Submitted to Statistical Science

On some connections between Esscher's tilting, saddlepoint approximations, and optimal transportation: a statistical perspective[1]

Davide La Vecchia, Elvezio Ronchetti and Andrej Ilievski

Abstract. We showcase some unexplored connections between saddlepoint approximations, measure transportation, and some key topics in information theory. To bridge these different areas, we review selectively the fundamental results available in the literature and we draw the connections between them. First, for a generic random variable we explain how the Esscher's tilting (which is a result rooted in information theory and that lies at the heart of saddlepoint approximations) is connected to the solution of the dual Kantorovich problem (which lies at the heart of measure transportation theory) via the Legendre transform of the cumulant generating function. Then, we turn to statistics: we illustrate the connections when the random variable we work with is the sample mean or a statistic with known (either exact or approximate) cumulant generating function. The unveiled connections offer the possibility to look at the saddlepoint approximations from different angles, putting under the spotlight the links to e.g. convex analysis (via the notion of duality) or differential geometry (via the notion of geodesic). We feel these possibilities can trigger a knowledge transfer between statistics and other disciplines, like e.g. mathematics and machine learning. A discussion on some topics for future research concludes the paper.

Key words and phrases: Change of variable, Kullback-Leibler divergence, Geodesic, Optimal transportation map, Wasserstein distance.

1. INTRODUCTION

1.1 Theoretical motivation

A typical problem that arises in statistics is the need for approximating the distribution of some random quantities, depending on $n \in \mathbb{N}$ independent and identicallyl distributed (i.i.d.) random variables.

For instance, assume we are given a sample of size n and we have to derive a confidence interval for the parameter in a one-dimensional location model. To achieve this goal, we need to know the exact distribution of an estimator, such as e.g. the sample mean or another M-estimator of the location parameter. This distribution is known exactly only in very special

Davide La Vecchia is Associate Professor, Research Center for Statistics, Geneva School of Economics and Management, University of Geneva, Geneva, Switzerland (e-mail: davide.lavecchia@unige.ch). Elvezio Ronchetti is Professor Emeritus, Research Center for Statistics, Geneva School of Economics and Management, University of Geneva, Geneva, Switzerland (e-mail: elvezio.ronchetti@unige.ch). Andrej Ilisevski is master student in Statistics at ETH Zürich, Zürich, Switzerland (e-mail: andrej.ilievski.98@gmail.com).

[1] AMS classification: 28E99, 41A60, 46N10, 49K99, 62F12

settings, like e.g. the case when the estimator is a linear function of underlying Gaussian observations. More often, only approximations to the exact distribution are available.

A common approach to tackle this inferential issue is to define a first-order approximation of the distribution, based on a linearization of the estimator. Then, the behaviour of the linearized statistic is studied, as the sample size diverges to infinity. This leads, through the central limit theorem, to many asymptotic normality proofs, and the resulting first-order asymptotic distribution can be used as an approximation to the exact distribution of the estimator; see e.g. Serfling (2009).

Unfortunately, the accuracy of the first-order (Gaussian) asymptotic approximation deteriorates quickly in small samples. Moreover, it tends to be inaccurate in the tails of the distribution even in large samples. To cope with this low accuracy, several techniques have been developed and are available to achieve higher-order accuracy. These techniques include Edgeworth expansions, saddlepoint approximations, the jackknife, and the bootstrap, and provide various corrections to the first-order approximation. For a general survey, we refer to Barndorff-Nielsen and Cox (1989), Young (2009), and Small (2010).

In this paper, we focus on saddlepoint approximations, introduced in the seminal paper of Daniels (1954). Book-length presentations can be found in Field and Ronchetti (1990), Jensen (1995), Kolassa (2006), and Brazzale et al. (2007). These approximations are derived by deforming the contour of the integral defining the inverse Fourier transform of the density of the random variable (i.e. the estimator) of interest. Such a derivation hinges on *complex analysis* results; see e.g. Field and Ronchetti (1990), Ch. 3, or Reid (1988) for a review. We believe that this technical derivation buries some interesting theoretical aspects which link the saddlepoint theory to some recent advances in mathematics. It is our aim to unveil these connections, bridging the statistical theory of saddlepoint approximations to the mathematical theory of optimal measure transportation, as defined in Monge (1781) and Kantorovich (1942).

Considering the practical problem of finding the optimal way to move given piles of sand to fill up given holes of the same total volume, Gaspard Monge (1746-1818), one of the pioneers in the area of optimal transportation, initiated a profound theory anticipating several mathematical areas, including e.g. differential geometry, linear programming, and nonlinear partial differential equations. Monge's problem remained open until the 1940s, when it was revisited by Leonid Vitaliyevitch Kantorovich (1912-1986; Nobel Prize in Economics in 1975) in relation to the economic problem of optimal allocation of resources. We refer to Villani (2009) and to Santambrogio (2015) for a book-length introduction to the historical background, for the mathematical details and for some mathematical applications of measure transportation theory. We refer also to Galichon (2016), for applications in economics (e.g. the principal-agent models or the assignment problem of firms to managers), and to Galichon (2017) for a survey of some recent applications of optimal transport methods in econometrics.

As far as statistics is concerned, the use of measure transportation techniques is becoming more and more popular; see e.g. Chernozhukov et al. (2017) for recent research papers, and Panaretos and Zemel (2019) for a review or Panaretos and Zemel (2020) for a book-length presentation, with focus on the theory of statistics in Wasserstein spaces and on Fréchet means. For instance, measure transportation theory is related to the so-called Wasserstein distance, which is applied in probability and statistics to derive weak convergence and convergence of moments, and which can be easily bounded to derive concentration inequalities. Recently, Hallin (2017), Hallin et al. (2020b) and Hallin et al. (2020c) propose the application of the measure transportation results of Chernozhukov et al. (2017), Hallin et al. (2020a) and del Barrio et al. (2020) to define multivariate version of ranks and signs, which are suitable for semi-parametric inference for multivariate time series. Moreover, measure transportation theory is rapidly becoming pivotal for machine learning research. Many data analysis techniques in computer vision, imaging (e.g. for color/texture processing or histograms comparisons), and more general machine learning problems about regression, classification, and generative modeling are often based on optimal transportation theory; see Peyré and Cuturi (2019). The Wasserstein distance is also useful for contrasting complex objects and can be applied to signal processes and engineering; see Kolouri et al. (2017) for a survey.

In spite of this growing body of literature on the statistical applications of measure transportation theory, there is no paper which identifies the close connections between the Monge-Kantorovich results, the theory of saddlepoint approximations, and information theory. It is our aim to fill that gap. We review the fundamental results available in the literature, with the purpose of drawing the theoretical connections between them. This effort has its motivation not only in the intrinsic intellectual challenge of relating different branches of mathematics and statistics, but also in the methodological added value. Indeed, working across the fields of information theory, statistics and measure transportation allows us to identify the potential transfer of existing knowledge (and technology) developed in one field to the other two fields. The selected review presented in this paper and the unveiled connections lay down the theoretical foundation to trigger that transfer. Finally, we believe that our results can have an impact on some areas of statistical education: they offer novel approaches to introduce higher-order techniques, giving the opportunity of looking at statistical and mathematical problems from different angles.

1.2 Outline

We consider the derivation of the saddlepoint approximation via the method of the conjugate density, which hinges on *convex analysis* results. The key tool of our development is the Legendre transform. We first consider an abstract setting in §2: we work with a generic random variable and we set up the basic notions. In §3 we turn to the statistical framework and explain how the links discussed in §2 work for well-known statistics, whose cumulant generating function (c.g.f.) is available in closed form. In §4, we extend the connections to M-estimators, which are general tools for conducting parametric inference. In §5 we mention some possible future research directions. The code (R and MATLAB) to replicate the numerical exercises of this paper is available at https://github.com/dvdlvc/MyGitHub/tree/Saddlepoint_MeasureTransportation.

2. KEY MATHEMATICAL ASPECTS

2.1 Esscher's tilting, saddlepoint, and conjugate density

2.1.1 Definitions. Let $X \sim F_X$, where $dF_X(x) = f_X(x)dx$, or equivalently, let us assume that X has a measure μ (which is absolutely continuous w.r.t. the Lebesgue measure), whose support is $\mathcal{X} \subseteq \mathbb{R}$. Then, given $t \in \mathbb{R}$, we define the conjugate density,

$$h_t(x) = C(t)\exp\{\upsilon(t)(x-t)\}f_X(x), \tag{1}$$

where $\upsilon(t)$ is chosen such that $E_{h_t}[X] = t$, E_{h_t} represents the expected value taken w.r.t. h_t. Eq. (1) defines an embedding of f_X into an exponential family and transforms the whole density—it does not change only the mean, rather it changes all existing moments: Eq. (20) below illustrates this aspect for the variance. The function $C : \mathbb{R} \to \mathbb{R}^+$ is defined as $C(t) = \exp\{\upsilon(t)t - K_X[\upsilon(t)]\}$, with

$$K_X(\upsilon) = \log E_{f_X}[\exp\{\upsilon X\}]$$

representing the c.g.f. of X and $C(t)$ is such that h_t integrates to one. In fact, $\upsilon(t)$ is the saddlepoint at t, obtained by solving

$$K'_X(\upsilon) = t; \tag{2}$$

see Field and Ronchetti (1990) p. 34-35.

The conjugate density highlights a connection between $\upsilon(t)$ and information theory. Indeed, (1) illustrates that $\upsilon(t)$ defines a transformation of the original density f_X into h_t. Theorem 2.1 in Kullback (1997) shows that h_t is the solution to the information theoretic problem:

$$\min_{g\in\mathcal{G}} \mathrm{KL}(g, f_X), \quad \text{s.t.} \quad g(x) \geq 0, \quad \int_{\mathcal{X}} g(x)dx = 1, \quad E_g[X] = t\,, \tag{3}$$

where $\mathcal{G}$ contains all the densities having support $\mathcal{X}$ and finite second moment, while KL represents the (backward) Kullback-Leibler divergence

$$\mathrm{KL}(g, f_X) = \int_{\mathcal{X}} g(x) \log \frac{g(x)}{f_X(x)} dx.$$

The proof of this result is available in Appendix A. We refer to Kremer (1982) p. 59 and to Esscher (1932) for a discussion.

The transformation $f_X \mapsto h_t$, commonly called Esscher's transformation (Kremer (1982)) or exponential tilting (Barndorff-Nielsen and Cox (1989), p. 105), is related to the Legendre transform (henceforth, denoted by the symbol *) of K_X:

$$K_X^*(t) = \sup_{\upsilon} \{\upsilon t - K_X(\upsilon)\}, \tag{4}$$

where the maximising value $\upsilon(t)$ is the saddlepoint. The function $-K_X^*(t)$ is called the (point) entropy of the density f_X; see McCullagh (2018). In the literature on convex analysis, the term convex conjugate is also applied; see Rockafellar (2015), §12. We notice that

$$K_X^*(t) = \upsilon(t)t - K_X(\upsilon(t)) = \log[\exp\{\upsilon(t)t - K_X(\upsilon(t)\}] = \log C(t). \tag{5}$$

2.1.2 Graphical illustration. In what follows it is convenient to consider the set of υ-values, denoted by Υ, for which $K_X(\upsilon) < \infty$ as being, in a sense, dual to the sample space of possible averages of identically distributed Xs. Corresponding to $K_X(\upsilon)$ defined on Υ, there is a dual function $K_X^*(t)$ defined on $\mathcal{X}$ such that $K_X'(\upsilon)$ and $\partial_t K_X^*(t)$ are functional inverses.

