\int_X g d\gamma - \log\left(\int_X e^g d\theta\right)\right\} = \sup_{\psi \in \Psi_b(X)} \left\{\int_X \psi d\gamma - \log\left(\int_X e^{\psi} d\theta\right)\right\}$$

Since $C_b(X) \subset \Psi_b(X)$, the LHS is smaller or equal than the RHS. For the opposite inequality, given $\varepsilon > 0$, since γ, θ are tight, we can find a compact set K such that $\gamma(K^c) < \varepsilon$, $\theta(K^c) < \varepsilon$. Pick $\psi \in \Psi_b(X)$. Then $\psi|_K$ is also a measurable function, so by weak Lusin's theorem, since the measure $\gamma + \theta$ is finite, we can find a closed subset $F \subset K$ such that the restriction $\psi|_F$ is in fact continuous and $(\gamma + \theta)(K \backslash F) < \varepsilon$. Let us call $\psi|_F = g$. Since F is a closed set, we can use Tietze's extension theorem and find a continuous $\tilde{g} : X \to \mathbb{R}$ extending g with $\sup_{x \in X} |\tilde{g}(x)| = \sup_{y \in F} |g(y)|$. Since $\psi|_F = g$, we actually have constructed a continuous function $\tilde{g} : X \to \mathbb{R}$ with the property $\tilde{g}|_F = \psi|_F$, $\|\tilde{g}\|_\infty \leq \|\psi\|_\infty$ and $(\gamma + \theta)(K \backslash F) < \varepsilon$. It follows that

$$\gamma(F^c) \leq \gamma(K^c) + \gamma(K \backslash F) \leq 2\varepsilon \quad \text{and} \quad \theta(F^c) \leq \theta(K^c) + \theta(K \backslash F) \leq 2\varepsilon$$

and we claim that there is a constant C independent of ε such that:

$$\int_X \psi d\gamma - \log \left(\int_X e^\psi d\theta \right) \leq \int_X \tilde{g} d\gamma - \log \left(\int_X e^{\tilde{g}} d\theta \right) + C\varepsilon$$

In fact, we have that

$$\int_X (\psi - \tilde{g}) d\gamma = \int_{F^c} (\psi - \tilde{g}) d\gamma \leq 2\|\psi\|_\infty \gamma(F^c) = 4\|\psi\|_\infty \varepsilon$$

and

$$\int_X e^{\tilde{g}} d\theta = \int_F e^\psi d\theta + \int_{F^c} e^{\tilde{g}} d\theta \leq \int_X e^\psi d\theta + e^{\|\tilde{g}\|_\infty} \theta(F^c) \leq$$

$$\leq \int_X e^\psi d\theta + e^{\|\psi\|_\infty} 2\varepsilon$$

By taking logarithms and using the mean value theorem on the positive interval $(\int_X e^\psi d\theta, \int_X e^\psi d\theta + e^{\|\psi\|_\infty} 2\varepsilon)$ we find some ξ_ε in this interval such that:

$$\log \left(\int_X e^{\tilde{g}} d\theta \right) \leq \log \left(\int_X e^\psi d\theta + e^{\|\psi\|_\infty} 2\varepsilon \right) =$$

$$= \frac{1}{\xi_\varepsilon} e^{\|\psi\|_\infty} 2\varepsilon + \log \left(\int_X e^\psi d\theta \right) \leq \frac{1}{\int_X e^\psi d\theta} e^{\|\psi\|_\infty} 2\varepsilon + \log \left(\int_X e^\psi d\theta \right)$$

so if we take $C = 4\|\psi\|_\infty + \frac{2}{\int_X e^\psi d\theta} e^{\|\psi\|_\infty}$ the claim is proved.
Now we can take first the supremum over $g \in C_b(X)$ and then let $\varepsilon \to 0$ so we get

$$\int_X \psi d\gamma - \log \left(\int_X e^\psi d\theta \right) \leq \sup_{\tilde{g} \in C_b(X)} \left\{ \int_X \tilde{g} d\gamma - \log \left(\int_X e^{\tilde{g}} d\theta \right) \right\}$$

and since $\psi \in \Psi_b(X)$ was arbitrary, we conclude that

$$\sup_{\psi\in\Psi_b(X)}\left\{\int_X \psi d\gamma - \log\left(\int_X e^{\psi} d\theta\right)\right\} \leq \sup_{\tilde{g}\in C_b(X)}\left\{\int_X \tilde{g} d\gamma - \log\left(\int_X e^{\tilde{g}} d\theta\right)\right\}$$

We now define $R(\gamma,\theta)$ to be the common value of these suprema. Then $R(\gamma,\theta) \geq 0$, since we can take for example $g \equiv 0$. By part (α) of Proposition 7, we get that for all $k \in \Psi_b(X)$ it holds

$$H(\gamma|\theta) \geq -\int_X k d\gamma - \log\left(\int_X e^{-k} d\theta\right) .$$

Replacing k by $\psi := -k$ and then taking the supremum over all $\psi \in \Psi_b(X)$ we have:

$$H(\gamma|\theta) \geq \sup_{\psi\in\Psi_b(X)}\left\{\int_X \psi d\gamma - \log\left(\int_X e^{\psi} d\theta\right)\right\} = R(\gamma,\theta) .$$

In order to show the opposite inequality, we may of course assume that $R(\gamma,\theta) < +\infty$. In that case, we will prove that $\gamma \ll \theta$. Fix $r > 0$ and a Borel set A with $\theta(A) = 0$. Since for any $\psi \in \Psi_b(X)$ it holds

$$\int_X \psi d\gamma - \log\left(\int_X e^{\psi} d\theta\right) \leq R(\gamma,\theta) < \infty$$

then for $\psi := r\mathbb{I}_A$ we get that

$$\int_X \psi d\gamma - \log\left(\int_X e^{\psi} d\theta\right) = \int_X \psi d\gamma - \log(1) = r\gamma(A) \leq R(\gamma,\theta) < \infty$$

and by letting $r \to \infty$ we obtain that $\gamma(A) = 0$.
We can define the Radon-Nikodym derivative $f := \frac{d\gamma}{d\theta}$. Suppose that f is bounded and everywhere positive. Then $\psi := \log(f)$ is bounded and measurable hence we may apply the above formula for this ψ to get the desired inequality:

$$\int_X \log(f) d\gamma - \log\left(\int_X e^{\log(f)} d\theta\right) \leq R(\gamma,\theta) \implies$$

$$\int_X \log(f) d\gamma - \log\left(\int_X \frac{d\gamma}{d\theta} d\theta\right) \leq R(\gamma,\theta) \implies$$

$$H(\gamma|\theta) = \int_X \log(f) d\gamma \leq R(\gamma, \theta)$$

If f is everywhere positive but not bounded, then set $f_n = f \wedge n$ and $\psi := \log(f_n)$. By monotone convergence, we obtain:

$$H(\gamma|\theta) = \int_X \log(f) d\gamma = \lim_{n\to\infty} \int_X \log(f_n) d\gamma \leq$$

$$\leq R(\gamma, \theta) + \lim_{n\to\infty} \log\left(\int_X f_n d\theta\right) = R(\gamma, \theta) \ .$$

For the general case where f is neither bounded nor everywhere positive, define for $t \in [0, 1]$:

$$\gamma_t = t\theta + (1-t)\gamma \quad \text{and} \quad f_t = \frac{d\gamma_t}{d\theta} = t + (1-t)f \ .$$

For every $t \in (0, 1]$, f_t is everywhere positive, so by the previous argument we have $H(\gamma_t|\theta) \leq R(\gamma_t, \theta)$. We claim now that $\lim_{t\to 0} H(\gamma_t|\theta) = H(\gamma|\theta)$ and $\lim_{t\to 0} R(\gamma_t, \theta) = R(\gamma, \theta)$. Indeed, since $s\log(s)$ is convex on $(0, \infty)$ it holds that

$$H(\gamma_t|\theta) = \int_X f_t \log(f_t) d\theta \leq (1-t) \int_X f \log(f) d\theta = (1-t) H(\gamma|\theta) \ .$$

On the other hand, since $\log(s)$ is concave and strictly increasing, it also holds that $\log(f_t) \geq \log(t) \vee [(1-t)\log(f)]$, therefore:

$$H(\gamma_t|\theta) = t \int_X \log(f_t) d\theta + (1-t) \int_X f(\log(f_t)) d\theta \geq t\log(t) + (1-t)^2 H(\gamma|\theta)$$

Combining the above two displays, we get that $\lim_{t\to 0} H(\gamma_t|\theta) = H(\gamma|\theta)$. Now, since $R(\gamma, \theta) \leq H(\gamma|\theta)$, it holds that $R(\theta, \theta) = 0$. Also, it is a matter of routine calculations to show that the mapping $t \in [0, 1] \mapsto R(\gamma_t, \theta)$ is convex and lower semicontinuous. Hence, for $t \in [0, 1]$ we have:

$$0 \leq R(\gamma_t, \theta) \leq tR(\theta, \theta) + (1-t)R(\gamma, \theta) = (1-t)R(\gamma, \theta) \leq R(\gamma, \theta) < \infty$$

which means that this mapping is also bounded. Since any convex, lower semicontinuous, bounded function on a closed interval is continuous, we obtain that $t \in [0, 1] \mapsto R(\gamma_t, \theta)$ is continuous, and this yields that $\lim_{t\to 0} R(\gamma_t, \theta) = R(\gamma_0, \theta) = R(\gamma, \theta)$. The proof of Donsker-Varadhan variational formula is complete.