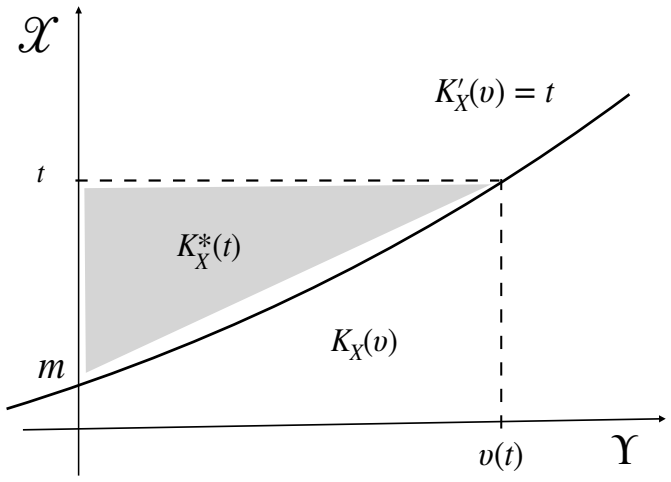

FIG 1. *Graphical illustration of the Legendre transform of K_X, when $\mathcal{X} \subseteq \mathbb{R}$. For the sake of the graphical representation, we assume $m > 0$. Figure adapted from McCullagh (2018), Ch. 6.*

Looking at Figure 1, the curved solid line represents $K_X'(\upsilon)$ plotted against υ and is such that $K_X'' > 0$. The intercept is equal to $K_X'(\upsilon)|_{\upsilon=0} = m = E_{f_X}(X)$. The area under the curve from zero to $\upsilon(t)$ is $K_X(\upsilon) = \int_0^{\upsilon} K_X'(\xi) d\xi$. For any given value of $t \in \mathcal{X}$, $\upsilon(t)t$ is the area of the rectangle whose opposite corners are at the origin $(0,0)$ and at $(\upsilon(t), t)$. The figure illustrates that the relation $\sup_\upsilon\{\upsilon t - K_X(\upsilon)\}$ is satisfied at $\upsilon(t)$, where $K_X'(\upsilon) = t$. The shaded area above the curve is equal to $\upsilon(t)K'(\upsilon(t)) - K(\upsilon(t))$. Viewed as a function of t, this is the Legendre transform of K_X, as in (5).

2.2 Measure transportation

2.2.1 Definitions. In the 18th century, Monge (1781) formulated a mathematical problem that in modern language can be expressed as follows. Let μ and ν denote two probability measures over $(\mathbb{R}, \mathcal{B})$, where $\mathcal{B}$ represents the Borel sigma-field. Let $c : \mathbb{R}^2 \to \mathbb{R}$ be a Borel-measurable cost function such that $c(x, y)$ is the cost of transporting x to y. Then, for $X \sim \mu$, $Y \sim \nu_t$, solve

$$\inf \left\{ M(\mathcal{T}) := \int_{\mathbb{R}} c(x, \mathcal{T}(x)) d\mu \quad \text{such that} \quad \mathcal{T} : X \to Y \right\}. \tag{6}$$

We say that the solution to (6) is the mapping $\mathcal{T}$ such that $\mathcal{T}_{\#}\mu = \nu_t$, to be read as $\mathcal{T}$ pushes μ forward to ν_t. Specifically, if μ is a Borel measure on $\mathcal{X} \subseteq \mathbb{R}$ and $\mathcal{T}$ is a Borel map $\mathcal{X} \to \mathcal{X}$, then $\mathcal{T}_{\#}\mu = \nu_t$ stands for the image measure (or push-forward) of μ by $\mathcal{T}$: this is a Borel measure on $\mathcal{X}$, defined by $(\mathcal{T}_{\#}\mu)[A] = \mu[\mathcal{T}^{-1}(A)]$ for any Borel set A.

The map $\mathcal{T}$ appearing in all these statements is called the optimal transportation map. Informally, one can say that $\mathcal{T}$ transports the mass represented by the measure μ, to the mass represented by the measure ν_t. The transportation map which solves (6) (for a given c) is called an optimal transportation map (for the selected c).

Monge's problem (defined using $c(x, y) = |x - y|$) remained open until the 1940s, when it was revisited by Kantorovich (1942), who introduced the notion of coupling in the following way. Let μ and ν_t belong to the family $\mathcal{G}$, and let $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a Borel-measurable cost function. The objective is to minimize

$$\mathrm{KP}(\mu, \nu_t) = \inf_{\gamma \in \Gamma(\mu, \nu_t)} \int_{\mathcal{X} \times \mathcal{X}} c(x, y) d\gamma(x, y), \tag{7}$$

where the infimum is over all pairs (X, Y) of (μ, ν_t), belonging to $\Gamma(\mu, \nu_t)$, the set of probability measures γ on $\mathcal{X} \times \mathcal{X}$, satisfying $\gamma(A \times \mathcal{X}) = \mu(A)$ and $\gamma(\mathcal{X} \times B) = \nu_t(B)$, for Borel sets A, B. Kantorovich's problem is more general than Monge's problem, since it allows mass splitting.

We consider Kantorovich's dual formulation to the primal optimization problem (7). The derivation of the dual problem follows standard duality arguments, that we sketch in Appendix B. Here we say that the Kantorovich's dual problem is

$$\begin{aligned} \mathrm{KD}(\mu, \nu_t) \; &= \; \sup_{\phi, \varphi} \left(\int_{\mathcal{X}} \varphi(y) d\nu_t(y) - \int_{\mathcal{X}} \phi(x) d\mu(x) \right) \\ &\text{s.t. } \varphi(y) - \phi(x) \leq c(x, y), \quad \forall (x, y). \end{aligned} \tag{8}$$

Under suitable conditions on c (e.g. for a convex $c(x, y) = (1/2)(x - y)^2$, see Th. 1.40 in Santambrogio (2015)), we have $\mathrm{KP}(\mu, \nu_t) = \mathrm{KD}(\mu, \nu_t)$, namely there is no duality gap. The solution to $\mathrm{KD}(\mu, \nu_t)$ is

$$\begin{aligned} \varphi(y) &= \inf_x [\phi(x) + c(x, y)] \\ \phi(x) &= \sup_y [\varphi(y) - c(x, y)]. \end{aligned} \tag{9}$$

The equations in (9) imply that the functions ϕ and φ are related to each other through the so-called c-transform, whose functional form depends on c. The c-transform of φ is indicated by φ^c and functions (φ, φ^c) satisfying

$$\sup_{\varphi} \left(\int_{\mathcal{X}} \varphi(y) d\nu_t(y) - \int_{\mathcal{X}} \varphi^c(x) d\mu(x) \right)$$

are called Kantorovich potentials for the transport from μ to ν_t.

For the quadratic cost, we have $c(x, y) = -xy$ (see below) and the c-transform coincides with the Legendre transform of ϕ. Therefore, (ϕ, ϕ^*) is a pair of Kantorovich potentials for μ to ν_t—a graphical intuition can be obtained looking again at Figure 1 and replacing: K_X by ϕ, v by x, and t by y.

The support of γ (henceforth denoted as, spt(γ)) is defined as the smallest closed set on which γ is concentrated. If we fix a point $(x_0, y_0) \in \text{spt}(\gamma)$, we have $\phi(x_0) = \varphi(y_0) - c(x_0, y_0)$. Thus, the measure γ is concentrated on the graph which expresses the map associating y_0 to each x_0 and this map is the optimal transport $\mathcal{T}$—more precisely, for the last statement to be true, we need that the cost function c satisfies the twist condition (see Santambrogio (2015) p. 14, Definition 1.16). We refer to e.g. Villani (2009) for a discussion oriented towards the mathematics, while we refer to Chernozhukov et al. (2017), Hallin (2017), and Panaretos and Zemel (2019) for a discussion in a statistical context. We refer also to Ch. 1 of Panaretos and Zemel (2020) for an introduction to the problem of optimal transport in probabilistic terms.

We consider often the quadratic cost $c(x, y) = (1/2)(x - y)^2$, which satisfies the twist condition. We remark that solving the primal problem KP(μ, ν_t) with the quadratic cost is equivalent to solving the same problem with $c(x, y) = -xy$. To see this, let us consider the elementary equality $(1/2)(x-y)^2 = (1/2)(x^2+y^2-2xy)$, which implies that the interaction between x and y in the quadratic cost function is the same as in $c(x, y) = -xy$ (recall that in KP(μ, ν_t) the marginal of X and Y are fixed). As a result, solving the problem KP(μ, ν_t) is equivalent to finding the $\sup_\gamma E_\gamma(XY)$, where the supremum is over all coupling (X, Y) of (μ, ν_t), so the problem is to maximize the correlation between the random variables X and Y. Within this setting, one can prove (see e.g. Galichon (2016), Th. 4.8) that ϕ is a convex function and it is such that

$$\phi(x) = \beta_1 + \int_{x_0}^{x} \mathcal{T}(u)du, \tag{10}$$

for $x, x_0 \in \mathcal{X}$ and the constant $\beta_1 \in \mathbb{R}$. Differentiating (10), we obtain $\mathcal{T} = \nabla\phi$, namely the optimal transport map is the gradient of ϕ.

Moreover, for $\mathcal{X} \subseteq \mathbb{R}$, when μ has c.d.f. F_X and $Y \sim \nu_t$ with c.d.f. H_t, one can prove (see e.g. Santambrogio (2015) Ch. 2) that

$$\mathcal{T} = H_t^{-1} \circ F_X. \tag{11}$$

Finally, we remark that the optimal mapping $\mathcal{T}$ is such that, for all ν_t-integrable functions ω, we have

$$\int_{\mathcal{X}} \omega(y)d\nu_t(y) = \int_{\mathcal{X}} \omega(\mathcal{T}(x))d\mu(x), \tag{12}$$

which expresses that the $\mathcal{T}$ yields a change of variable $X \mapsto Y$.

2.2.2 Wasserstein distance, Wasserstein space, and geodesics. The solution to KP(μ, ν_t) with $c(x, y) = |x - y|^p$ defines the p-Wasserstein distance, for $p \geq 1$,

$$W_p(\mu, \nu_t) = \left(\inf_{\gamma \in \Gamma(\mu, \nu_t)} \int_{\mathcal{X}\times\mathcal{X}} |x - y|^p d\gamma(x, y) \right)^{1/p}.$$

For any $p \geq 1$, W_p is a metric on the set of Borel probability measures on $\mathcal{X}$, with finite p-th moment. For $p = 2$, the squared 2-Wasserstein distance is

$$W_2^2(\mu, \nu_t) = \inf_{\gamma \in \Gamma(\mu, \nu_t)} \int_{\mathcal{X}\times\mathcal{X}} (1/2)(x - y)^2 d\gamma(x, y),$$

where the infimum is taken over all pairs (X, Y) of (μ, ν_t), belonging to $\Gamma(\mu, \nu_t)$. Wasserstein distances are ordered in the sense that $p \geq q \geq 1$, $W_p(\mu, \nu_t) \geq W_q(\mu, \nu_t)$.

When $\mathcal{X}$ is a convex subset of $\mathbb{R}$, let us denote by $\mathcal{P}(\mathcal{X})$ the set of all probability measures defined on $\mathcal{X}$ and admitting finite second moment. The resulting Wasserstein space $(\mathcal{P}(\mathcal{X}), W_2)$ is a metric space. We briefly recall the key aspects of Wasserstein space that are relevant for our arguments; we refer e.g. to Panaretos and Zemel (2020) Ch. 2 for additional details. The space $(\mathcal{P}(\mathcal{X}), W_2)$ is a geodesic space. We recall that a geodesic space refers to a metric space $\mathcal{M}$ in which every pair of points $x, y \in \mathcal{M}$ is connected by a continuous curve $s \in [0, 1] \to x(s) \in \mathcal{M}$ which satisfies $\|x - x(s)\| = s\|x - y\|$ and

$\|x(s) - y\| = (1-s)\|y - x\|$. Such a curve is called a geodesic (segment), which is such that $x(0) = x$ and $x(1) = y$; see e.g. McCann and Guillen (2011) for a survey in the context of optimal transportation. Thus, in $(\mathcal{P}(\mathcal{X}), W_2)$, for $\mu, \nu_t \in \mathcal{P}(\mathcal{X})$, there exists a continuous path going from μ to ν_t, such that its length is the distance between the two measures.

To elaborate further, we recall that in the Wasserstein space the geodesics are easily characterized and they are given by the so-called displacement interpolation (a.k.a. McCann interpolation). Specifically, the geodesic in $(\mathcal{P}(\mathcal{X}), W_2)$ is obtained exploiting the geodesic properties of $(\mathcal{X}, \|x - y\|)$: for $s \in [0,1]$ and given $\mathcal{T}$, we set $\mathcal{T}_s(x) = (1-s)x + s\mathcal{T}(x)$. This is a simple linear interpolation of the transport map and the identity function, thus uniqueness of $\mathcal{T}$ implies uniqueness of the geodesic. We interpret $\mathcal{T}_s$ as the position at time s of the mass initially at x. We remark that $\mathcal{T}_0 \equiv Id$ (the identity function), while $\mathcal{T}_1 \equiv \mathcal{T}$. In the 2-Wasserstein space, the geodesic $\gamma : [0,1] \to \mathcal{P}(\mathcal{X})$ is the parameterized curve from μ to ν_t:

$$\gamma(s) = \mathcal{T}_{s\#}\mu. \tag{13}$$

Eq. (13) indicates that the geodesic depends on the optimal transport map via $\mathcal{T}_s$. The velocity of each particle is $\partial_s \mathcal{T}_s(x) = \mathcal{T}(x) - x$, while its acceleration is $\partial_s^2 \mathcal{T}_s(x) \equiv 0$. Thus, we conclude that the mass of μ is transported to the mass of ν_t at a constant speed, along $\gamma(s)$. We refer to §5.2 for additional comments.

2.2.3 Graphical illustration. In Figure 2, we consider the problem of transporting μ onto a new measure ν_t, using the quadratic cost. The original measure μ (top left panel) is a mixture of Gaussians and it is moved to the target measure ν_t (top right panel) by the map $\mathcal{T}$ (bottom left panel, obtained by numerical method to invert the c.d.f. of ν_t), which does the push-forward of μ toward ν_t. Comparing the p.d.f. of the original random variable $X \sim \mu$ to the one of the target $Y \sim \nu_t$, we see that the probability mass has been moved from the left to the right. In the bottom right panel, we display some p.d.f.s located along $\gamma(s)$, for some selected values of s: the panel depicts the different steps though which $\mathcal{T}$ moves the mass in μ, along the geodesic—in fact, the mass displacement happens continuously in s.

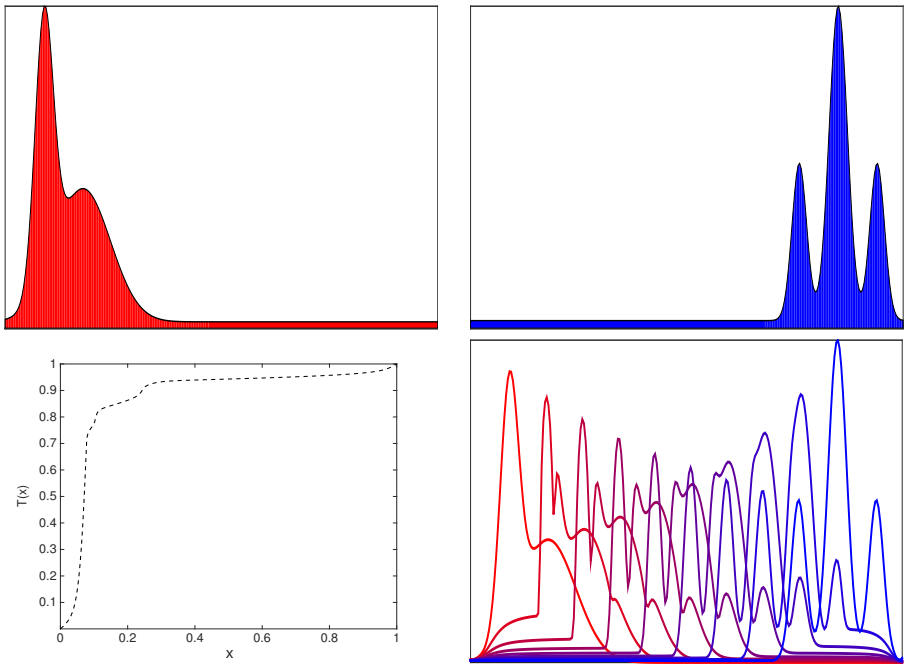

FIG 2. *Top left panel: original measure μ. Top right panel: target measure ν_t. Bottom left panel: monotone transport map $\mathcal{T}$ (x- and y-axis are normalized to $[0,1]$). Bottom right panel: sequence of p.d.f.s where each measure μ_s is obtained via $\mathcal{T}_s$ (displacement interpolation). The plot is inspired by Peyré and Cuturi (2019).*

2.3 Connecting optimal transportation theory to saddlepoint theory

We can think of the conjugate density method in the following way: we start from $X \sim \mu$, having density f_X, and obtain $Y = \mathcal{T}(X)$ with $Y \sim \nu_t$, where the measure ν_t is absolutely continuous w.r.t. the Lebesgue measure, it has support $\mathcal{X}$ and a c.d.f. H_t, such that $dH_t(y) = h_t(y)dy$. From this perspective, the saddlepoint $\upsilon(t)$ is related to a measure transportation of μ onto ν_t. Moreover, from §2.1 we know that K_X^* defines $\upsilon(t)$ and it yields h_t, which minimizes $\mathrm{KL}(g, f_X)$ as in (3). Then, a question naturally arises:

"What is the link between the saddlepoint $\upsilon(t)$ and the optimal transportation problem as defined in (6) and/or (7)?"

To answer this question, we need to investigate the link between the optimal transportation map $\mathcal{T}$ and $\upsilon(t)$. To this end, we state the following:

PROPOSITION 2.1. *Let X and Y be two random variables such that $X \sim \mu$ and $Y \sim \nu_t$. Both measures have support $\mathcal{X} \subseteq \mathbb{R}$ and are absolutely continuous w.r.t. the Lebesgue measure. The c.d.f. associated to μ is F_X, having p.d.f. f_X and c.g.f. equal to $K_X(\upsilon) = \log E_{f_X}[\exp\{\upsilon X\}]$, whose Legendre transform is $K_X^*(t)$. The measure ν_t is such that $E_{\nu_t}[Y] = t$, for $t \in \mathcal{X}$, the c.d.f. of Y is H_t, while its p.d.f. h_t is the conjugate density*

$$h_t(x) = C(t)\exp\{\upsilon(t)(x-t)\}f_X(x),$$

with $\upsilon(t)$ satisfying $K_X'(\upsilon) = t$ and $C(t) = \exp\{K_X^*(t)\}$ as in (5). Let $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be $c(x, y) = -xy$. Then, there exists a unique optimal transport plan γ solution to (7), which is of the form $(Id, \mathcal{T})_{\#}\mu$, where $\mathcal{T}$ is as in (11).

The proof is available in Appendix C and it follows from some well-known results in measure transportation theory. To elaborate on the statement, some comments are in order.

Comment (i). For each $t \in \mathcal{X}$, the conjugate density h_t is the solution to the information theoretic problem (3). Specifically, for each t we start from f_X and we associate to it h_t, which is characterized by the saddlepoint $\upsilon(t)$, which in turn is connected to K_X and K_X^*, in the sense of (2) and (4). By the very definition of Legendre transform, the pair (K_X, K_X^*) satisfies $K_X(\upsilon(t)) + K_X^*(t) = \upsilon(t)t$, for all $(\upsilon(t), t)$, so that (2) holds.

Comment (ii). Once h_t is known, we can build upon the optimal transportation theory and look for the optimal plan which solves $\mathrm{KP}(\mu, \nu_t)$. For the quadratic cost, the solution of (7) (the original Monge's problem, without mass splitting) is the plan induced by the optimal map $\mathcal{T} = \nabla\phi$, where the Kantorovich potential ϕ is as in (10) and the pair (ϕ, ϕ^*) is the solution of $\mathrm{KD}(\mu, \nu_t)$. Making use of Proposition 2.1, we have $H_t^{-1} \circ F_X(x) = \nabla\phi(x)$, where the expression of H_t is characterized by $K_X'(\upsilon(t))$: this unveils the link between $\mathcal{T}$ and $K_X'(\upsilon(t))$. Combining these considerations, we conclude that $\upsilon(t)$ is related to the solution of the information theoretic problem (3) and, at the same time, it also characterizes the solution of the optimal mass transportation problem (7)

Comment (iii). The fact that we can now focus on the map $\mathcal{T}$ which transforms X into Y offers a completely new perspective from which we can look at the saddlepoint approximation. The change of variable yielded by $\mathcal{T}$ is related to the Esscher's tilting and it is characterized by the saddlepoint. This opens the door to another link between optimal transportation theory and statistics. Indeed, in probability and statistics, it is customary to write the Jacobian formula for $X \mapsto Y$ and, in the univariate case, the density of $Y = \mathcal{T}(X)$ satisfies:

$$f_Y(y) = f_X\left[\mathcal{T}^{-1}(y)\right]\left|\frac{\partial\mathcal{T}^{-1}(y)}{\partial y}\right|. \tag{14}$$

Clearly, $f_Y(y) \equiv h_t(y)$ and Eq. (14) connects the conjugate density with the Jacobian formula expressed in terms of optimal transportation map.

2.4 Example: exponential random variable

Let us consider a random variable having an exponential p.d.f. with rate one, namely $X \sim \exp(1)$ so $X \sim \mu$, and define a target variable Y, having mean t, so $Y \sim \nu_t$.

Conjugate density method (illustration of Comment (i)). The c.g.f of X is

$$K_X(\upsilon) = \log(1/(1-\upsilon)),$$

which is defined for $0 < \upsilon < 1$, and solving (2) yields the saddlepoint $\upsilon(t) = 1 - 1/t$. From (1) it follows that the conjugate density is

$$h_t(y) = C(t) \exp\{\upsilon(t)(y-t) - y\}, \tag{15}$$

where $C(t) = \exp\{t - 1 - \log t\} = \exp\{K_X^*(t)\}$, where $K_X^*(t) = (t-1) - \log(t)$. Simple algebraic manipulations yield

$$h_t(y) = (1/t) \exp\{-y/t\},$$

namely $Y \sim \exp(1/t)$. The c.d.f. $H_t(x) = \int_0^x h_t(u)du$ is related to the measure ν_t.