(d)Because of Theorem 9, we know that the stochastic kernels $\alpha(dy|x), \beta(dy|x)$ are measurable functions from X into $\mathcal{P}(Y)$. Since $H(\cdot|\cdot)$

is lower semicontinuous (by property (b)), it is also measurable, therefore the measurability of the mapping $x \in X \mapsto H(\alpha(\cdot|x)|\beta(\cdot|x))$ follows. We will now prove that

$$H(\alpha|\beta) = H(\alpha_1|\beta_1) + \int_X H(\alpha(\cdot|x)|\beta(\cdot|x))\alpha_1(dx)$$

Step 1: Suppose first that the RHS is finite. Under this assumption, $\alpha_1 \ll \beta_1$ and there is an α_1-null set Γ such that $H(\alpha(\cdot|x)|\beta(\cdot|x))$ is finite for $x \in \Gamma^c$. Hence, for $x \in \Gamma^c$, $\alpha(\cdot|x) \ll \beta(\cdot|x)$ as measures on Y. By redefining $\alpha(\cdot|x)$ on the null set Γ, we can assure that $\alpha(\cdot|x) \ll \beta(\cdot|x)$ for every $x \in X$. Let

$$\psi(x) := \frac{d\alpha_1}{d\beta_1}(x) \ .$$

Theorem 11 ensures us that there exists a version of the Radon-Nikodym derivative

$$\zeta(x,y) = \frac{d\alpha(\cdot|x)}{d\beta(\cdot|x)}(y)$$

which is nonnegative and measurable on $X \times Y$. For any Borel subsets A of X and B of Y, using the above derivatives and Fubini's theorem we get :

$$\alpha(A \times B) = \int_A \alpha(B|x)\alpha_1(dx) = \int_A \left(\int_B \zeta(x,y)\beta(dy|x) \right) \psi(x)\beta_1(dx) =$$

$$= \int_{A\times B} \psi(x)\zeta(x,y)\beta(dx \times dy)$$

This yields that $\alpha \ll \beta$ and

$$\frac{d\alpha}{d\beta}(x,y) = \psi(x)\zeta(x,y) \ .$$

Consequently we have:

$$H(\alpha_1|\beta_1) + \int_X H(\alpha(\cdot|x)|\beta(\cdot|x))\alpha_1(dx) =$$

$$= \int_X \log(\psi(x))\alpha_1(dx) + \int_X \left(\int_Y \log(\zeta(x,y))\alpha(dy|x) \right) \alpha_1(dx) =$$

$$= \int_{X\times Y} \log(\psi(x))\alpha(dx \times dy) + \int_{X\times Y} \log(\zeta(x,y))\alpha_1(dx) \otimes \alpha(dy|x) =$$

$$= \int\limits_{X\times Y} \log[\psi(x)\zeta(x,y)]\alpha(dx\times dy) = H(\alpha|\beta) \ .$$

Step 2: Suppose now that the LHS is finite. This yields that $\alpha \ll \beta$ so we can define

$$\phi(x,y) := \frac{d\alpha}{d\beta}(x,y) \ .$$

Since $\alpha \ll \beta$, we have that $\alpha_1 \ll \beta_1$ and we may define

$$\psi(x) := \frac{d\alpha_1}{d\beta_1}(x) \ .$$

If $A \subset X$ and $B \subset Y$ are Borel subsets, then

$$\int\limits_A \alpha(B|x)\psi(x)\beta_1(dx) = \int\limits_A \alpha(B|x)\alpha_1(dx) = \alpha(A\times B) =$$

$$= \int\limits_{A\times B} \phi(x,y)\beta(dx\times dy) = \int\limits_A \left(\int\limits_B \phi(x,y)\beta(dy|x)\right)\beta_1(dx) \ .$$

This implies that there exists a β_1-null set Γ such that for all $x \in \Gamma^c$ it holds

$$\psi(x)\alpha(B|x) = \int\limits_B \phi(x,y)\beta(dy|x) \ .$$

Thus, for all $x \in \Gamma^c \cap \{\psi > 0\}$, $\alpha(\cdot|x) \ll \beta(\cdot|x)$ and for such x and for all $y \in Y$:

$$\zeta(x,y) := \frac{d\alpha(\cdot|x)}{d\beta(\cdot|x)}(y) \quad \text{equals} \quad \frac{\phi(x,y)}{\psi(x)} \ .$$

In other words, for all $x \in \Gamma^c \cap \{\psi > 0\}$ the various derivatives are related by

$$\phi(x,y) = \psi(x)\zeta(x,y) \ .$$

We have

$$\alpha_1(\{\psi = 0\}) = \alpha(\{\psi = 0\}\times Y) = \int\limits_{\{\psi=0\}} \alpha(Y|x)\psi(x)\beta_1(dx) = 0$$

thus $\alpha_1(\{\psi > 0\}) = 1$, and since $\alpha_1 \ll \beta_1$ and $\beta_1(\Gamma^c) = 1$, we also have $\alpha_1(\Gamma^c) = 1$. Finally we obtain:

$$H(\alpha|\beta) = \int\limits_{X\times Y} \log(\phi(x,y))\alpha(dx\times dy) =$$

$$= \int\limits_{(\Gamma^c\cap\{\psi>0\})\times Y} \log(\psi(x)\zeta(x,y)\alpha_1(dx)\otimes\alpha(dy|x) =$$

$$= \int\limits_{\Gamma^c \cap \{\psi > 0\}} \log(\psi(x))\alpha_1(dx) + \int\limits_{\Gamma^c \cap \{\psi > 0\}} \left(\int\limits_Y \log(\zeta(x,y))\alpha(dy|x) \right) \alpha_1(dx) =$$

$$= H(\alpha_1|\beta_1) + \int\limits_X H(\alpha(\cdot|x)|\beta(\cdot|x))\alpha_1(dx) \ .$$

For the last claim of property (d), it suffices to take $\alpha = \theta \otimes \sigma$ and $\beta = \theta \otimes \tau$.

□

Before we finish our discussion about these basic properties of the relative entropy we must mention that in case where $\nu \ll \mu$, i.e. $H(\nu|\mu) < \infty$, there is a more useful version for the duality formula of the entropy namely that:

$$H(\nu|\mu) = \sup \left\{ \int\limits_X u d\nu - \log \left(\int\limits_X e^u d\mu \right) : \int\limits_X e^u d\mu < +\infty \quad , \int\limits_X u^- < +\infty \right\}$$

where $u^- = \max\{-u, 0\}$ and $\int u d\nu \in (-\infty, \infty]$ is well defined for u with $\int\limits_X u^- < +\infty$. The proof of this formula relies on Fenchel transformation of the convex function $h(t) = t\log(t) - t + 1$ and can be found in Proposition B.1 on the work "Transport Inequalities. A survey" by Gozlan and Leonard. Informally we can explain using this formula for the entropy why the statement of Theorem 6 is easier to prove at the level of entropic cost than trying to prove it directly for the Wasserstein distance. If we take $u = -(V + W)$, then $\int e^u d\mu = \int e^{-V-W} e^V d\gamma_d = \int d\nu < +\infty$, so if we also assume that $\int (-(V + W))^- d\rho < +\infty$ then:

$$H(\rho|\mu) \geq \int\limits_X -(V + W)d\rho \geq \int\limits_X -(V + W)d\nu = H(\nu|\mu)$$

as soon as $\rho \leq_c \nu$. Remember that our main goal is to infer that $W_2(\rho, \mu) \geq W_2(\nu, \mu)$, so comparison is easier in the entropy level when we deal with a log concavity condition on the relative density.

The next proposition gives us more information about the optimal coupling π^ε for $\mathcal{T}_H^\varepsilon(\mu, \nu)$:

Proposition 13. *Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ be such that $H(\mu|\gamma_d) < +\infty$ and $H(\nu|\gamma_d) < +\infty$.*

(1) There exists a unique coupling $\pi^\varepsilon \in C(\mu, \nu)$ such that

$$\mathcal{T}_H^\varepsilon(\mu, \nu) = H(\pi^\varepsilon|R^\varepsilon) < +\infty$$

(2) *There exist two measurable functions* $f^\varepsilon, g^\varepsilon : \mathbb{R}^d \to \mathbb{R}^+$ *such that* $\log(f^\varepsilon) \in L^1(\mu), \log(g^\varepsilon) \in L^1(\nu)$ *and*

$$\pi^\varepsilon(dxdy) = f^\varepsilon(x)g^\varepsilon(y)R^\varepsilon(dxdy) \ .$$

Proof. We endow the set $\mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ with the topology of weak convergence of probability measures. Under this topology, the set $C(\mu,\nu)$ is compact and because of Theorem 12-property (b), we know that the function $\pi \mapsto H(\pi|R^\varepsilon)$ is lower semicontinuous, hence it attains its minimum at some point π^ε of $C(\mu,\nu)$. We claim that the coupling $\pi_0 = \mu \otimes \nu$ is such that $H(\pi_0|R^\varepsilon) < +\infty$. In fact, we can apply the chain rule for the entropy (Theorem 12-property (d)) for $\alpha = \mu \otimes \nu$, $\alpha_1 = \mu$, $\alpha(dy|x) = \nu$, $\beta = R^\varepsilon$, $\beta_1 = \gamma_d$, $\beta(dy|x) = r^\varepsilon_x$. Then

$$H(\mu \otimes \nu|R^\varepsilon) = H(\mu|\gamma_d) + \int\limits_{\mathbb{R}^d} H(\nu|r^\varepsilon_x)d\mu(x)$$

The first term on the RHS is finite due to our assumption. The second term equals:

$$\int\limits_{\mathbb{R}^d}\left(\int\limits_{\mathbb{R}^d}\log(\frac{d\nu}{dr^\varepsilon_x})d\nu(y)\right)d\mu(x) = \int\limits_{\mathbb{R}^d}\left(\int\limits_{\mathbb{R}^d}\log(\frac{d\nu}{d\gamma_d}) - \log(\frac{dr^\varepsilon_x}{d\gamma_d})d\nu(y)\right)d\mu(x) =$$

$$= \int\limits_{\mathbb{R}^d} H(\nu|\gamma_d)d\mu(x) - \int\limits_{\mathbb{R}^d}\left(\int\limits_{\mathbb{R}^d}\log(\frac{dr^\varepsilon_x}{d\gamma_d})d\nu(y)\right)d\mu(x)$$

Again by assumption, the first integral equals $H(\nu|\gamma_d)\cdot 1$. For the finiteness of the second integral we write

$$r^\varepsilon_x(dy) = \frac{e^{-\frac{\|y-e^{-\frac{\varepsilon}{2}}x\|^2}{(2\pi(1-e^{-\varepsilon}))^{\frac{d}{2}}}}}{(2\pi(1-e^{-\varepsilon}))^{\frac{d}{2}}}\frac{e^{\frac{\|x\|^2}{2}}}{(2\pi)^{\frac{d}{2}}}\gamma_d(dy)$$

so when we take the logarithm of the derivative $dr^\varepsilon_x/d\gamma_d$ we will have to deal with the quantity $\|y - e^{-\frac{\varepsilon}{2}}x\|^2$. But it holds that

$$\|y - e^{-\frac{\varepsilon}{2}}x\|^2 \leq K\left(\|y\|^2 + e^{-\varepsilon}\|x\|^2\right)$$

for some constant K, so we can use the fact that the measures μ and ν have finite second moments and deduce that also the second integral is finite. Hence the claim is proved. Since the value $H(\pi^\varepsilon|R^\varepsilon)$ is the minimum value of $H(\cdot|R^\varepsilon)$, we get that $H(\pi^\varepsilon|R^\varepsilon) < +\infty$. Uniqueness follows from the strict convexity of $H(\cdot|R^\varepsilon)$. For the proof of the assertion (2), see for example [9, Corollary 3.2]. In the case that matters to us, namely when μ and ν satisfy our log-convexity/concavity assumptions we give a self-contained proof in Section 3. □

The following result states that in the setting of Theorem 1, a lot more can be said about the functions f^ε and g^ε:

Theorem 14. *With the same notation as in Proposition 13, let μ be a probability measure of the form $\mu(dx) = e^{V(x)}\gamma_d(dx)$ with finite second moment and ν be a compactly supported probability measure of the form $\nu(dx) = e^{-W(x)}\gamma_d(dx)$, with V, W convex and also V bounded from below. Then there exist a log-convex function $f^\varepsilon : \mathbb{R}^d \to [1, +\infty)$ and a log-concave function $g^\varepsilon : \mathbb{R}^d \to (0, +\infty)$ such that the unique optimal coupling $\pi^\varepsilon \in C(\mu, \nu)$ has the form $\pi^\varepsilon(dxdy) = f^\varepsilon(x)g^\varepsilon(y)R^\varepsilon(dxdy)$. Furthermore, $\log(f^\varepsilon) \in L^1(\mu)$, $\log(g^\varepsilon) \in L^1(\nu)$ and it holds*

$$\mathcal{T}_H^\varepsilon(\mu, \nu) = H(\pi^\varepsilon | R^\varepsilon) = \int_{\mathbb{R}^d} \log(f^\varepsilon) d\mu + \int_{\mathbb{R}^d} \log(g^\varepsilon) d\nu \ .$$

In the next Section we will prove Theorem 14. Having this in our hands, we can continue the proof of Theorem 6:

Proof of Theorem 6. We use the duality inequality of the relative entropy, i.e. if α, β are two probability measures with $H(\alpha|\beta) < +\infty$, then for any measurable function h such that $\int h d\alpha < +\infty$ and $\int e^h d\beta < +\infty$ it holds

$$H(\alpha|\beta) \geq \int h d\alpha - \log \left(\int e^h d\beta \right) \ . \tag{3}$$

Let η be a probability measure with $\eta \leq_c \nu$. If for every coupling $\pi \in C(\mu, \eta)$ it holds that $H(\pi|R^\varepsilon) = +\infty$, then obviously

$$+\infty = \inf_{\pi \in C(\mu,\eta)} H(\pi|R^\varepsilon) = \mathcal{T}_H^\varepsilon(\mu, \eta) \geq \mathcal{T}_H^\varepsilon(\mu, \nu)$$

so there is nothing special to prove in that case. Otherwise, pick $\pi \in C(\mu, \eta)$ with $H(\pi|R^\varepsilon) < +\infty$. Applying the duality inequality above to $\alpha = \pi$, $\beta = R^\varepsilon$ and $h(x, y) = \log(f^\varepsilon(x)g^\varepsilon(y))$, we get:

$$\begin{aligned}
H(\pi|R^\varepsilon) &\geq \int \log(f^\varepsilon(x)) + \log(g^\varepsilon(y))\pi(dxdy) - \log \left(\int f^\varepsilon(x)g^\varepsilon(y)dR^\varepsilon(x, y) \right) \\
&= \int \log(f^\varepsilon(x))\mu(dx) + \int \log(g^\varepsilon(y))\eta(dy) - \log \left(\int d\pi^\varepsilon(x, y) \right) \\
&\geq \int \log(f^\varepsilon(x))\mu(dx) + \int \log(g^\varepsilon(y))\nu(dy) = H(\pi^\varepsilon|R^\varepsilon) = \mathcal{T}_H^\varepsilon(\mu, \nu)
\end{aligned}$$

where the second inequality came from the fact that $\log(g^\varepsilon)$ is concave and $\eta \leq_c \nu$. Minimizing over π we obtain $\mathcal{T}_H^\varepsilon(\mu, \eta) \geq \mathcal{T}_H^\varepsilon(\mu, \nu)$. $\square$

Before we go on to the next section, we will mention some perspectives. One could wonder whether this idea of proof could also help us establish a version of Caffarelli's theorem in settings other than $\mathbb{R}^d$, for example on manifolds or free probability, e.g. see [21], [33] and [18] for the Schrödinger problem in a wider geometric setting. Another question is about non-local quantitative regularity estimates, such as those in [27,28]. The role of 1-Lipschitz bounds in Theorem 2 is very specific and it is not known if there is an analogue of that equivalence adapted to other types of regularity bounds. However, one could possibly prove stable a priori bounds for $\varepsilon \log(f^\varepsilon)$ and then pass to the limit. Of special interest is whether we can find integrated gradient bounds for non-uniformly convex potentials, since such bounds can be used to prove Poincaré inequalities [32,26]. In [12] there is a stability result for Caffarelli's theorem and one could focus on improving the quantitative bounds.

3. PROOFS OF THE MAIN RESULTS

In this section we develop the basic ideas needed to proof Theorem 6. These ideas were first introduced in a paper by Fortet [15]. Fortet's work was revisited in [13].
We will first give some notation. For $\varepsilon > 0$ and for a non-negative measurable function ψ, let P^ε be the function:

$$P^\varepsilon \psi(x) = \frac{1}{(2\pi)^{d/2}} \frac{1}{(1-e^{-\varepsilon})^{d/2}} \int\limits_{\mathbb{R}^d} \psi(y + e^{-\varepsilon/2}x) e^{-\frac{\|y\|^2}{2(1-e^{-\varepsilon})}} dy \ , \quad x \in \mathbb{R}^d$$

We claim that $\{P^\varepsilon\}_{\varepsilon \geq 0}$ is the transition semigroup of the solution of the Ornstein-Uhlenbeck stochastic differential equation discussed in section 2.1, namely it holds that:

$$P^\varepsilon \psi(x) = \mathbb{E}[\psi(Z_\varepsilon)|Z_0 = x] \ , \ x \in \mathbb{R}^d$$

Indeed, by the disintegration theorem (see for example Theorem 5.4 in [24]) it holds that

$$\mathbb{E}[\psi(Z_\varepsilon)|Z_0 = x] = \int\limits_{\mathbb{R}^d} \psi(y)\mu(x, dy)$$

where for fixed x, we have the probability $\mu(x, \cdot) = \mathbb{P}[Z_\varepsilon \in \cdot | Z_0 = x]$. Now, since we know the joint density of the random vector (Z_0, Z_ε) and the density of Z_0 we can compute explicitly the measure $\mu(x, B)$:

$$\mu(x, B) = \mathbb{P}[Z_\varepsilon \in B | Z_0 = x] = \int\limits_B \frac{f_{Z_0, Z_\varepsilon}(x, y)}{f_{Z_0}(x)} dy =$$

$$= \int\limits_B \frac{\frac{e^{-\frac{\|x\|^2}{2}}}{(2\pi)^{d/2}} \frac{e^{-\frac{\|y - xe^{-\varepsilon/2}\|^2}{2(1-e^{-\varepsilon})}}}{(2\pi(1-e^{-\varepsilon}))^{d/2}}}{\frac{e^{-\frac{\|x\|^2}{2}}}{(2\pi)^{d/2}}} dy = \int\limits_B \frac{e^{-\frac{\|y - xe^{-\varepsilon/2}\|^2}{2(1-e^{-\varepsilon})}}}{(2\pi(1-e^{-\varepsilon}))^{d/2}} dy$$

which means that

$$\mathbb{E}[\psi(Z_\varepsilon)|Z_0 = x] = \int\limits_{\mathbb{R}^d} \psi(y) \frac{e^{-\frac{\|y - xe^{-\varepsilon/2}\|^2}{2(1-e^{-\varepsilon})}}}{(2\pi(1-e^{-\varepsilon}))^{d/2}} dy$$

as desired.

Now suppose that $f^\varepsilon, g^\varepsilon$ are measurable non-negative functions such that $\pi^\varepsilon(dxdy) = f^\varepsilon(x)g^\varepsilon(y)R^\varepsilon(dxdy)$ belongs to $C(\mu,\nu)$. Writing down the marginal conditions, we can see that f^ε and g^ε are related to each other by the identities:

$$f^\varepsilon(x)P^\varepsilon g^\varepsilon(x) = e^{V(x)} \qquad \text{and} \qquad g^\varepsilon(y)P^\varepsilon f^\varepsilon(y) = e^{-W(y)} \ . \tag{4}$$

In fact, let us try to prove the first one (a similar argument can be applied to the second)

$$\mu(A) = \pi^\varepsilon(A\times\mathbb{R}^d) = \int\limits_{A\times\mathbb{R}^d} f^\varepsilon(x)g^\varepsilon(y)dR^\varepsilon(x,y) =$$

$$= \int\limits_{A\times\mathbb{R}^d} f^\varepsilon(x)g^\varepsilon(y)\gamma_d(x)r^\varepsilon_x(y)d(\lambda^d\times\lambda^d)(x,y) =$$

$$= \int\limits_A f^\varepsilon(x)\frac{e^{-\frac{\|x\|^2}{2}}}{(2\pi)^{d/2}}\int\limits_{\mathbb{R}^d} g^\varepsilon(y)\frac{e^{-\frac{\|y-xe^{-\varepsilon/2}\|^2}{2(1-e^{-\varepsilon})}}}{(2\pi(1-e^{-\varepsilon}))^{d/2}}d\lambda^d(y)d\lambda^d(x) \ .$$

But also

$$\mu(A) = \int_A e^{V(x)}\frac{e^{-\frac{\|x\|^2}{2}}}{(2\pi)^{d/2}}d\lambda^d(x)$$

hence

$$e^{V(x)} = f^\varepsilon(x)P^\varepsilon g^\varepsilon(x) \ .$$

These relations suggest to define the functional Φ^ε as follows: for every measurable function $h:\mathbb{R}^d\to\mathbb{R}\cup\{+\infty\}$

$$\Phi^\varepsilon(h) = V - \log\left(P^\varepsilon\left(e^{-W}\frac{1}{P^\varepsilon(e^h)}\right)\right) \ .$$

This means that a pair $(f^\varepsilon, g^\varepsilon)$ satisfies (4) if and only if $g^\varepsilon = e^{-W}\frac{1}{P^\varepsilon(f^\varepsilon)}$ and $f^\varepsilon = e^{h^\varepsilon}$ with h^ε being a fixed point of Φ, i.e.

$$h^\varepsilon = \Phi(h^\varepsilon) \ .$$

We could find the unknown function h^ε as a limit when $n\to+\infty$ of a sequence $\{h_n\}_n$ satisfying the recursion

$$h_{n+1} = \Phi^\varepsilon(h_n) \tag{5}$$

and some initial condition h_0. The above recursion is at the core of Sinkhorn's algorithm to approximate numerically the optimal transport via entropic regularization, see also [1,10].
The next Lemma assures us that if we begin with some initial convex function h_0, then h^ε will be also convex:

Lemma 15. *If $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is convex, then $\Phi^\varepsilon(h)$ is also convex.*