Change of variable and Kantorovich potential (illustration of Comment (ii) and Comment (iii)). Now, let us express the connection to the optimal transportation problem. To this end, we notice that $H_t(x) = 1 - \exp\{-x/t\}$ so $H_t^{-1}(u) = -t\log(1-u)$, for $u \in (0,1)$. From Proposition 2.1, we have $Y = \mathcal{T}(X) = H_t^{-1} \circ F_X(X) = tX$. The Jacobian formula (Comment (iii)) yields that the p.d.f. of Y is

$$f_Y(y) = (1/t) \exp\{-y/t\}.$$

Hence, looking at (15) we have that $f_Y(y) \equiv h_t(y)$. From (10), we have, for $c_1 \in \mathbb{R}$, that

$$\phi(x) = c_1 + t \int_0^x u \, du = c_1 + \frac{x^2}{2} K_X'(\upsilon(t)),$$

which shows that the Kantorovich potential is explicitly related to K_X (Comment (ii)). Finally, the Legendre transform of ϕ is

$$\phi^*(y) = \frac{1}{2}\left(\frac{y^2}{K_X'(\upsilon(t))}\right) - c_1.$$

Graphical interpretation. In Figure 3, we provide a graphical illustration of the optimal transportation map. In the top left panel, we display the histogram of an observed sample $\{x_i\}_{i=1}^{40}$ drawn from an exponential with rate one (related to the measure μ). The bottom right panel illustrates the optimal map $\mathcal{T}(x) = tx$, applied to each x_i: the map is the push-forward of the original measure μ onto ν_t, for $t = 2$, i.e., $\mathcal{T}_{\#}\mu = \nu_2$. The map is plotted in the form of vectors acting on the observed data, where each arrow indicates the source and destination of the mass being transported for each x_i. The arrows illustrate that the map $\mathcal{T}$ acts more on those x_i's which are between the origin and the vertical dotted line, which represents the mean of the target random variable Y. Each arrow represents a geodesic on the ground space $\mathcal{X} = (0, \infty)$: it depicts that x_i is transported to a unique target point y_i by the optimal transport $\mathcal{T} = H_t^{-1} \circ F_X$. In the top right panel, we display the histogram of the sample $\{y_i\}_{i=1}^{40}$ as obtained applying $\mathcal{T}$ to each x_i.

3. THE CONNECTIONS FOR SOME SIMPLE STATISTICS

How are the connections unveiled in §2 related to the inference problem of deriving an approximation to some statistics of interest? To answer this question we start from the sampling distribution of the sample mean. We let $X \sim f$ (for the ease-of-notation, we use f rather than f_X to denote the density of X), with f related to a measure μ. We assume we are given a random sample $X_1, ..., X_n$ of i.i.d. copies of X, whose K_X is the well-defined.

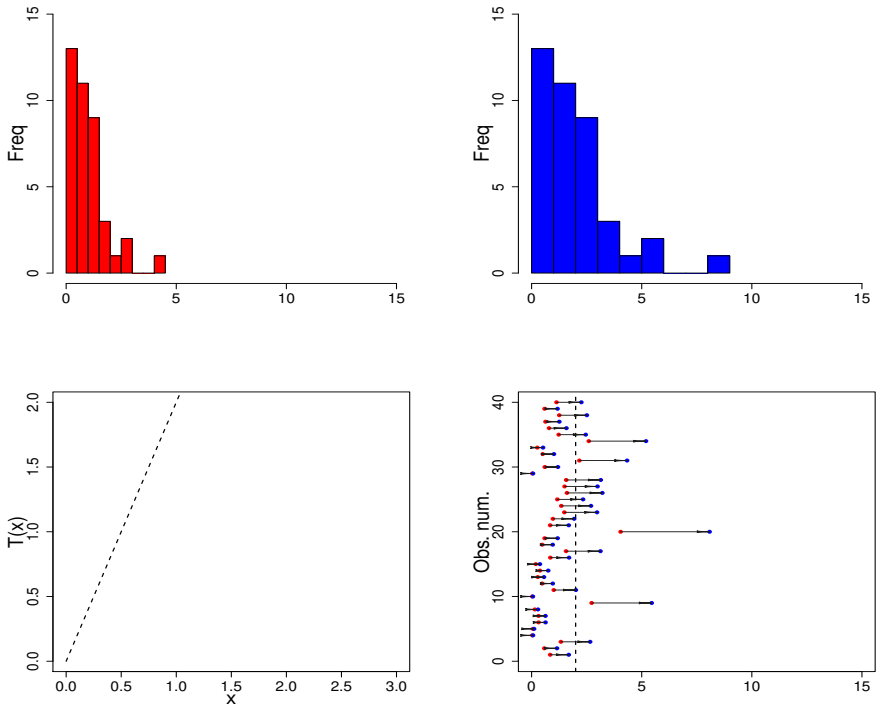

FIG 3. *Optimal transportation for the exponential random variable. Top left panel, histogram of 40 random observations drawn from an exponential with rate one (related to the measure μ). Top right panel: histogram of 40 random observations as transformed by $\mathcal{T}$, which yields an exponential with rate 1/2. Bottom left panel: the optimal transportation map $\mathcal{T}(x) = 2x$, for $x \in [0, 3]$. Bottom right panel: optimal transportation map applied to each x_i: each arrow is a geodesic, indicating the source and destination of the transported particle x_i. The vertical dotted line at 2 represents the mean of the target measure υ_t.*

3.1 The sample mean

Let $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ be the mean of $X_1, ..., X_n$ and let us denote its density by f_n. A first improvement on the standard asymptotic normal distribution of the standardized mean is provided by the Edgeworth expansion. It is obtained by Fourier inversion of a Taylor expansion of the characteristic function of the mean around 0. This leads to an expansion of the distribution of the standardized mean, where the dominant term is the standard normal distribution followed by additive terms of order $n^{-r/2}$, $r = 1, 2, \ldots$. In the case of the sample mean, the Edgeworth expansion including the first three terms is

$$\begin{aligned} f_{\text{Edg}}(x) &= \phi(x) \\ &\quad + n^{-1/2}\frac{\lambda_3}{6}(x^3 - 3x)\phi(x) \\ &\quad + n^{-1}\left[\frac{\lambda_4}{24}(x^4 - 6x^2 + 3) + \frac{\lambda_3^2}{72}(x^6 - 15x^4 + 45x^2 + 15)\right]\phi(x), \end{aligned}$$

with λ_3 and λ_4 being the standardized cumulants of X of order three and four. For details, we refer e.g. to Barndorff-Nielsen and Cox (1989) and Field and Ronchetti (1990), Ch. 2. Hall (1992), Ch.2, and Kolassa (2006), Ch.3.

By construction the approximation to f_n obtained by f_{Edg} typically works well in the center of the distribution, but deteriorates quickly in the tails.

3.1.1 Saddlepoint approximation for the mean. Let $K_{\bar{X}_n}(\upsilon) = \log E_{f_n}[\exp\{\upsilon\bar{X}_n\}]$ be the c.g.f. of $\bar{X}_n$. By standard Fourier inversion, the density f_n is obtained as

$$f_n(t) = \frac{n}{2\pi i}\int_{-i\infty}^{i\infty} \exp\{n(K_X(\upsilon) - \upsilon t)\}d\upsilon. \tag{16}$$

The integral in (16) is typically not available in a closed form. However, an approximation to f_n can be obtained by applying the method of steepest descent, followed by the Cauchy's theorem (to deform the path of the integral in the complex domain) and Watson's lemma (to control the error of the approximation) accordingly; see Daniels (1954), p. 633 for mathematical details. The resulting approximation to f_n is the saddlepoint density approximation f_{sad}, namely

$$f_{\text{sad}}(t) = \left[\frac{n}{2\pi K_X''(\upsilon(t))}\right]^{1/2} \exp\{n[K_X(\upsilon(t)) - \upsilon(t)t]\}, \tag{17}$$

where $K_X''(\upsilon(t))$ is the second derivative of K_X computed at $\upsilon(t)$. The approximation f_{sad} is such that

$$f_n(t) = f_{\text{sad}}(t)\{1 + O(n^{-1})\}. \tag{18}$$

See also Barndorff-Nielsen and Cox (1979) and Goutis and Casella (1999).

A few remarks are in order. The use of f_{sad} in (18) yields some well-known advantages over other routinely applied approximations. For instance, the Gaussian approximation is first-order accurate, with an absolute error of order $O(n^{-1/2})$, and it performs poorly in the tails. The combination of these two aspects entails large approximation errors in small samples and/or for large values of t in moderate samples. The Edgeworth expansion features higher-order accuracy, performs well in small samples and it yields an absolute error of order $O(n^{-1})$. However, f_{Edg} can become negative in the tails.

The saddlepoint approximation overcomes these problems. Indeed, by construction, f_{sad} is a density-like object which cannot become negative and it keeps its accuracy in the tails, providing accurate small sample approximations. With this regard, notice that the error in (18) is of order $O(n^{-1})$ and it is of relative type—to be contrasted with the absolute error entailed by the asymptotic theory and by the Edgeworth expansion. Moreover, Daniels (1954) proves that the size of the error holds uniformly in $t \in \mathbb{R}$. All these features are peculiar of f_{sad}.

3.1.2 Connection to measure transportation. Daniels (1954) provided an alternative derivation of f_{sad} using the method of the conjugate density h_t and this establishes the connection of this construction with measure transportation.

The basic idea of the conjugate density method is to recenter the density of $\bar{X}_n$ at the point of interest t and use a normal approximation in the recentered problem which leads to the approximation to $f_n(t)$. To perform this construction, we first embed the original density f into an exponential family, and then look for h_t which is the closest to f in KL distance and has a mean of t. Then, we compute a low-order Edgeworth expansion (see Bhattacharya and Rao (1986)) for the tilted density to obtain $f_{\text{sad}}(t)$. Finally, we repeat this procedure for every $t \in \mathbb{R}$. See also Reid (1988), p. 215. To clarify the connections between this procedure and the measure transportation, we consider three steps.

(i) Eq. (1) gives the expression of the conjugate density, which is defined by means of the saddlepoint $\upsilon(t)$. This implies that, by construction, h_t is such that $E_{h_t}[X] = t$ and we have

$$K_X'(\upsilon) = t \iff E_{h_t}[X] = t. \tag{19}$$

Moreover, simple algebra shows that

$$\begin{aligned} K_X''(\upsilon(t)) &= \frac{1}{E_f[e^{\upsilon(t)(X-t)}]} \int_{\mathbb{R}} (x-t)^2 e^{\upsilon(x-t)} f(x)dx \\ &= E_{h_t}(X^2) - [E_{h_t}(X)]^2 = var_{h_t}[X] =: \sigma^2(t). \end{aligned} \tag{20}$$

The statistic $\bar{X}_n$ is a linear combination of X_is and its c.g.f. is

$$K_{\bar{X}_n}(\upsilon) = \log E_{f_n}[\exp\{\upsilon \bar{X}_n\}] = nK_X(\upsilon/n). \tag{21}$$

Thus

$$K_{\bar{X}_n}'(\upsilon(t)) = t \iff E_{h_t}[\bar{X}_n] = t, \tag{22}$$

which illustrates that $\upsilon(t)$ performs a re-centering of the measure of $\bar{X}_n$, let us label it $\mu^{(n)}$, at the point of interest t.

(ii) The combination of (19) with (22) implies that $\upsilon(t)$ is such that $\mu^{(n)}$ has to be transported onto $\nu_t^{(n)}$, having mean t. To perform this measure transportation, we exploit the linearity of $\bar{X}_n$ in the X_is. Thus, we set KD(μ, ν_t) as in (8) with cost function $c(x, y) = -xy$. In this formulation, each $X_i \sim \mu$ and the measure μ is transported to ν_t, which has density h_t. The solution to KD(μ, ν_t) is obtained via the Legendre transform of $\phi(x)$, whose gradient is $\mathcal{T} = H_t^{-1} \circ F$ and depends on $K'_X(\upsilon(t))$.

(iii) The link between K_X and $K_{\bar{X}_n}$ in (21) induces the following

$$\begin{aligned} K^*_{\bar{X}_n}(t) &= \sup_\upsilon\{\upsilon t - K_{\bar{X}_n}(\upsilon)\} = \sup_\upsilon\{\upsilon t - nK_X(\upsilon/n)\} \\ &= n\sup_\upsilon\{(\upsilon/n)t - K_X(\upsilon/n)\} = n\sup_\upsilon\{\upsilon t - K_X(\upsilon)\} \\ &= nK^*_X(t), \end{aligned} \tag{23}$$

and $K^{*\prime}_{\bar{X}_n}(t) = nK^{*\prime}_X(t), K^{*\prime\prime}_{\bar{X}_n}(t) = nK^{*\prime\prime}_X(t)$. Then, from the definition of $C(t)$ and by (5), we obtain $C(t) = \exp\{\upsilon(t)t - K_X[\upsilon(t)]\} = \exp\{K^*_X(t)\}$, and

$$K^{*\prime}_X(t) = \frac{C'(t)}{C(t)} = \upsilon(t),$$

$$K^{*\prime\prime}_X(t) = \upsilon'(t) = \frac{1}{\sigma^2(t)} = \frac{1}{K''_X[\upsilon(t)]}.$$

Finally, using these expressions in (17), we can express the saddlepoint density approximation of f_n at t via the Legendre transform of K_X or $K_{\bar{X}_n}$

$$\begin{aligned} f_{\text{sad}}(t) &= \left(\frac{n}{2\pi\sigma^2(t)}\right)^{1/2} \exp\{n[K_X(\upsilon(t)) - \upsilon(t)t]\} \\ &= \frac{1}{\sqrt{2\pi}}\left(nK^{*\prime\prime}_X(t)\right)^{1/2} \exp\{-nK^*_X(t)\} \\ &= \frac{1}{\sqrt{2\pi}}\left(K^{*\prime\prime}_{\bar{X}_n}(t)\right)^{1/2} \exp\{-K^*_{\bar{X}_n}(t)\}. \end{aligned} \tag{24}$$

Therefore, our unveiled connections put under the spot light the key role of $K^*_{\bar{X}_n}$, which appears directly in f_{sad}.

3.1.3 Example (cont'd): mean of exponential random variables. Let us consider $\bar{X}_n$, the mean of n i.i.d. copies of X having an exponential density.

Density approximation. The plots in Figure 4 complete the information provided in Figure 3. In the top panel of Figure 4 we illustrate, for $n = 10$, the inaccuracy of the asymptotic theory, which performs poorly in the center and also in both tails of the distribution. The Edgeworth expansion yields accuracy improvements on the asymptotic theory. However, in the left tail the Edgeworth approximation becomes negative. In the same Figure we depict how the Esscher's tilting and the use of the saddlepoint density approximation overcome this problem. To illustrate this aspect, we compare the Edgeworth and the saddlepoint approximations to the exact density of $\bar{X}_n$—as far as the saddlepoint density is concerned, the formulae for the Legendre transform, the saddlepoint and the conjugate density are available in §2.4. Using these expressions in (24) we obtain f_{sad}. For each n, the exact (true, say) density of $\bar{X}_n$ is known (the Gamma distribution).

We consider $n = 10, 50, 250$. In the bottom panels of Figure 3 we display the relative error, computed as 100*(approx density - true density)/(true density), as entailed by each approximation for different sample sizes. The plots illustrate the improvement that the tilting of the

underlying distribution yields in terms of approximation accuracy, especially in the tails. For instance, when $n = 50$, the relative error entailed by the Edgeworth expansion and by the saddlepoint approximation are similar for $x \in [0.8, 1.2]$, but outside this central region the error of the Edgeworth is higher than the one entailed by the saddlepoint approximation. Looking at the plots, we see that the accuracy of Edgeworth improves as n increases, but for $n = 250$ the relative error entailed by the Edgeworth approximation in the left tail of the distribution is still higher than the error entailed by the saddlepoint approximation.

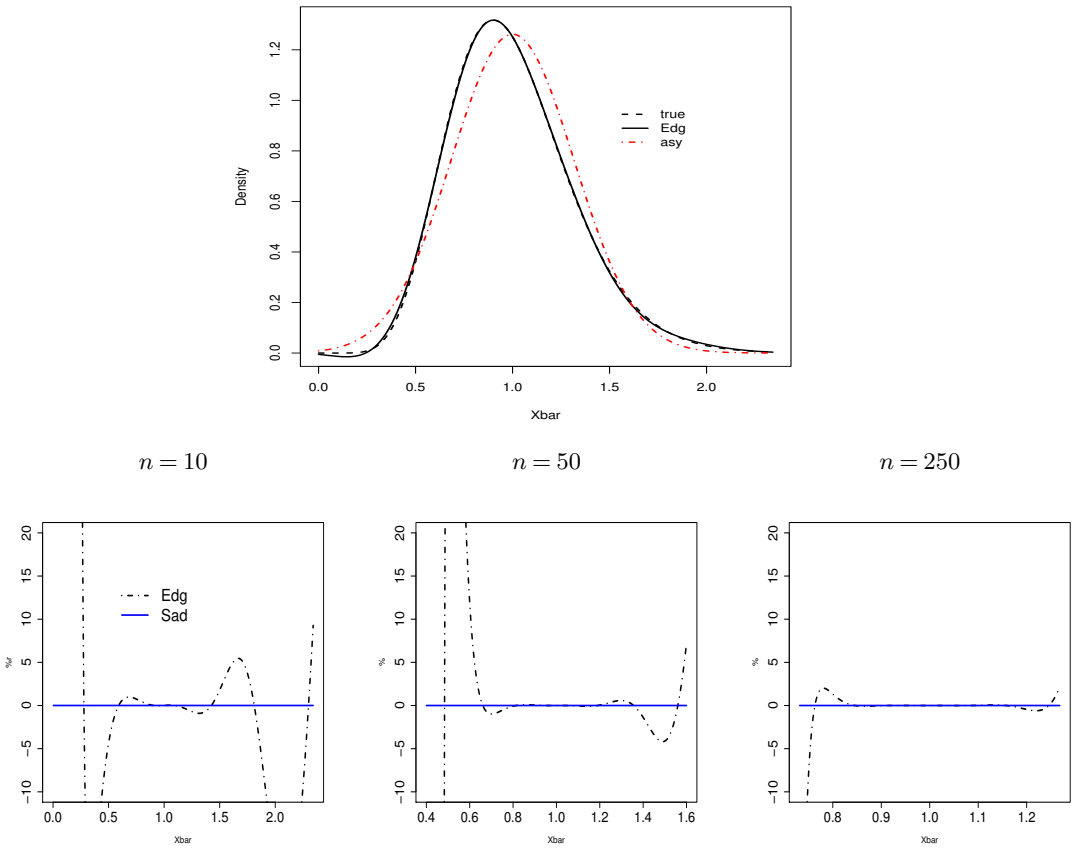

FIG 4. *Top panel: density of the sample mean ($\bar{X}_n$) of $n = 10$ i.i.d. random variable distributed as an exponential one. True density (dashed line), Edgeworth expansion (continuous line) and Gaussian asymptotic approximation (dash-dotted line). Bottom panels: relative error in percentage (y-axis) for the Edgeworth (dot-dashed line) and saddlepoint (continuous line) approximation to the density of the sample mean of n i.i.d. random variable distributed as an* $\exp(1)$. *The comparison is for different sample sizes.*

Saddlepoint test via Legendre transform. We consider the mean of n i.i.d. random variables $X_i \sim \exp(\alpha)$. Assume we want to test the hypothesis $\mathcal{H}_0 : \alpha = 1$ versus $\mathcal{H}_1 : \alpha > 1$. Following Robinson et al. (2003), to perform the test, we make use of the saddlepoint test statistic based on the Legendre transform of X (see §2.4 for its expression) evaluated at $\bar{X}_n$, i.e. $2nK_X^*(\bar{X}_n) = 2K_{\bar{X}_n}^*(\bar{X}_n)$. Robinson et al. (2003) prove that under $\mathcal{H}_0$ the distribution of this test statistic can be approximated by a $\chi^2(1)$ distribution, with relative error of order $O(n^{-1})$. This yields an accurate approximation of the level of the test, even for small sample sizes. The left panel of Figure 5 illustrates this aspect for $n = 10$. To obtain the plot, we simulate 5000 samples, drawing from $\exp(1)$ and for each sample we compute the test statistic. Finally, we do a QQplot, comparing the quantiles of the distribution of the test statistic with the quantiles of the $\chi^2(1)$ distribution. We see that the quantiles of the test statistic are close

to the ones of the $\chi^2(1)$: the approximation is remarkably accurate for the 0.95 and 0.975 quantiles, which are the quantiles typically applied for hypothesis testing. Similar pictures are available for $n = 50$ and 250, where the accuracy is even better.

Beside the behaviour under the null, another key aspect is the power of the saddlepoint test. To investigate it, in the right panel of Figure 5 we plot the power curves for $n = 10, 50, 250$. To obtain the plots, we proceed as described in the paragraph about the QQplot. The main difference is that, for each sample size, we simulate 5000 samples using a sequence of alternative hypotheses and we consider the frequency of non acceptance of $\mathcal{H}_0$. For each sample size, we have random drawings of size n from an $\exp(1+\delta)$, where $1+\delta$ represents the value of α under the alternative hypothesis and we consider $\delta \in [0, 0.8]$. We see that the test has good power already for $n = 10$. Clearly, the larger the sample size, the higher the power.

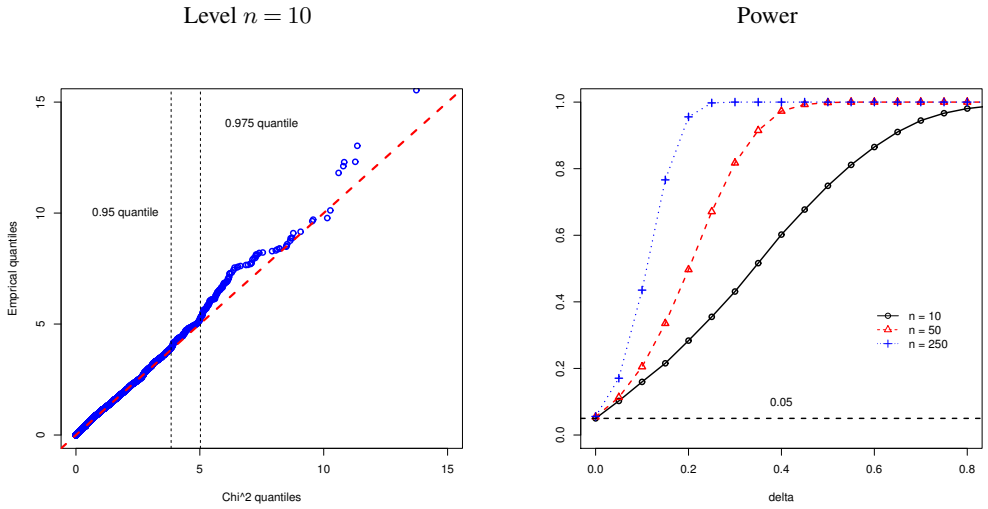

FIG 5. *QQplot for the test of the sample mean of n i.i.d. random variable distributed as an $\exp(\alpha)$. Left panel, $n = 10$ and $\alpha = 1$ (density under $\mathcal{H}_0$). Right panel, power for $n = 10, 50, 250$.*

3.2 General statistics with known cumulant generating function

Consider now a real-valued statistic $U_n = U(X_1, ..., X_n)$ having a density f_n and c.g.f. $K_{U_n}(\upsilon) = \log E_{f_n}[\exp\{\upsilon U_n\}]$, which is well-defined and known in a closed-form. It is important to note that U_n needs not be linear in the X_i's as it is the case with the sample mean. Then, by Fourier inversion we obtain:

$$f_n(t) = \frac{n}{2\pi i}\int_{-i\infty}^{i\infty} \exp\{nR_n(\upsilon) - n\upsilon t\}d\upsilon, \tag{25}$$

where

$$R_n(\upsilon) = K_{U_n}(n\upsilon)/n. \tag{26}$$

This is a generalization of (16): if $U_n = n^{-1}\sum_{i=1}^n X_i$, we have $R_n(\upsilon) \equiv K_X(\upsilon)$ and (25) coincides with (16). Following Easton and Ronchetti (1986), the saddlepoint equation is defined by $R_n'(\upsilon) = t$ and the saddlepoint density approximation $f_{\text{sad}}(t)$ is as in (18) and (17), with K_X replaced by R_n.