Proof. Step 1: If f is log-convex, then $P^\varepsilon(f)$ is log-convex. To prove this statement, we must check that for every $x_1, x_2 \in \mathbb{R}^d$ and $t \in [0,1]$

$$P^\varepsilon(f)(tx_1 + (1-t)x_2) \le [P^\varepsilon(f)(x_1)]^t [P^\varepsilon(f)(x_2)]^{1-t} \ .$$

For simplicity, write $M(y) = \frac{1}{C} e^{-\frac{\|y\|^2}{2(1-e^{-\varepsilon})}}$ where $C = (2\pi)^{d/2}(1-e^{-\varepsilon})^{d/2}$. Then:

$$\int_{\mathbb{R}^d} f[y + e^{-\varepsilon/2}(tx_1 + (1-tx_2))] e^{M(y)} dy =$$

$$= \int_{\mathbb{R}^d} f[t(y + e^{-\varepsilon/2}x_1) + (1-t)(y + e^{-\varepsilon/2}x_2)] e^{M(y)} dy \le$$

$$\le \int_{\mathbb{R}^d} [f(y + e^{-\varepsilon/2}x_1)]^t [f(y + e^{-\varepsilon/2}x_2)]^{1-t} e^{M(y)} dy =$$

$$= \int_{\mathbb{R}^d} [f(y + e^{-\varepsilon/2}x_1) e^{M(y)}]^t [f(y + e^{-\varepsilon/2}x_2) e^{M(y)}]^{1-t} dy$$

where the first inequality came from the log-convexity of f. Now we can apply Hölder's inequality with $p = \frac{1}{1-t}$ and $q = \frac{1}{t}$ and get that the last expression is smaller or equal to

$$\left(\int_{\mathbb{R}^d} \left\{ [f(y + e^{-\varepsilon/2}x_1) e^{M(y)}]^t \right\}^q dy \right)^{1/q} \cdot$$

$$\cdot \left(\int_{\mathbb{R}^d} \left\{ [f(y + e^{-\varepsilon/2}x_2) e^{M(y)}]^{1-t} \right\}^p dy \right)^{1/p}$$

which of course equals

$$[P^\varepsilon(f)(x_1)]^t \cdot [P^\varepsilon(f)(x_2)]^{1-t}$$

Step 2: If ψ is log-concave, then $P^\varepsilon(\psi)$ is log-concave. This follows from Prekopa's Theorem (see [34]), which states that the convolution of two log-concave functions is again log-concave. We just need to show that P^ε is in fact a convolution operator. To this end, we use the change of variables $u = -y \cdot e^{\varepsilon/2}$. Then

$$P^\varepsilon \psi(x) = \frac{(-1)^d}{(2\pi)^{d/2}} \frac{(e^{-\varepsilon/2})^d}{(1-e^{-\varepsilon})^{d/2}} \int_{\mathbb{R}^d} \psi(e^{-\varepsilon/2}(x-u)) e^{-\frac{\|u\|^2 \cdot e^{-\varepsilon}}{2(1-e^{-\varepsilon})}} du$$

So if we let $\tilde{\psi}(z) = \psi(e^{-\varepsilon/2}z)$ and $E(z) = e^{-\frac{\|z\|^2 \cdot e^{-\varepsilon}}{2(1-e^{-\varepsilon})}}$, then $P^\varepsilon\psi(x) = C \cdot (\tilde{\psi} * E)(x)$, where $C = \frac{(-1)^d}{(2\pi)^{d/2}} \frac{(e^{-\varepsilon/2})^d}{(1-e^{-\varepsilon})^{d/2}}$ We deduce than if ψ is log-concave, then also $\tilde{\psi}$ is log-concave and since of course E is log-concave we get that $P^\varepsilon\psi$ is log-concave.

$\square$

Remark 16. *A good question is if we can use this scheme of proof directly at the level of the Kantorovich dual optimal transport problem, rather than on the regularized version. The answer might be no, as in the limit while the minimizers in the dual formulation of entropic transport , suitably rescaled, converge to the Kantorovich potentials, the fixed point problem becomes degenerate in the limit and only selects c-convex functions (here the cost function c is just the quadratic distance), and we lose uniqueness. This is why Sinkhorn's algorithm approximates numerically the regularized problem* (see [10]).

The next Lemma presents two other very useful properties of the operators Φ^ε and P^ε:

Lemma 17.

(1) The operator P^ε maps $L^2(\gamma_d)$ to $L^2(\gamma_d)$ and it is symmetric.

(2) The operator Φ^ε is monotone in the sense that if $h \le k$ then $\Phi^\varepsilon(h) \le \Phi^\varepsilon(k)$.

(3) For any measurable $h : \mathbb{R}^d \to \mathbb{R}$, it holds

$$\int \exp(h(x) - \Phi^\varepsilon(h)(x)) d\mu(x) \le 1$$

with equality if h is bounded from above. (Here, the measure μ is the one of Theorem 1)

Proof. (2) Let $h \le k$. Then

$$e^h \le e^k \implies 0 < P^\varepsilon(e^h) \le P^\varepsilon(e^k) \implies 0 < \frac{e^{-W}}{P^\varepsilon(e^k)} \le \frac{e^{-W}}{P^\varepsilon(e^h)} \implies$$

$$\implies 0 < P^\varepsilon\left(\frac{e^{-W}}{P^\varepsilon(e^k)}\right) \le P^\varepsilon\left(\frac{e^{-W}}{P^\varepsilon(e^h)}\right) \implies$$

$$\implies \log\left(P^\varepsilon\left(\frac{e^{-W}}{P^\varepsilon(e^k)}\right)\right) \le \log\left(P^\varepsilon\left(\frac{e^{-W}}{P^\varepsilon(e^h)}\right)\right) \implies$$

$$\implies V - \log\left(P^\varepsilon\left(\frac{e^{-W}}{P^\varepsilon(e^h)}\right)\right) \le V - \log\left(P^\varepsilon\left(\frac{e^{-W}}{P^\varepsilon(e^k)}\right)\right) \implies$$

$$\implies \Phi^\varepsilon(e^h) \leq \Phi^\varepsilon(e^k).$$

(1) Take a function $h \in L^2(\gamma_d)$. Following the notation of Lemma 15, the functions $\tilde{h}(z) = h(e^{-\varepsilon/2}z)$ and $E(z) = e^{-\frac{\|z\|^2 \cdot e^{-\varepsilon}}{2(1-e^{-\varepsilon})}}$ are also in $L^2(\gamma_d)$. Hölder's inequality gives us that the convolution $\tilde{h} * E$ is a bounded function, which means that $|P^\varepsilon h(x)| = C \cdot |(\tilde{h} * E)(x)| \leq M$, for some constant M. Hence

$$\int_{\mathbb{R}^d} |P^\varepsilon h(x)|^2 d\gamma_d(x) \leq M^2 \gamma_d(\mathbb{R}^d) = M^2 < +\infty$$

so $P^\varepsilon : L^2(\gamma_d) \to L^2(\gamma_d)$.
For the symmetry property, we must check that

$$\langle P^\varepsilon \psi, g \rangle_{L^2(\gamma_d)} = \langle \psi, P^\varepsilon g \rangle_{L^2(\gamma_d)} \ .$$

Following the arguments of Lemma 15, denote by $K = \frac{1}{(2\pi)^{d/2}} \frac{1}{(1-e^{-\varepsilon})^{d/2}}$.

$$\langle P^\varepsilon \psi, g \rangle = K \int_{\mathbb{R}^d} g(x) \int_{\mathbb{R}^d} \tilde{\psi}(u) \exp\left\{ -\frac{\|x-u\|^2 e^{-\varepsilon}}{2(1-e^{-\varepsilon})} \right\} (-1)^d e^{-\frac{d\varepsilon}{2}} du \ d\gamma_d(x)$$

$$= K \int_{\mathbb{R}^d} \tilde{\psi}(u) \int_{\mathbb{R}^d} g(x) \exp\left\{ -\frac{\|x-u\|^2 e^{-\varepsilon}}{2(1-e^{-\varepsilon})} \right\} (-1)^d e^{-\frac{d\varepsilon}{2}} d\gamma_d(x) \ du$$

$$= K \int_{\mathbb{R}^d} \psi(e^{-\varepsilon/2}u) \int_{\mathbb{R}^d} g(x) \exp\left\{ -\frac{\|x-u\|^2 e^{-\varepsilon}}{2(1-e^{-\varepsilon})} - \frac{\|x\|^2}{2} \right\} (-1)^d e^{-\frac{d\varepsilon}{2}} dx \ du$$

$$= K \int_{\mathbb{R}^d} \psi(\xi)(-1)^d \int_{\mathbb{R}^d} g(x) \exp\left\{ -\frac{\|x-e^{\varepsilon/2}\xi\|^2 e^{-\varepsilon}}{2(1-e^{-\varepsilon})} - \frac{\|x\|^2}{2} \right\} dx \ d\xi$$

$$= K \int_{\mathbb{R}^d} \psi(\xi)(-1)^d \int_{\mathbb{R}^d} g(x) \exp\left\{ -\frac{\|x-e^{\varepsilon/2}\xi\|^2 e^{-\varepsilon}}{2(1-e^{-\varepsilon})} - \frac{\|x\|^2}{2} + \frac{\|\xi\|^2}{2} \right\} dx d\gamma_d(\xi)$$

Similarly,

$$\langle \psi, P^\varepsilon g \rangle = K \int_{\mathbb{R}^d} \psi(\xi) \int_{\mathbb{R}^d} \tilde{g}(x) \exp\left\{ -\frac{\|\xi-x\|^2 e^{-\varepsilon}}{2(1-e^{-\varepsilon})} \right\} (-1)^d e^{-\frac{d\varepsilon}{2}} dx \ d\gamma_d(\xi)$$

$$= K \int_{\mathbb{R}^d} \psi(\xi) \int_{\mathbb{R}^d} g(e^{-\varepsilon/2}x) \exp\left\{ -\frac{\|\xi-x\|^2 e^{-\varepsilon}}{2(1-e^{-\varepsilon})} \right\} (-1)^d e^{-\frac{d\varepsilon}{2}} dx \ d\gamma_d(\xi)$$

$$= K \int_{\mathbb{R}^d} \psi(\xi)(-1)^d \int_{\mathbb{R}^d} g(y) \exp\left\{ -\frac{\|\xi-e^{\varepsilon/2}y\|^2 e^{-\varepsilon}}{2(1-e^{-\varepsilon})} \right\} dy \ d\gamma_d(\xi)$$

and by noticing that

$$-\frac{\|\xi - e^{\varepsilon/2}x\|^2 e^{-\varepsilon}}{2(1-e^{-\varepsilon})} = -\frac{\|x - e^{\varepsilon/2}\xi\|^2 e^{-\varepsilon}}{2(1-e^{-\varepsilon})} - \frac{\|x\|^2}{2} + \frac{\|\xi\|^2}{2}$$

we get the desired symmetry property.
(3) Fix some positive number α and some measurable $h : \mathbb{R}^d \to \mathbb{R}$. By the proof of property (1), we see that we used Fubini's theorem only when we computed the quantity $\langle P^\varepsilon \psi, g\rangle_{L^2(\gamma_d)}$, which means that for the symmetry property it suffices to have that $\langle P^\varepsilon \psi, g\rangle_{L^2(\gamma_d)} < +\infty$. Since always

$$\frac{P^\varepsilon(e^{h\wedge\alpha})}{P^\varepsilon(e^h)} \leq 1$$

we get that

$$\int\limits_{\mathbb{R}^d} P^\varepsilon(e^{h\wedge\alpha})\frac{e^{-W}}{P^\varepsilon(e^h)} d\gamma_d \leq 1$$

so, by applying the symmetry property we get:

$$\int\limits_{\mathbb{R}^d} P^\varepsilon(e^{h\wedge\alpha})\frac{e^{-W}}{P^\varepsilon(e^h)} d\gamma_d = \int\limits_{\mathbb{R}^d} e^{h\wedge\alpha}\left(P^\varepsilon \frac{e^{-W}}{P^\varepsilon(e^h)}\right) d\gamma_d = \int\limits_{\mathbb{R}^d} e^{h\wedge\alpha - \Phi^\varepsilon(h)} d\mu \ .$$