The interpretation of the saddlepoint procedure in terms of measure transportation remains essentially the same as in §3.1.2. Briefly, we have $U_n \sim \mu^{(n)}$ and we move its mass into a new measure $\nu_t^{(n)}$ such that $E_{\nu_t^{(n)}}[U_n] = t$. The function

$$R_n^*(t) = \sup_{\upsilon}\{\upsilon t - R_n(\upsilon)\}, \tag{27}$$

is the Legendre transform of R_n and the pair (R_n, R_n^*) characterizes the solution to the KD$(\mu_\theta^{(n)}, \nu_t^{(n)})$ problem with the quadratic cost function. The main difference with the construction in §3.1.2 lies in the fact that we can no longer exploit the linearity of the functional in the X_i's. Thus, we have to perform the Esscher's tilting directly on f_n and obtain its conjugate density, say $h_{n,t}$, via R_n^*. Doing so (see formula (2.9) in Easton and Ronchetti (1986)) we obtain

$$h_{n,t}(x) = C_n(t) f_n(x) \exp\{n[R_n(\upsilon(t) - \upsilon t]\},$$

with $\upsilon(t)$ is defined via $R'_{U_n}(\upsilon(t)) = t$ and $C_n(t)$ is an integration constant. Under $h_{n,t}$, similarly to (22), we have the desired recentering,

$$R'_{U_n}(\upsilon(t)) = t \iff E_{h_{n,t}}[U_n] = t. \tag{28}$$

4. M-ESTIMATORS

We consider the general case in which the c.g.f. of the statistic we are working with is not available in closed form. Among the possible statistics with this property, we focus on M-estimators, whose saddlepoint density approximations are derived in Field (1982). For the sake of exposition, we confine ourselves to the case $\theta \in \Theta \subset \mathbb{R}$. In §5.4, we discuss the case of multivariate parameter.

4.1 Notation

Let $X \sim F$, where $dF(x) = f(x)dx$, or equivalently, let us assume that X has measure μ, whose support is $\mathcal{X} \subseteq \mathbb{R}$. We are interested in conducting inference on some parameter $\theta(F)$. Given a random sample $X_1, ..., X_n$ of i.i.d. copies of X, an M-estimator of θ is the solution U_n to

$$\sum_{i=1}^{n} \psi(X_i; \theta) = 0, \tag{29}$$

where $\psi : \mathcal{X} \times \Theta$ and $\Theta \subseteq \mathbb{R}$. The population version of (29) is $E_f[\psi(X; \theta)] = 0$, where E_f represents the expected value taken w.r.t. to f. M-estimators include the maximum likelihood estimator as a special case, when F belongs to a parametric family of distributions $\{F_\theta\}$ and $\psi(x; \theta) = \partial_\theta \log f_\theta(x)$. Moreover, setting $\psi(x; \theta) = x - \theta$ in (29), U_n becomes the sample mean. We refer to Huber (1981), Huber and Ronchetti (2009), Hampel et al. (1986), van der Vaart (1998), and Serfling (2009) for book-length presentations.

Standard first-order asymptotic results establish the asymptotic normality of M-estimators via the first-order von Mises expansion:

$$U_n - \theta(F) = \frac{1}{n} \sum_{i=1}^{n} IF_\psi(X_i; \theta, F) + o_p(n^{-1/2}), \tag{30}$$

where IF_ψ is the influence function (Hampel (1974))

$$IF_\psi(x; \theta, F) = \psi(x; \theta)/M(\theta), \tag{31}$$

with $M(\theta) = E_f[-\partial_\theta \psi(X; \theta)]$. The expansion in (30) is the starting point for first-order asymptotics: under suitable assumptions (see e.g. Huber (1981) Ch. 6.3),

$$n^{1/2}(U_n - \theta) \xrightarrow{D} \mathcal{N}(0, V(\theta)),$$

where $\xrightarrow{D}$ represents the convergence in distribution and $\mathcal{N}(0, V(\theta))$ is a Gaussian distribution with expectation zero and variance $V(\theta) = Q(\theta)/M^2(\theta)$, and $Q(\theta) = E_f[\psi^2(X; \theta)]$.

The expansion in (30) shows that the distributional properties of the M-estimator depend on the estimating function ψ; see also La Vecchia et al. (2012) and La Vecchia (2016) for a review. This provides the intuition for the fact that, to derive higher-order asymptotic properties of U_n, in particular its f_{sad}, we need to work with ψ. Thus, we let $K_\psi(\upsilon; \theta) = \log E_f[\exp\{\upsilon\psi(X; \theta)\}$ be the c.g.f. of $\psi(X; \theta)$ and we use $K_\psi(\upsilon; \theta)$ to show

how the arguments of §2 and §3 adapt to the case of M-estimators. The main difference with the previous sections is that the c.g.f. of U_n is (typically) unavailable in closed form. To overcome this problem we need to approximate K_{U_n} using (30) and obtain a saddlepoint equation based on K_ψ.

4.2 Saddlepoint approximation for M-estimators

Theorem 1 in Field (1982) shows that the density $f_n(t)$ of U_n can be approximated by an expansion as in (18), where the saddlepoint density approximation is

$$f_{\text{sad}}(t) = (n/2\pi)^{1/2} C^{-n}(t) \left| E_{h_{\psi,t}} \left[\frac{\partial \psi(X;t)}{\partial t} \right] \right| \left[E_{h_{\psi,t}} \left[\psi^2(X;t) \right] \right]^{-1/2}, \tag{32}$$

where

$$h_{\psi,t}(x) = C(t) e^{\upsilon(t)\psi(x;t)} f(x), \tag{33}$$

is the conjugate density, with $C(t) = \exp\{-K_\psi(\upsilon(t);t)\}$ and $\upsilon(t)$ is the solution of

$$\partial_\upsilon K_\psi(\upsilon;t) = 0, \tag{34}$$

or equivalently $E_{h_{\psi,t}}[\psi(X;t)] = 0$.

4.3 Connections to measure transportation

Notice that the saddlepoint $\upsilon(t)$ is such that

$$\begin{aligned} \partial_\upsilon K_\psi(\upsilon;t) = 0 &\iff E_{h_{\psi,t}}[\psi(X;t)] = 0 \\ &\iff E_f[\psi(X;t)\exp\{\upsilon\psi(X;t)\}] = 0. \end{aligned} \tag{35}$$

In (35), $E_{h_{\psi,t}}$ and E_f represent the expected value computed w.r.t. the conjugate and the original density, respectively. As noticed in Field (1982), p. 678, $h_{\psi,t}(x)$ is the conjugate density of the linearized version of U_n, as obtained in (30). Field (1982) shows that, even if we neglect the remainder in (30), the order of approximation error entailed by the saddlepoint approximation is $O(n^{-1})$. Moreover, $\upsilon(t)$ is such that the conjugate density is centering (up to the remainder) U_n at t, namely

$$E_{h_{\psi,t}}[\psi(X;t)] = 0 \iff E_{h_{\psi,t}}[U_n] = t + O(n^{-1}). \tag{36}$$

Starting from (35), we introduce a modified version of the Kullback-Leibler divergence problem in (3). For fixed $t \in \Theta$, we can prove (see Appendix A) that $h_{\psi,t}(x)$ in (33) is a solution of the following information-theoretic problem:

$$\min_{g \in \mathcal{G}} \left\{ \int_{\mathcal{X}} g(x) \log \frac{g(x)}{f(x)} dx \right\}, \text{ s.t. } g(x) \geq 0, \textstyle\int_{\mathcal{X}} g(x)dx = 1, E_g[\psi(X;t)] = 0, \tag{37}$$

where $\mathcal{G}$ contains all the densities having support $\mathcal{X}$ and such that ψ has finite second moment. Therefore, the saddlepoint solution to (35) yields an optimal property in terms of information theory: this illustrates the link between saddlepoint and information theory as discussed in §2.2.1.

To make the connection with measure transportation, we parallel the argument in §3.2. Let $U_n \sim \mu^{(n)}$: the Esscher's tilting yields a transformation of $\mu^{(n)}$ into a new measure $\nu_t^{(n)}$, which is such that (36) holds and

$$E_{\nu_t^{(n)}}[U_n] = t + O(n^{-1}).$$

We specify $\text{KD}(\mu_\theta^{(n)}, \nu_t^{(n)})$ using $c(x,y) = -xy$: its solution is characterised by the Legendre transform of K_{U_n}, which is not available in closed form. To overcome this problem, we approximate (see (30)) U_n by the mean of IF_ψs, where IF_ψ is as in (31) and we obtain

$$\tilde{K}_{U_n}(\upsilon) = nK_\psi(\upsilon/n;\theta).$$

Then, we set

$$\tilde{f}_n(t) = \frac{n}{2\pi i}\int_{-i\infty}^{i\infty} \exp\{n\tilde{R}_n(\upsilon) - n\upsilon t\}d\upsilon, \tag{38}$$

where $\tilde{R}_n(\upsilon) = \tilde{K}_{U_n}(n\upsilon)/n$. The saddlepoint is defined by

$$\tilde{R}_n'(\upsilon) = t, \tag{39}$$

which is the maximizer υ of

$$\tilde{R}_n^*(t) = \sup_{\upsilon}\{\upsilon t - \tilde{R}_n(\upsilon)\},$$

the Legendre transform of the approximated c.g.f. of U_n. The pair $(\tilde{R}_n, \tilde{R}_n^*)$ characterizes the solution to KD$(\tilde{\mu}^{(n)}, \nu_t^{(n)})$, where $\tilde{\mu}^{(n)}$ is the "measure" related to the "density" $\tilde{f}_n$, as in (38). Along the lines of Proposition 2.1, we have $\mathcal{T} = H_{\psi,t}^{-1} \circ F_X$, where $H_{\psi,t}$ is the c.d.f. of the conjugate density in (33). The map $\mathcal{T}$ yields a change of variable from $X \sim f$ to $Y \sim h_{\psi,t}$.