The last quantity goes to $\int e^{h-\Phi^\varepsilon(h)} d\mu$, as $\alpha \to +\infty$, so using the above inequality we get that $\int e^{h-\Phi^\varepsilon(h)} d\mu \leq 1$. In the case where h is a bounded function, we get that

$$P^\varepsilon(e^{h\wedge\alpha})(x) \xrightarrow{\alpha\to+\infty} P^\varepsilon(e^h)(x) \ , \quad \forall x \ \text{ (by dominated convergence)}$$

and

$$\lim_{\alpha\to+\infty} \int\limits_{\mathbb{R}^d} P^\varepsilon(e^{h\wedge\alpha})\frac{e^{-W}}{P^\varepsilon(e^h)} d\gamma_d = \int\limits_{\mathbb{R}^d} e^{-W} d\gamma_d = 1$$

as desired.

□

The existence of a coupling of the desired form can be established under more general conditions on μ and ν:

Theorem 18. *Let μ a probability measure of the form $\mu(dx) = e^{V(x)}\gamma_d(dx)$, with $V : \mathbb{R}^d \to \mathbb{R}$ convex and bounded from below, and let ν a probability measure of the form $\nu(dx) = e^{-W(x)}\gamma_d(dx)$, with $W : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ convex such that the set $\{W < -m\}$ is bounded, for $m = \inf_x\{V(x)\} \leq 0$. Then there exist a log-convex function $f^\varepsilon : \mathbb{R}^d \to [1, +\infty)$ and a log-concave function $g^\varepsilon : \mathbb{R}^d \to (0, +\infty)$ such that the measure π^ε defined by $\pi^\varepsilon(dxdy) = f^\varepsilon(x)g^\varepsilon(y)R^\varepsilon(dxdy)$ belongs to $C(\mu, \nu)$.*

Proof of Theorem 18. Let us show that there exists a convex function $\bar{h} : \mathbb{R}^d \to \mathbb{R}^+$ such that $\Phi^\varepsilon(\bar{h}) = \bar{h}$. Then, by defining $f^\varepsilon = e^{\bar{h}}$ and $g^\varepsilon = e^{-W}/P^\varepsilon(f^\varepsilon)$, we see that f^ε is log-convex, g^ε is log-concave (use the fact that P^ε preserves log-convexity) and satisfy relations (4). We define inductively the sequence $\{h_n\}_{n\geq 0}$ as follows: $h_0 = 0$ and if $n \geq 0$

$$h_{n+1} = [\Phi^\varepsilon(h_n)]^+ \wedge n \ . \tag{6}$$

It is clear that $h_n \in [0, n-1]$. We show inductively that h_n is a non-decreasing sequence. In fact, $h_1 = 0 = h_0$, so in particular $h_0 \leq h_1$. If $h_n \leq h_{n+1}$ then using that Φ^ε is monotone (Lemma 17, property (2)) we get that:

$$h_{n+2} = [\Phi^\varepsilon(h_{n+1})]^+ \wedge (n+1) \geq [\Phi^\varepsilon(h_n)]^+ \wedge (n+1) \geq [\Phi^\varepsilon(h_n)]^+ \wedge (n) = h_{n+1}$$

We denote by h_∞ the pointwise limit of h_n as $n \to +\infty$, which takes values in $\mathbb{R}^+ \cup \{+\infty\}$. We will prove that h_∞ satisfies the following fixed point equation:

$$h_\infty = [\Phi^\varepsilon(h_\infty)]^+ \tag{7}$$

In fact, by monotone convergence we see that $P^\varepsilon(e^{h_n})(x) \to P^\varepsilon(e^{h_\infty})(x)$ We want to apply dominated convergence to show that

$$P^\varepsilon\left(e^{-W}\frac{1}{P^\varepsilon(e^{h_n})}\right) \xrightarrow{n\to\infty} P^\varepsilon\left(e^{-W}\frac{1}{P^\varepsilon(e^{h_\infty})}\right) \ . \tag{8}$$

For fixed x, we write

$$P^\varepsilon\left(e^{-W}\frac{1}{P^\varepsilon(e^{h_n})}\right)(x) = \frac{1}{C}\int_{\mathbb{R}^d} \frac{e^{-W(y+xe^{-\varepsilon/2})}}{P^\varepsilon(e^{h_n})(y+xe^{-\varepsilon/2})} e^{-\frac{\|y\|^2}{2(1-e^{-\varepsilon})}} dy$$

and let $f_n^{(x)}(y)$ be the function inside the integral. Since the sequence $\{h_n\}$ is non-decreasing and bounded from below by 0, it follows that for all x:

$$\cdots \frac{1}{P^\varepsilon(e^{h_n})(x)} \leq \cdots \leq \frac{1}{P^\varepsilon(1)} = \frac{1}{C\int e^{-\frac{\|y\|^2}{2(1-e^{-\varepsilon})}} dy} = K < \infty$$

Now, suppose that $y \in \mathbb{R}^d$ is such that $y + xe^{-\varepsilon/2} \in \{z \in \mathbb{R}^d : W(z) < -m\}$ By assumption, the last set is bounded, so we can find some radius r_1 such that $y + xe^{-\varepsilon/2} \in B(0, r_1)$. Since W is lower-semicontinuous (by convention, all convex functions we are dealing with are lower-semicontinuous), it attains a minimum on the set $\overline{B(0, r_1)}$, hence

$$-W(y + xe^{-\varepsilon/2}) \leq M, \quad \text{for } -M = \min\{W(z) : z \in B(0, r_1)\}$$

and we deduce that
$$e^{-W(y+xe^{-\varepsilon/2})} \leq e^M$$
in that case.
If y is such that $y + xe^{-\varepsilon/2} \notin \{z \in \mathbb{R}^d : W(z) < -m\}$, then trivially we get that
$$e^{-W(y+xe^{-\varepsilon/2})} \leq e^m \ .$$
Finally, by taking $\Lambda = \max\{e^M, e^m\}$, we get that
$$e^{-W(y+xe^{-\varepsilon/2})} \leq \Lambda, \quad \forall y \in \mathbb{R}^d$$
and
$$|f_n^{(x)}(y)| \leq \Lambda \cdot K \cdot e^{-\frac{\|y\|^2}{2(1-e^{-\varepsilon})}} \in L^1(\mathbb{R}^d)$$
so the dominated convergence theorem can be applied.
The relation (8) gives us that
$$[\Phi^\varepsilon(h_n)(x)]^+ \xrightarrow{n\to\infty} [\Phi^\varepsilon(h_\infty)(x)]^+$$
hence
$$[\Phi^\varepsilon(h_n)(x)]^+ \wedge n \xrightarrow{n\to\infty} [\Phi^\varepsilon(h_\infty)(x)]^+ \ .$$
But $[\Phi^\varepsilon(h_n)(x)]^+ \wedge n = h_{n+1}(x) \to h_\infty(x)$, so we deduce (7).
Now we will prove that h_∞ is in fact a fixed point of Φ^ε. This will follow if we prove that $\Phi^\varepsilon(h_\infty)(x) < +\infty$ for all x. Indeed, let's consider this statement as true for now. Because of (7), we get also that $h_\infty(x) < +\infty$ for all x, so we can apply Lemma 17-property (3) to see that
$$\int e^{h_\infty - \Phi^\varepsilon(h_\infty)} d\mu \leq 1$$
Again by (7) it holds that $h_\infty \geq \Phi^\varepsilon(h_\infty)$, so easily we get that $h_\infty = \Phi^\varepsilon(h_\infty)$ μ-almost everywhere, so this holds also λ^d-almost everywhere (because μ and λ^d are mutually absolutely continuous). Using the same arguments as before, one can also find that $\Phi^\varepsilon(h_\infty)$ is continuous, so, by (7), also h_∞ is continuous. This means that the equality $h_\infty = \Phi^\varepsilon(h_\infty)$ holds in fact everywhere. To finish the proof that h_∞ is a fixed point of Φ^ε, it remains to prove that $\Phi^\varepsilon(h_\infty)(x) < +\infty$ for all x. By contradiction, assume there is a point $x_0 \in \mathbb{R}^d$ such that $\Phi^\varepsilon(h_\infty)(x_0) = +\infty$. Using the arguments above, we see that
$$e^{-W(y+x_0e^{-\varepsilon/2})} \leq \Lambda$$
so if $\Phi^\varepsilon(h_\infty)(x_0) = +\infty$, then necessarily $P^\varepsilon(e^{h_\infty}) = +\infty$ almost everywhere, which in turn implies that $\Phi^\varepsilon(h_\infty) \equiv +\infty$. Because

$h_\infty \geq \Phi^\varepsilon(h_\infty)$, we conclude also that $h_\infty \equiv +\infty$. Now let us show that there exists n_0 such that for all $n \geq n_0$

$$\inf_{x\in\mathbb{R}^d} \Phi^\varepsilon(h_n)(x) \geq 0 \ . \tag{9}$$

For any $x \in \mathbb{R}^d$, we write (denote by $C = (2\pi)^{d/2}(1-e^{-\varepsilon})^{d/2}$ and $m = \inf_x V(x)$):

$$P^\varepsilon\left(e^{-W}\frac{1}{P^\varepsilon(e^{h_n})}\right)(x) = \frac{1}{C}\int\limits_{\mathbb{R}^d} e^{-W(y)}\frac{1}{P^\varepsilon(e^{h_n})}(y)e^{-\frac{\|y-xe^{-\varepsilon/2}\|^2}{2(1-e^{-\varepsilon})}}dy$$

$$= \frac{1}{C}\int\limits_{\{W<-m\}} e^{-W(y)}\frac{1}{P^\varepsilon(e^{h_n})}(y)e^{-\frac{\|y-xe^{-\varepsilon/2}\|^2}{2(1-e^{-\varepsilon})}}dy$$

$$+\frac{1}{C}\int\limits_{\{W\geq-m\}} e^{-W(y)}\frac{1}{P^\varepsilon(e^{h_n})}(y)e^{-\frac{\|y-xe^{-\varepsilon/2}\|^2}{2(1-e^{-\varepsilon})}}dy$$

$$\leq \frac{1}{C}\int\limits_{\{W<-m\}} e^{-\frac{\|y-xe^{-\varepsilon/2}\|^2}{2(1-e^{-\varepsilon})}}dy \sup_{z\in\{W\leq -m\}} e^{-W(z)}\frac{1}{P^\varepsilon(e^{h_n})}(z)$$

$$+\frac{e^m}{C}\int\limits_{\{W\geq-m\}} e^{-\frac{\|y-xe^{-\varepsilon/2}\|^2}{2(1-e^{-\varepsilon})}}dy$$

where we used the fact that $P^\varepsilon(e^{h_n}) \geq 1$, since $h_n \geq 0$. The sequence of functions is a non-increasing of continuous functions converging pointwise to 0 (for the continuity, write

$$P^\varepsilon(e^{h_n})(x) = \frac{1}{C}\int (e^{h_n})(y)e^{-\frac{\|y-xe^{-\varepsilon/2}\|^2}{2(1-e^{-\varepsilon})}}dy$$

and use that $e^{h_n} \leq e^{n-1}$ together with dominated convergence). According to Dini's theorem, the convergence is uniform on the compact set $K := \overline{\{W \leq -m\}}$. Since W is by convention lower-semicontinuous, it is bounded from below on K, therefore, there exists some n_0 such that $\sup_{z\in K} e^{-W(z)}\frac{1}{P^\varepsilon(e^{h_n})}(z) \leq e^m$, for all $n \geq n_0$. Plugging this into the above inequality, we get that

$$P^\varepsilon\left(e^{-W}\frac{1}{P^\varepsilon(e^{h_n})}\right)(x) \leq e^m \ .$$

Hence, for all $n \geq n_0$

$$\inf_{x\in\mathbb{R}^d} \Phi^\varepsilon(h_n)(x) = \inf_{x\in\mathbb{R}^d}\left\{V(x) - \log\left(P^\varepsilon\left(e^{-W}\frac{1}{P^\varepsilon(e^{h_n})}\right)(x)\right)\right\} \geq 0$$

which is inequality (9).
Choose n_0 such that $\Phi^\varepsilon(h_{n_0}) \geq 0$. Then, $h_{n_0+1} \leq \Phi^\varepsilon(h_{n_0}) \leq \Phi^\varepsilon(h_{n_0+1})$ and since h_{n_0+1} is bounded, again Lemma 17-property (3) gives us that $\int e^{h_{n_0+1}-\Phi^\varepsilon(h_{n_0+1})} d\mu = 1$, implying that $h_{n_0+1} = \Phi^\varepsilon(h_{n_0+1})$ everywhere (again we used continuity of $\Phi^\varepsilon(h_{n_0+1})$). Now similarly we see that

$$h_{n_0+2} \overset{(9)}{=} \Phi^\varepsilon(h_{n_0+1}) \wedge (n_0+1) = h_{n_0+1} \wedge (n_0+1) = h_{n_0+1}$$

therefore, inductively, $h_n \equiv h_{n_0+1}$ for all $n \geq n_0+1$. This implies that $h_\infty \equiv h_{n_0+1} \in [0, n_0]$, which is a contradiction. Thus $\Phi^\varepsilon(h_\infty)(x) < +\infty$ everywhere and the claim is proved.
The function h_∞ which satisfies the fixed point equation is in general not convex. In order to produce a convex function, let $k_0 = h_\infty^{**}$ be the Fenchel transform of h_∞, that is $k_0(x) = \sup_{y\in\mathbb{R}^d}\{\langle x, y\rangle - h_\infty^*(y)\}$ and $h_\infty^*(z) = \sup_{t\in\mathbb{R}^d}\{\langle t, z\rangle - h_\infty(t)\}$. By definition, $k_0 \leq h_\infty$ and since $h_\infty \geq 0$ and $h_\infty^{**} = \sup\{f_\alpha : f_\alpha \text{ is affine and } f_\alpha \leq h_\infty\}$ (see [35], Theorem 12.2 and Corollary 12.1.1), we see that $k_0 \geq 0$. Now define inductively $\{k_n\}_{n\geq 1}$ by $k_{n+1} = \max\{\Phi^\varepsilon(k_n), k_0\}$. According to Lemma 15, the operator Φ^ε preserves convexity, hence k_n is convex for all n. The sequence k_n is non-decreasing, and since Φ^ε is monotone, it holds that $k_n \leq h_\infty$ for all n. This means that k_n converges pointwise to a limit k_∞, which, as a pointwise limit of convex functions is also convex and since $k_n \leq h_\infty < \infty$, it is also finite valued. Reasoning as above, we see that $k_\infty = \max\{\Phi^\varepsilon(k_\infty), k_0\}$ and particularly, $k_\infty \geq \Phi^\varepsilon(k_\infty)$. Using again Lemma 17-property (3), necessarily it must hold that $k_\infty = \Phi^\varepsilon(k_\infty)$, i.e. k_∞ is a fixed point of Φ^ε.

□

Proof of Theorem 14. We wish to apply Theorem 18, and so we must check that the set $\{W < -m\}$ is bounded, where $m = \inf_x\{V(x)\}$. Indeed, since $\frac{d\nu}{d\gamma_d} = e^{-W}$ we have the following inclusions

$$\{W < -m\} \subseteq \{W = +\infty\}^c \subseteq \operatorname{supp}(\nu)$$

and the latter set is bounded, since it is a compact subset of $\mathbb{R}^d$. Now, according to Theorem 18, there exists a coupling

$$\pi^\varepsilon(dxdy) = f^\varepsilon(x) g^\varepsilon(y) R^\varepsilon(dxdy) \in C(\mu, \nu)$$

such that f^ε is log-convex and g^ε is log-concave. It remains to prove the optimality of this coupling for $\mathcal{T}_H^\varepsilon(\mu, \nu)$. We have that $P^\varepsilon(f^\varepsilon)(y) g^\varepsilon(y) = e^{-W(y)}$, so $\log g^\varepsilon(y) = -W(y) - \log P^\varepsilon(f^\varepsilon)(y)$. Since P^ε preserves log-convexity, the function $\log P^\varepsilon(f^\varepsilon)$ is convex, hence continuous and bounded on the compact set $\operatorname{supp}(\nu)$. Also, since $\{W < +\infty\} \subseteq \operatorname{supp}(\nu)$, we get that the convex function W is real valued inside the support of ν, hence also bounded there, which gives the integrability of $\log g^\varepsilon$ with respect to ν. Also, since $\log f^\varepsilon \geq 0$, the integral $\int \log f^\varepsilon d\mu$ is well defined in $[0, +\infty]$. Choose a coupling $\pi \in C(\mu, \nu)$ such that $H(\pi|R^\varepsilon) < +\infty$ (for example choose $\pi = \mu \otimes \nu$). Using inequality (3) with $\alpha = \pi, \beta = R^\varepsilon, h(x, y) = \log f^\varepsilon(x) + \log g^\varepsilon(y)$ we get:

$$+\infty > H(\pi|R^\varepsilon) \geq \int_{\mathbb{R}^d \times \mathbb{R}^d} \log f^\varepsilon(x) + \log g^\varepsilon(y)\pi(dxdy) =$$

$$= \int_{\mathbb{R}^d} \log f^\varepsilon(x) d\mu(x) + \int_{\mathbb{R}^d} \log g^\varepsilon(y) d\nu(y)$$

which proves also the integrability of $\log f^\varepsilon$ with respect to μ. We compute the entropy $H(\pi^\varepsilon|R^\varepsilon)$:

$$H(\pi^\varepsilon|R^\varepsilon) = \int \log\left(\frac{d\pi^\varepsilon}{dR^\varepsilon}\right) d\pi^\varepsilon = \int \log f^\varepsilon(x) d\mu(x) + \int \log g^\varepsilon(y) d\nu(y)$$

which proves the optimality of π^ε.

$\square$

We will now state and prove two technical Lemmas that were used in the proof of Theorem 1 (see page 11).

Proof of Lemma 5. Since $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, it suffices to show that $H(\mu|\gamma_d) < +\infty$, which is the same as proving that V is μ-integrable. Now, V is a convex function, hence it is bounded from below by some affine function, namely there exist $a, b \in \mathbb{R}$ and $w \in \mathbb{R}^d$ such that

$$V(x) \geq a \cdot \langle x, w \rangle + b, \ \forall x \in \mathbb{R}^d \ .$$

This yields that the function $[V]^- = \max\{-V, 0\}$ is μ-integrable. In fact:

$$\int [V]^- d\mu = \int_{\{V \geq 0\}} [V]^- d\mu + \int_{\{V<0\}} [V]^- d\mu = 0 + \int_{\{V<0\}} -V d\mu \leq$$

$$\leq \int_{\{V<0\}} |a| \cdot \langle x, w \rangle + |b| \ d\mu(x) \leq |a| \cdot \|w\| \int_{\{V<0\}} (\|x\|^2 + 1) d\mu(x) + |b|$$

and the last expression is finite since $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. Furthermore, since the convex function V is such that $\int e^{V(x)} d\gamma_d(dx) = 1$, we can use Lemma 2.1 of [19] to deduce that $[V]^+(x) \leq \frac{\|x\|^2}{2}$, for all x, hence $[V]^+$ is also μ-integrable. Similarly, to check that ν has finite Shannon entropy, we must check that W is ν-integrable. Reasoning as above, $[W]^-$ is

ν-integrable because W is bounded from below by some affine function. Also,

$$\int [W]^+ d\nu = \int \max\{W, 0\} \cdot e^{-W} d\gamma_d = \int \max\{-(-W) \cdot e^{-W}, 0\} d\gamma_d =$$

$$= \int [\log(e^{-W})e^{-W}]^- d\gamma_d \leq \frac{1}{e}$$

where the last inequality came from the fact that $g(s) = s\log(s) \geq -\frac{1}{e}$, for all s.