5. FURTHER RESULTS, DISCUSSION, AND OUTLOOK

5.1 General Legendre transform

The rigorous proof of the connections between measure transportation and saddlepoint approximation for M-estimators remains the object of future research. In particular, we need to understand how the connections are affected by the remainder term in (30) and in (36). We feel that suitable assumptions have to be introduced to fully understand the impact that the use of an approximate c.g.f. has on the optimal transport map (or on the optimal transportation plan). Nevertheless, (39) provides an interesting research direction. Indeed, it is equivalent to solving $\partial_\upsilon K_\psi(\upsilon;t)|_{\upsilon=\upsilon(t)} = 0$, which in turn is equivalent to finding $\sup_\upsilon\{-K_\psi(\upsilon;t)\}$. This leads to the definition of

$$K_\psi^\dagger(t) = \sup_{\upsilon}\{-K_\psi(\upsilon;t)\}, \tag{40}$$

which is a "Legendre-type transform" of K_ψ. Simple calculations show that $\partial_t K_\psi(\upsilon;t) = \tilde{M}(t)\,\upsilon$, where

$$\tilde{M}(t) = E_{h_{\psi,t}}[\partial\psi(X;t)/\partial t]$$

and

$$\frac{d}{dt}K_\psi^\dagger(t) = -\frac{d}{dt}K_\psi(\upsilon(t);t) = -\partial_\upsilon K_\psi(\upsilon(t);t)\,\frac{d}{dt}\upsilon(t) - \partial_t K_\psi(\upsilon(t);t) = -\tilde{M}(t)\,\upsilon(t).$$

Therefore,

$$\frac{d}{dt}K_\psi^\dagger(\partial_\upsilon K_\psi(\upsilon(t);t)) = \frac{d}{dt}K_\psi^\dagger(0) = 0,$$

which generalizes the characterization between the derivatives of a function and its Legendre transform.

The function $K_\psi^\dagger$ has been used in the definition of the saddlepoint test based on M-estimators introduced by Robinson et al. (2003), with test statistic

$$2nK_\psi^\dagger(\hat{\beta}) = 2n\sup_{\upsilon}\{-K_\psi(\upsilon;\hat{\beta})\} = 2n\{-K_\psi(\upsilon(\hat{\beta});\hat{\beta})\}, \tag{41}$$

where $\hat{\beta}$ is the M-estimator defined by the score function ψ. This test statistic is asymptotically χ^2-distributed under the null hypothesis. Our numerical illustration in §3.1.3 provides an example in the case of M-estimation of location via the sample mean, where $\psi(x;t) = x - t$ thus $K_\psi^\dagger(t) \equiv K_X^*(t)$ and it coincides with the Legendre transform. Another important application in quantile regression can be found below. Moreover, Ronchetti and Welsh (1994) apply $K_\psi^\dagger$ (as obtained using the empirical measure of the observations) to define the empirical

saddlepoint approximation of M-estimators; Monti and Ronchetti (1993) apply it to unveil the connection between empirical saddlepoint techniques and empirical likelihood; Gatto (2017) uses $K_\psi^\dagger$ to define two tests on the mean direction of the von Mises–Fisher distribution: the tests perform well even in large dimensional settings, with small sample sizes. Additional applications of this and related tests can be found in Toma and Leoni-Aubin (2010), Toma and Broniatowski (2011), Aeberhard et al. (2017), Holcblat and Sowell (2019), La Vecchia and Ronchetti (2019) and Jiang et al. (2021). Here, we briefly recall the results in Ronchetti and Sabolová (2016).

Example (quantile regression). Let $Y_1, \dots, Y_n$ be observations following

$$Y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + u_i, \qquad i = 1, \dots, n$$

where $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip}) \in \mathbb{R}^p$, $x_{i1} \equiv 1$, $\boldsymbol{\beta} \in \mathbb{R}^p$, $u_i \sim G$ with density g, where bold symbols denote vectors and $^\top$ denotes vector transposition. The regression quantile estimator for $\boldsymbol{\beta}$ (see Koenker and Bassett (1978), Koenker (2005)) is

$$\hat{\boldsymbol{\beta}}_\alpha = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^n \rho_\alpha(Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}),$$

where $\rho_\alpha(u) = |u| \{(1-\alpha)\, \mathrm{I}[u < 0] + \alpha\, \mathrm{I}[u > 0]\}$. It is an M-estimator defined by the score function

$$\psi(y; \boldsymbol{\beta}) = \psi_\alpha(y - \boldsymbol{x}^\top \boldsymbol{\beta}),$$

where $\psi_\alpha(u) = \alpha I[u > 0] - (1-\alpha) I[u < 0] = \alpha - I[u < 0]$. The estimator $\hat{\boldsymbol{\beta}}_\alpha$ is consistent for $\boldsymbol{\beta}_\alpha = (\beta_1 + G^{-1}(\alpha), \beta_2, \dots, \beta_p)$.

Ronchetti and Sabolová (2016) define a saddlepoint test for the regression quantile estimator. To derive K_ψ and $K_\psi^\dagger$, the authors assume for convenience that $(Y_i, \mathbf{x}_i)$ are i.i.d. with density $g(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}) k(\mathbf{x}_i)$, where $k(\cdot)$ is the density of $\mathbf{x}_i$. Then, they obtain:

$$K_\psi(\boldsymbol{\upsilon}; \boldsymbol{\beta}_\alpha) = \log E e^{\boldsymbol{\upsilon} \psi(Y_i; \boldsymbol{\beta}_\alpha) \mathbf{x}_i} = \log E e^{\boldsymbol{\upsilon}^\top \mathbf{x}_i (\alpha - \mathrm{I}[Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_\alpha < 0])}$$

$$= \log \int \left\{ e^{\alpha \boldsymbol{\upsilon}^\top \mathbf{x}_i} k(\mathbf{x}_i) \left[e^{-\boldsymbol{\upsilon}^\top \mathbf{x}_i} G\left(\mathbf{x}_i^\top (\boldsymbol{\beta}_\alpha - \boldsymbol{\beta})\right) + 1 - G\left(\mathbf{x}_i^\top (\boldsymbol{\beta}_\alpha - \boldsymbol{\beta})\right) \right] \right\} \mathrm{d}\mathbf{x}_i$$

and putting to zero the derivative with respect to $\boldsymbol{\upsilon}$, they obtain that the saddlepoint must satisfy

$$\boldsymbol{\upsilon}(\hat{\boldsymbol{\beta}}_\alpha)^\top \mathbf{x}_i = -\log \left\{ \frac{\alpha}{1-\alpha} \frac{1 - G\left(\mathbf{x}_i^\top (\hat{\boldsymbol{\beta}}_\alpha - \boldsymbol{\beta})\right)}{G\left(\mathbf{x}_i^\top (\hat{\boldsymbol{\beta}}_\alpha - \boldsymbol{\beta})\right)} \right\}, \qquad i = 1, \dots, n.$$

Using the last two equations in (41), the saddlepoint test statistic for testing the null hypothesis $\boldsymbol{\beta}_\alpha = \boldsymbol{\beta}_{\alpha 0}$ is

$$2n K_\psi^\dagger(\hat{\boldsymbol{\beta}}_\alpha) = -2n \log E_{\mathbf{x}} \left[\left(\frac{G\left(\mathbf{x}^\top (\hat{\boldsymbol{\beta}}_\alpha - \boldsymbol{\beta}_0)\right)}{\alpha} \right)^\alpha \left(\frac{1 - G\left(\mathbf{x}^\top (\hat{\boldsymbol{\beta}}_\alpha - \boldsymbol{\beta}_0)\right)}{1-\alpha} \right)^{1-\alpha} \right],$$

where $\boldsymbol{\beta}_0$ is the regression parameter corresponding to $\boldsymbol{\beta}_{\alpha 0}$, i.e. $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_{\alpha 0} - (G^{-1}(\alpha), 0, \dots, 0)^\top$. The expectation over $\mathbf{x}$ can be estimated by the average over the observed $\mathbf{x}_i$'s. As an illustration, in Table 1 we compare the exact quantiles $(0.9, 0.95, 0.99)$ obtained by simulation with the χ^2_6 quantiles of the distribution under the null hypothesis for four different tests, i.e. the saddlepoint test, the Wald test, the likelihood-ratio test, and a test referred to *rank-Koenker* as implemented in `R` in the package `quantreg` using the command `rq(..., se = "rank")`, see R Core Team (2013). Under the null hypothesis, these tests have asymptotically a χ^2_6 distribution. We see that the Wald and the likelihood-ratio tests are inaccurate even under normality. The saddlepoint test and the rank-Koenker test are the most reliable

	$n=21$			$n=51$		
	0.9	0.95	0.99	0.9	0.95	0.99
$\mathcal{N}(0,1)$						
Sad	0.9424	0.9710	0.9875	0.9510	0.9808	0.9975
Wald	0.6122	0.6725	0.7641	0.7208	0.7768	0.8551
LR	0.7283	0.8062	0.9027	0.8090	0.8742	0.9463
rank-Koenker	0.9709	0.9939	0.9999	0.9244	0.9666	0.9961

TABLE 1
$\mathcal{H}_0 : \boldsymbol{\beta}_\alpha = (3 + G^{-1}_{\mathcal{N}(0,1)}(\alpha), 1, 2, 3, 4, 5)$, $\alpha = 0.25$, *N = 50000; from Ronchetti and Sabolová (2016)*

across distributions, even for $n = 21$.

The example and the papers mentioned above illustrate that the use of $K^\dagger_\psi$ is well-established in the statistical literature. These results can represent the starting point to study the mathematical properties of $K^\dagger_\psi$. This might offer new perspectives in convex analysis and/or measure transportation theory. A related challenge is the study of nonlinear estimating function ψ: we conjecture that $K^\dagger_\psi$ is a c-transform as in (9), where c may depend on t via the constraint $E_{h_{\psi,t}}[\psi(X;t)] = 0$. The validation of this conjecture and its possible implications in mathematical statistics require further investigations.

5.2 Geodesics, saddlepoints and geometry

In Section 2.2.2, we discuss the notion of geodesic. We believe that this notion introduces an important change of perspective in the literature on saddlepoint approximations. The usual information theoretic approach to the derivation of $v(t)$ does not consider (by the very nature of the variational approach adopted to solve the KL minimization problem) the aspects related to the action (and/or to the motion) needed to tilt μ into ν_t. The links to the optimal transportation theory discussed in this paper put under the spot light the change of variable yielded by $\mathcal{T}$. This shifts the focus to the geodesic in (13) and it offers new insights for the possibility of connecting the theory of saddlepoint approximations to the differential geometry on Riemannian manifolds, using a novel geometric approach which is simpler than the one related to the use of the KL divergence. We illustrate these points in the next example.

Example (univariate Gaussian). In the optimal transportation literature, when working with the quadratic cost, the Gaussian case deserves special considerations, since many quantities are available in closed-form. Thus, we consider $X \sim \mu$, where μ is a univariate Gaussian with mean m_1 and standard deviation σ_1. The c.g.f. $K_X(v) = vm_1 + v^2\sigma_1^2/2$. For $t \in \mathbb{R}$, the saddlepoint solves $K'_X(v) = t$ and $v(t) = \sigma_1^{-2}(t - m_1)$. Elementary algebra yields

$$h_t(x) = \frac{1}{2\pi\sigma_1^2} \exp\left\{-\frac{(x-t)^2}{2\sigma_1^2}\right\},$$

which is the p.d.f. of the Gaussian random variable $Y \sim \nu_t$, where $E_{h_t}[Y] = t$, $V_{h_t}(Y) = \sigma_1^2$ and $Y = \mathcal{T}(X)$, with $\mathcal{T} = H_t^{-1} \circ F_X$ (here, H_t is the c.d.f. of Gaussian with mean t and variance σ_1^2). One can prove (see Peyré and Cuturi (2019), p. 33) that $\mathcal{T} : x \mapsto x - m_1 + t$, which can be rewritten in terms of the Legendre transform of K_X evaluated at the saddlepoint as $\mathcal{T} : x \mapsto x - m_1 + K'_X(v(t))$. This optimal transportation mapping induces a shift of the original Gaussian random variable X and does not affect the variance. Since h_t is obtained via the Esscher's tilting, a natural research question is related to the comparison between the geometry of the KL divergence and the geometry of W_2 over the manifold of Gaussians. In particular, we focus on the geodesics in the two different geometries, proceeding as in Remark 8.2 of Peyré and Cuturi (2019). To begin with, we consider two generic univariate Gaussians $\mathcal{N}(m_1, \sigma_1^2)$ and $\mathcal{N}(t, \sigma_2^2)$. As in Carter et al. (2007) and in Costa et al. (2015), we notice that, when $t = m_1 + \delta_1$ and $\sigma_2 = \sigma_1 + \delta_2$ for infinitesimal (δ_1, δ_2), then the

KL divergence can be approximated by the hyperbolic distance in the Poincaré half-plane. The latter induces a hyperbolic geometry, where the geodesics are half circles. Thus, the geodesic between $\mathcal{N}(m_1, \sigma_1^2)$ and $\mathcal{N}(t, \sigma_2^2)$ contains Gaussians which do not have a constant standard deviation. This consideration remains valid also in the case of the Esscher's tilted Gaussian, where we move from μ to ν_t, with $m_1 \neq t$ and $\sigma_1 \equiv \sigma_2$. Therefore, over the space of Gaussian parameters (m, σ), the geodesic connecting μ to ν_t is not the horizontal line connecting (m_1, σ_1) to (t, σ_1). Rather, the geodesic is a half-circle, along which the standard deviation is not constant. We compare this with the geometry associated to optimal transport with quadratic cost, for which we have $W_2^2(\mu, \nu_t) = (t - m)^2$. The expression illustrates that the squared 2-Wasserstein distance is related to the (squared) Euclidean distance between the mean parameters. In this simpler geometry, the standard deviation of all the Gaussian densities located on the geodesic (13) remains constant. In Figure 6, we provide a graphical sketch of these concepts.