□

The next Lemma was also used to prove Theorem 1:

Lemma 19. *Let $\nu(dx) = e^{-W(x)}\gamma_d(dx)$ with $W : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ convex and $\eta \leq_c \nu$. Suppose also that ν has compact support. Define, for all $\theta \in (0, \pi/2)$,*

$$\nu_\theta = \mathrm{Law}((\cos\theta)X + (\sin\theta)Z) \quad \textit{and} \quad \eta_\theta = \mathrm{Law}((\cos\theta)Y + (\sin\theta)Z)$$

where $X \sim \nu$, $Y \sim \eta$ and Z is independent of X, Y and the law α of Z is given by $\alpha(dz) = \frac{1}{C}\mathbb{I}_B\gamma_d(dz)$, where B is the Euclidean unit ball and C an appropriate normalizing constant. Then, for all $\theta \in (0, \pi/2)$:

(1) the probability ν_θ has density of the form e^{-W_θ} with respect to γ_d, with $W_\theta : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ convex

(2) the probability measures ν_θ and η_θ are compactly supported

(3) it holds $\eta_\theta \leq_c \nu_\theta$

(4) η_θ has finite Shannon entropy

(5) $W_2(\mu, \nu_\theta) \xrightarrow{\theta \to 0} W_2(\mu, \nu_0)$ and $W_2(\mu, \eta_\theta) \xrightarrow{\theta \to 0} W_2(\mu, \eta_0)$

Proof. (1) We claim that the density f_θ of ν_θ with respect to the Lebesgue measure is given by

$$f_\theta(x) = \frac{1}{C'} \int\limits_B e^{-W(\frac{x-(\sin\theta)y}{\cos\theta})} e^{-\frac{\|x-(\sin\theta)y\|^2}{2\cos^2\theta}} e^{-\frac{\|y\|^2}{2}} dy$$

where $C' = C(\cos\theta)^d(2\pi)^d$. Indeed, for example if we restrict ourselves in the case $d = 1$, we compute the law of $(\cos\theta)X + (\sin\theta)Z$:

$$\mathbb{P}\{\omega : (\cos\theta)X(\omega) + (\sin\theta)Z(\omega) \leq a\} =$$

$$= \int\limits_{-1}^{1} \int\limits_{-\infty}^{\frac{a-(\sin\theta)t}{\cos\theta}} \frac{1}{C}\frac{1}{2\pi} e^{-W(s)} e^{-\frac{s^2}{2}} e^{-\frac{t^2}{2}} d\lambda(s) d\lambda(t) =$$

$$= \int_{-1}^{1} \int_{-\infty}^{a} \frac{1}{C \cdot \cos\theta} \frac{1}{2\pi} e^{-W(\frac{u-(\sin\theta)t}{\cos\theta})} e^{-\frac{(u-(\sin\theta)t)^2}{2\cos^2\theta}} e^{-\frac{t^2}{2}} d\lambda(u) d\lambda(t) =$$

$$= \int_{-\infty}^{a} \frac{1}{C \cdot \cos\theta} \frac{1}{2\pi} \int_{-1}^{1} e^{-W(\frac{u-(\sin\theta)t}{\cos\theta})} e^{-\frac{(u-(\sin\theta)t)^2}{2\cos^2\theta}} e^{-\frac{t^2}{2}} d\lambda(t) d\lambda(u)$$

A simple calculation now shows that

$$e^{\frac{\|x\|^2}{2}} \cdot f_\theta(x) = \frac{1}{C'} \int_B e^{-W(\frac{x-(\sin\theta)y}{\cos\theta})} e^{-\frac{\|(\sin\theta)x-y\|^2}{2\cos^2\theta}} dy$$

The function inside the integral is log-concave (as a function of $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$) so according to Prekopa's Theorem (see [34], Theorem 6), since B is a convex set, we deduce that also the right hand side is log-concave. Now choose $\log(e^{\frac{\|x\|^2}{2}} \cdot f_\theta(x)) = -W_\theta$ and the item (1) is proved.

(2) We have that

$$\mathrm{supp}[\mathrm{Law}((\cos\theta)X)] = \cos\theta \cdot \mathrm{supp}[\nu] \subseteq \bar{B}(0, r)$$

for some appropriate radius $r > 0$, because $\mathrm{supp}[\nu]$ is compact. Similarly

$$\mathrm{supp}[\mathrm{Law}((\sin\theta)Z)] = \sin\theta \cdot \mathrm{supp}[\alpha] \subseteq \bar{B}(0, \sin\theta)$$

Since the law ν_θ is just the convolution of the laws of $(\cos\theta)X$ and $(\sin\theta)Z$, it holds that this law is supported on the Minkowski sum of the the above two supports, which means that

$$\mathrm{supp}(\nu_\theta) \subseteq \bar{B}(0, r + \sin\theta)$$

so ν_θ is compactly supported. A similar argument shows that η_θ is compactly supported, but first we must check that the support of η is also compact. Take the function

$$I(x) = \begin{cases} 0, & x \in \bar{B}(0, r) \\ +\infty, & x \notin \bar{B}(0, r) \end{cases} .$$

This is a clearly a convex and nonnegative function, so since $\eta \leq_c \nu$ we get that

$$0 \leq \int I(x) d\eta(x) \leq \int I(x) d\nu(x) = 0$$

From this we deduce the implication

$$x \notin \bar{B}(0, r) \implies x \notin \mathrm{supp}(\eta)$$

or equivalently

$$\text{supp}(\eta) \subseteq \bar{B}(0, r) \ .$$

As before we prove that the support of η_θ is compact.
(3) It suffices to prove that if A, B, C are random variables with C being independent from A and B and Law(A)$\leq_c$Law(B), then Law(A+C)$\leq_c$Law(B+C). Take an arbitrary convex function $f : \mathbb{R}^d \to \mathbb{R}$. Then:

$$\int\limits_{\mathbb{R}^d} f(x) d(\alpha * c)(x) = \int\limits_{\mathbb{R}^d} \int\limits_{\mathbb{R}^d} f(x+y) d\alpha(x) dc(y) \leq$$

$$\leq \int\limits_{\mathbb{R}^d} \int\limits_{\mathbb{R}^d} f(x+y) d\beta(x) dc(y) = \int\limits_{\mathbb{R}^d} f(x) d(\beta * c)(x)$$

which proves the claim. We used above that the function $x \mapsto f(x+y)$ is convex for every fixed y.
(4) As before, we can compute the distribution of $((\cos\theta)Y + (\sin\theta)Z)$, η_θ, and see that its density with respect to the Lebesgue measure equals:

$$g_\theta(x) = \frac{1}{K} \int\limits_{\frac{x-\sin\theta}{\cos\theta}}^{\frac{x+\sin\theta}{\cos\theta}} e^{-\frac{\|(\cos\theta)y - x\|^2}{2\sin^2\theta}} d\eta(y)$$

for some properly chosen constant K. This means that $g_\theta \leq \frac{1}{K}$. Moreover, since $s \cdot \log(s) \geq -\frac{1}{e}$ for every s, we deduce that $g_\theta \log(g_\theta) \geq -\frac{1}{e}$. Item (2) gives that $\text{supp}(\eta_\theta)$ is compact, hence:

$$H(\eta_\theta | \lambda^d) = \int\limits_{\text{supp}(\eta_\theta)} g_\theta \log(g_\theta) d\lambda^d < +\infty \ .$$

(5) We will prove that $W_2(\mu, \nu_\theta) \xrightarrow{\theta \to 0} W_2(\mu, \nu)$ by showing that $\nu_\theta \xrightarrow[\theta \to 0]{w} \nu$ and $\int \|y\|^2 d\nu_\theta(y) \xrightarrow{\theta \to 0} \int \|y\|^2 d\nu(y)$. To show the first convergence, we can use Scheffé lemma and check that we have almost everywhere pointwise convergence of the respective densities, i.e. $f_{\nu_\theta}(x) \xrightarrow{\theta \to 0} f_\nu(x)$ for λ^d-almost all x. Let us fix a point $x \in \mathbb{R}^d$ which does not belong to the boundary of the domain of W, $\partial(\text{Dom}(W))$, where $\text{Dom}(W) = \{x : W(x) < +\infty\}$. The set $\partial(\text{Dom}(W))$ is a set of zero Lebesgue measure (it is the boundary of a convex set). By the proof of item (1), the density f_{ν_θ} equals:

$$f_{\nu_\theta}(x) = \frac{1}{C'} \int\limits_B e^{-W(\frac{x-(\sin\theta)y}{\cos\theta})} e^{-\frac{\|x-(\sin\theta)y\|^2}{2\cos^2\theta}} e^{-\frac{\|y\|^2}{2}} dy$$

Since $y \in B$, we see that $\frac{x-(\sin\theta)y}{\cos\theta} \in \frac{x-(\sin\theta)\cdot B}{\cos\theta}$. For small enough θ (such that for example $\cos\theta > 1/2$), we can find a radius R such that the ball

$\bar{B}(2x, R)$ contains all the points $\frac{x-(\sin\theta)\cdot B}{\cos\theta}$. Since W is lower semicontinuous, it attains a minimum l on the ball $\bar{B}(2x, R)$, which means that

$$\sup_{\theta\in(0,\pi/4)}\sup_{y\in B}\left\{e^{-W(\frac{x-(\sin\theta)y}{\cos\theta})}\right\}\leq e^{-l}\ .$$

Similarly, the function

$$e^{-\frac{\|x-(\sin\theta)y\|^2}{2\cos^2\theta}}e^{-\frac{\|y\|^2}{2}}$$

is also bounded uniformly for $y \in B$, and all these facts yield that:

$$\sup_{\theta\in(0,\pi/4)}\sup_{y\in B}\left\{e^{-W(\frac{x-(\sin\theta)y}{\cos\theta})}e^{-\frac{\|x-(\sin\theta)y\|^2}{2\cos^2\theta}}e^{-\frac{\|y\|^2}{2}}\right\}\leq M$$

for some constant M. Moreover we claim that the following convergence holds for all $y \in B$:

$$\frac{1}{C'}e^{-W(\frac{x-(\sin\theta)y}{\cos\theta})}e^{-\frac{\|x-(\sin\theta)y\|^2}{2\cos^2\theta}}e^{-\frac{\|y\|^2}{2}}\xrightarrow{\theta\to 0^+}\frac{1}{C(2\pi)^d}e^{-W(x)}e^{-\frac{\|x\|^2+\|y\|^2}{2}}$$

where

$$C=\frac{1}{(2\pi)^{d/2}}\int\limits_B e^{-\frac{\|z\|^2}{2}}dz$$

In fact, the only thing we must check is that the amount $W(\frac{x-(\sin\theta)y}{\cos\theta})$ converges to $W(x)$, as $\theta\to 0$, (remember, we assumed that $x\notin\partial(\mathrm{Dom}(W))$. If our point x does not belong to $\mathrm{Dom}(W)$, then $W(x)=+\infty$, hence by the lower semicontinuity we get immediately that

$$\liminf_{\theta} W(\frac{x-(\sin\theta)y}{\cos\theta})\geq W(x)=+\infty$$

If the point x belongs to $\mathrm{Dom}(W)$, then we write $\mathrm{Dom}(W)=\partial(\mathrm{Dom}(W))\cup\mathrm{Dom}(W)^\circ$. Since the points that belong to the interior of $\mathrm{Dom}(W)$ are always continuity points of W, again the convergence holds true if $x\in\mathrm{Dom}^\circ$. The set $\partial(\mathrm{Dom}(W))$, as we said before, is a set of zero Lebesgue measure, hence the above convergence holds for λ^d-almost every $x\in\mathbb{R}^d$.
Now the dominated convergence theorem yields that

$$\lim_{\theta\to 0^+}\int\limits_{y\in B}\frac{1}{C'}e^{-W(\frac{x-(\sin\theta)y}{\cos\theta})}e^{-\frac{\|x-(\sin\theta)y\|^2}{2\cos^2\theta}}e^{-\frac{\|y\|^2}{2}}dy=\frac{1}{(2\pi)^{d/2}}e^{-W(x)}e^{-\frac{\|x\|^2}{2}}$$

and the right hand side is of course the density of ν with respect to the Lebesgue measure. Hence, $f_{\nu_\theta}(x)\xrightarrow{\theta\to 0}f_\nu(x)$ for λ^d-almost all x.
To prove the other claim, we note that

$$\|y\|^2 f_{\nu_\theta}(y)\xrightarrow{\theta\to 0}\|y\|^2 f_\nu(y)$$

for almost every y. Also, the integrals $\int \|y\|^2 f_{\nu_\theta}(y)dy$ can be restricted to the compact supports of ν_θ, which, because of item (2), are all subsets of $\bar{B}(0, 2r+1)$. If we denote by $G_\theta(x, y)$ the function

$$\|y\|^2 \frac{1}{C'} e^{-W(\frac{y-(\sin\theta)z}{\cos\theta})} e^{-\frac{\|y-(\sin\theta)z\|^2}{2\cos^2\theta}} e^{-\frac{\|z\|^2}{2}}$$

then it holds that

$$\sup_{\theta\in(0,\pi/4)} \sup_{(y,z)\in\bar{B}(0,2r+1)\times B} G_\theta(z, y) = A < +\infty$$

and of course

$$\int\limits_{\bar{B}(0,2r+1)} \int\limits_{B} A \; dzdy < +\infty$$

so the dominated convergence theorem yields that

$$\lim_{\theta\to 0} \int\limits_{\bar{B}(0,2r+1)} \int\limits_{B} G_\theta(x, y) \; dxdy =$$

$$= \int\limits_{\bar{B}(0,2r+1)} \int\limits_{B} \|y\|^2 \frac{1}{C(2\pi)^d} e^{-W(y)} e^{-\frac{\|y\|^2}{2}} e^{-\frac{\|x\|^2}{2}} \; dxdy =$$

$$= \int\limits_{\bar{B}(0,2r+1)} \|y\|^2 e^{-W(y)} \frac{1}{(2\pi)^{d/2}} e^{-\frac{\|y\|^2}{2}} dy = \int\limits_{\mathrm{supp}(\nu)} \|y\|^2 f_\nu(y)dy$$

which is the desired result. All the above give us the Wasserstein convergence

$$W_2(\mu, \nu_\theta) \xrightarrow{\theta\to 0} W_2(\mu, \nu)$$

and the same reasoning can be applied to show that

$$W_2(\mu, \eta_\theta) \xrightarrow{\theta\to 0} W_2(\mu, \eta)$$

$\square$

4. REFERENCES

1. Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna and Gabriel Peyré, *Iterative Bregman projections for regularized transportation problems*, SIAM J. Sci. Comput **37** (2015), no. 2, A1111-A1138. MR 3340204

2. Luis A. Caffarelli, *Monotonicity properties of optimal transportation and the FKG and related inequalities*, Comm. Math. Phys. **214** (2000), no. 3, 547-563. MR 1800860

3. Luis A. Caffarelli, *Erratum: "Monotonicity of optimal transportation and the FKG and related inequalities" [Comm. Math. Phys. 214 (2000), no. 3, 547-563; MR1800860 (2002c:60029)]*, Comm. Math. Phys. **225** (2002), no. 2, 449-450. MR 1889232

4. Guillaume Carlier, Vincent Duval, Gabriel Peyré and Bernhard Schmitzer, *Convergence of entropic schemes for optimal transport and gradient flows*, SIAM J. Math. Anal. **49** (2017), no. 2, 1385-1418. MR 3635459

5. Y. Chen, T. Georgiou and M. Pavon, *Entropic and displacement interpolation: a computational approach using the Hilbert metric*, SIAM J. Math. Analysis **76** (2016), no. 6, 2375-2396

6. Maria Colombo, Alessio Figalli and Yash Jhaveri, *Lipschitz changes of variables between perturbations of log-concave measures*, Ann. Sc. Norm. Super. Pisa Cl. Sci. (5) **17** (2017), no. 4, 1491-1519. MR 3752535

7. Dario Cordero-Erausquin, *Some applications of mass transport to Gaussian-type inequalities*, Arch. Ration. Mech. Anal. **161** (2002), no. 3, 257-269. MR 1894593

8. Thomas A. Courtade, Max Fathi and Ashwin Pananjady, *Quantitative stability of the entropy power inequality*, IEEE Trans. Inform. Theory **64** (2018), no. 8, 5691-5703. MR 3832330

9. I. Csiszár,*I-divergence geometry of probability distributions and minimization problems*, Ann. Probability **3** (1975), 146-158. MR 0365798

10. M. Cuturi, *Sinkhorn distances: Lightspeed computation of optimal transport*, in Advances in Neural Information Processing Systems (NIPS) (2013), no. 26, 2292-2300

11. Paul Dupuis, Richard S. Ellis, *A weak convergence approach to the theory of large deviations*, Wiley Series in Probability and Statistics, John Wiley and Sons, Inc., New York, 1997

12. Guido De Philippis and Alessio Figalli, *Rigidity and stability of*

Caffarelli's log-concave perturbation theorem, Nonlinear Anal. **154** (2017), 59-70. MR 3614644

13. Montacer Essid and Michele Pavon, *Traversing the Schrödinger Bridge Strait: Robert Fortet's Marvelous Proof Redux*, J. Optim. Theory Appl. **181** (2019), no. 1, 23-60. MR 3921394

14. Max Fathi, Nathael Gozlan and Maxime Prodhomme, *A proof of the Caffarelli contraction theorem via entropic regularization*, arXiv preprint arXiv:1904.06053v1, 12 April 2019

15. Robert Fortet, *Résolution d'un système d'équations de M. Schrödinger*, J. Math. Pures Appl. **19** (1940), 83-105. MR 0002692

16. Joel Frnaklin and Jens Lorenz, *On the scaling of multidimensional matrices*, Linear Algebra Appl. **114/115** (1989), 717-735. MR 986904

17. A. Genevay, *Entropy-regularized optimal transport for machine learning*, Ph.D. thesis, Université Paris-Dauphine, 2019

18. N. Gigli and L. Tamanini, *Benamou-Brenier and duality formulas for the entropic cost on RCD*(k,n) spaces*, arxiv preprint (2018)

19. N. Gozlan, M. Madiman, C. Roberto and P.M. Samson, *Deviation inequalities for convex functions motivated by the Talagrand conjecture*, Zap. Nauchn. Sem. S.-Petersburg. Otdel. Mat. Inst. Steklov. (POMI) **457** (2017), no. Veroyatnost' i. Statistika. 25, 168-182. MR 3723580

20. Nathael Gozlan and Nicolas Juillet, *On a mixture of Brenier and Strassen theorems*, arXiv preprint arXiv:1808.02681 (2018)

21. A. Guionnet and D. Shlyakhtenko, *Free monotone transport*, Invent. Math. **197** (2014), no. 3, 613-661. MR 3251831

22. Gilles Hargé, *A particular case of correlation inequality for the Gaussian measure*, Ann. Probab. **27** (1999), no. 4, 1939-1951. MR 1742895

23. Gilles Hargé, *Inequalities for the Gaussian measure and an application to Wiener space*, C. R. Acad. Sci. Paris Sér. I Math **333** (2001), no. 8, 791-794. MR 1868955

24. Olav Kallenberg, *Foundations of Modern Probability*, Second edition Probability and its applications series, Springer, January 2002

25. Young-Heon Kim and Emanuel Milman, *A generalization of Caffarelli's contraction theorem via (reverse) heat flow*, Math. Ann. **354** (2012), no. 3, 827-862. MR 2983070

26. Bo'az Klartag, *Poincaré inequalities and moment maps*, Ann. Fac. Sci. Toulouse Math. **22** (2013), no. 1, 1-41

27. A. V. Kolesnikov, *Global Hölder estimates for optimal transportation*, Mat. Zametki **88** (2010), no. 5, 708-728. MR 2868394

28. A. V. Kolesnikov, *On Sobolev regularity of mass transport and transportation inequalities*, Theory probab. Appl. **57** (2013), no. 2, 243-264. MR 3201654

29. Christian Léonard, *From the Schrödinger problem to the Monge-Kantorovich problem*, J. Funct. Anal. **262** (2012), no. 4, 1879-1920. MR 2873864

30. Christian Léonard, *A survey of the Schrödinger problem and some of its connections with optimal transport*, Discrete Contin. Dyn. Syst. **34** (2014), no. 4, 1533-1574. MR 3121631

31. Toshio Mikami, *Monge's problem with a quadratic cost by the zero noise limit of h-path processes*, Probab. Theory Related Fields **129** (2004), no. 2, 245-260. MR 2063377

32. Emanuel Milman, *on the role of convexity in isoperimetry, spectral gap and concentration*, Invent. Math. **177** (2009), no. 1, 1-43

33. Emanuel Milman, *Spectral estimates, contractions and hypercontractivity* J. Spectr. Theory **8** (2018), no. 2, 669-714. MR 3812812

34. András Prékopa, *On logarithmic concave measures and functions*, Acta Sci. Math. (Szeged) **34** (1973), 335-343. MR 0404557

35. R. Tyrrell Rockafellar, *Convex Analysis*, Princeton Landmarks in Mathematics and Physics, Princeton University Press (December 23, 1996)

36. E. Schrödinger, *Sur la théorie relativiste de l' électron et l'interprétation de la mécanique quantique*, Ann. Inst. H. Poincaré **2** (1932), no. 4, 269-310. MR 1508000

37. Cédric Villani, *Optimal transport*, Grundlehren der Mathematischen Wissenschaften, vol. 338, Springer-Verlag, Berlin, 2009, Old and new. MR 2459454