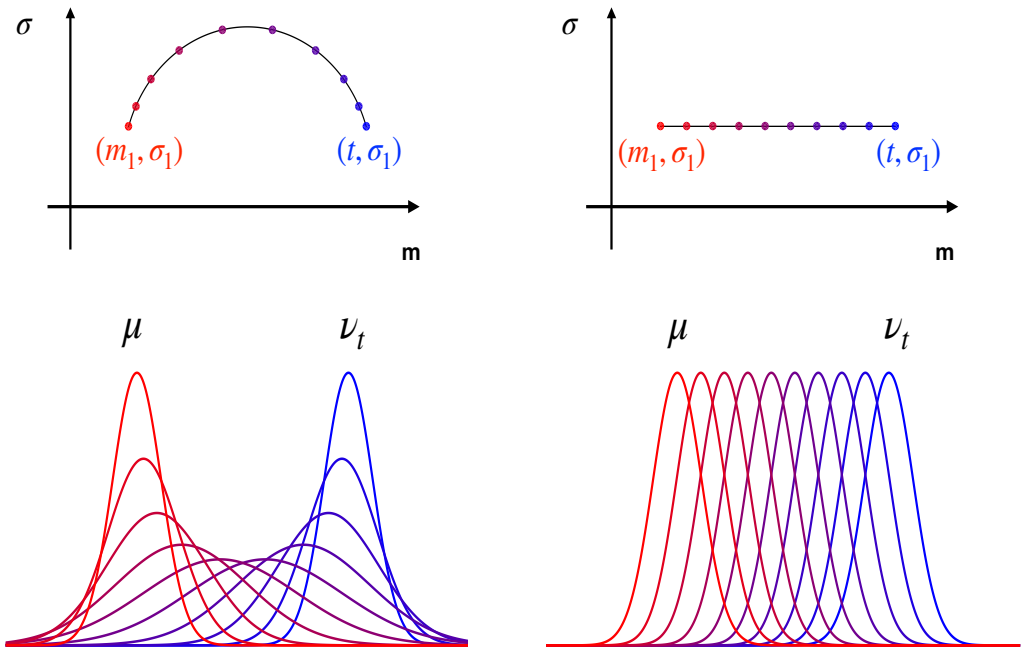

FIG 6. *Comparisons of interpolation between Gaussians (the original μ and the tilted ν_t) using KL hyperbolic geometry (left panels) and optimal transport geometry (right panels). Plot adapted from Peyré and Cuturi (2019).*

The above construction allows us to derive a link between the squared 2-Wasserstein distance and $K_X^*(t) = (t - m)^2/2\sigma^2$. Simple algebra yields:

$$W_2^2(\mu, \nu_t) = 2\sigma^2 K_X^*(t). \tag{42}$$

We think that (42) offers the possibility of connecting the saddlpeoint test of Robinson et al. (2003) (and its higher-order properties) to a test based on W_2^2. To see this, let us consider a Gaussian random variable $\mathcal{N}(m, \sigma_1^2)$ and define the null $\mathcal{H}_0 : m = m_0$ vs $\mathcal{H}_1 : m \neq m_0$, with known σ_1^2. As in the example of §3.1.3, we may think of defining a test via $2nK_X^*(\bar{X}_n)$, which has a χ_1^2 distribution. Eq. (42) implies that this is equivalent to define a test using $W_2^2(\mu_0, \nu_{\bar{X}_n})$, which gives a novel insight for the test of Robinson et al. (2003). Indeed, from (42) we have that $K_X^*(\bar{X}_n) \propto W_2^2(\mu_0, \nu_{\bar{X}_n})$. Thus, the test based on the Legendre transform of K_X has an interpretation in terms of measure transportation: it is related to the quadratic cost of transporting the probability mass in μ_0 (under the null) onto $\nu_{\bar{X}_n}$ (the measure related to the finite sample distribution of $\bar{X}_n$, as obtained using the saddlepoint approximation). We

feel this novel insight deserves further investigations. For instance, a road map can start form the understanding if (42) can be connected to the theory of tests developed in Del Barrio et al. (1999). Then, one may go beyond the Gaussian location case and work on M-estimators in the exponential family (e.g. for frequency domain time series analysis, along the lines as La Vecchia and Ronchetti (2019)), with the aim of studying the relationship between W_2^2 and $K_\psi^\dagger$.

Other connections to geometry can be devised, exploring some links to Amari's research. To this end, let $\mathbb{M}$ be a Riemaniann manifold equipped with a Riemannian metric, which defines an inner product on the tangent space $T_x\mathbb{M}$. If $x \in \mathbb{M}$ and $v(x) \in T_x\mathbb{M}$ are given, the exponential map $\exp_x v(x)$ is defined as $\gamma(1)$, where the parameterized curve $\gamma : [0,1] \to \mathbb{M}$ is the unique constant speed geodesic starting at $\gamma(0) = x$ with velocity $\dot\gamma(0) = v(0)$. As mentioned e.g. in Villani (2009), p. 364, the Jacobian determinant of the exponential map is related to the curvature of the manifold. Now, if we consider the manifold $\mathcal{P}(X)$ and the related Wasserstein space, the geodesic is as in (13): we have that $\gamma(0) = \mu$ and $\gamma(1) = \nu_t$. Thus, we foresee a connection between the curvature and (K_X, K_X^*). To our knowledge this connection is unexplored and its study may contribute to the literature on the interplay between differential geometry, higher-order asymptotics, and inference; see Kass (1989) with comments by Amari (1989) and Reid and Fraser (1989), and Amari (2016). The potential research outcome(s) related to the study of the link(s) between saddlepoint and information geometry (which studies the properties of a manifold of probability distributions) may be useful for various applications in statistics, time series analysis, machine learning, signal processing, and optimization. We refer to, Amari (2016), Ch. 11-13, and for recent developments see Amari et al. (2018). Amari's construction hinges on the Legendre transform (see Th. 4 in Amari et al. (2018)) and on the exponential family (see Amari (2016) Ch. 2). Our results unveil the connection between the optimal transportation map and the Legendre transform of the c.g.f. in the case of quadratic cost: one may think of combining these results. This is a challenging and almost unexplored research area.

5.3 Upper bounds for the Wasserstein distance and large deviations

Let $\{X_i\}$ for $i = 1, .., n$ be iid random variables defined on the same probability space $(\Omega, \mathcal{F}, F)$, where F has support $\mathcal{X} \subseteq \mathbb{R}$, for $\mathcal{X} = (a, b)$, where, if desired, $a = -\infty$ or $b = \infty$, or both. Assume that $E[X_i] = 0$, $V[X_i] = 1$. Let us further define the random variables $S_n := \sqrt{n}\bar X_n$ and Z_{sad} whose CDF is P_{sad}, the integral of the saddlepoint density p_{sad}. Moreover, let P_n be the exact CDF of S_n with derivative p_n. Then, under the same assumptions as in Th. 7.1, Th. 7.2, and Th. 7.3 of Daniels (1954)–which are needed to guarantee a uniform relative error for the saddleponit approximation of the distribution of the mean–we have

PROPOSITION 5.1. *(i) The Wasserstein distance between P_n and Φ is*

$$W_1(P_n, \Phi) = O(n^{-1/2}), \tag{43}$$

where Φ denotes the CDF of the standard Normal distribution.

(ii) The Wasserstein distance between P_n and P_{sad} is

$$W_1(P_n, P_{sad}) = O(n^{-1}). \tag{44}$$

The proof is available in the Appendix D. This proposition shows that P_{sad} is closer to the exact P_n than the normal Φ in Wasserstein distance, providing a transfer of a known result in statistics to measure transportation theory. We refer to Rio (2009) for other upper bounds on Wasserstein distances.

Beside this result, further links could be investigated, identifying new connections between concentration inequalities, Wasserstein distance and the theory of large deviations—as derived using the Legendre transform; see e.g. Chernoff (1952) and Bahadur (1971). These

connections are more involved than those mentioned in this paper and they require an extensive, separate investigation. To provide a starting point, we refer to Arcones (2006) (where the author studies large deviations in the setting of M-estimators) and to Léonard (2007) (where the author shows that the optimal transport cost related to the 2-Wasserstein distance can play the role of a rate function in large deviation theory).

5.4 Possible research directions in machine learning

Information theory (via entropy and Kullback-Leibler minimization) plays a pivotal role in the literature on machine learning, e.g. for patterns recognition; see Grenander et al. (2007) and Murphy (2012), among others. Moreover, measure transportation is advocated by machine learners for the analysis of large data sets; see Cuturi (2013) and Peyré and Cuturi (2019). In contrast, the use of saddlepoint techniques has been overlooked by the machine learning community and more computationally intensive methods (typically, the bootstrap) are preferred; see e.g. Murphy (2012) and James et al. (2013). We mention two possible research directions, where the saddlepoint techniques can be of help: (i) the need for speeding up machine learning algorithms, replacing resampling methods by saddlepoint approximations, without compromising accuracy; (ii) the problem of computing the optimal transportation map in the setting of multivariate random variables, which arises e.g. in image processing (color distribution transfer) and computer graphics (texture mixing).

For possible developments related to (i), we refer to Davison and Hinkley (1988) and Ronchetti and Welsh (1994). As far as (ii) is concerned, we refer e.g. to Pitié et al. (2007) who propose to break down the problem of computing a monotone transportation map between 2D (or N-D) variables into a succession of one-dimensional distribution transfer problems. The resulting map is easy-to-compute since each one-dimensional transportation map can be obtained in closed form via the probability integral transform. The convergence of the resulting algorithm is available only for Gaussian random variables; we refer to Santambrogio (2015), p. 81. With this regard, some techniques developed in the saddlepoint approximation literature for Gaussian multivariate random variables can be connected with the literature on multivariate transport maps. The next example provides an illustration.

Example (bivariate Gaussian). The definition of saddlepoint techniques for multivariate random variables goes through the same calculation steps as in the univariate setting; see e.g Kolassa (2006) Ch. 6 (see also Kolassa and Li (2010)) and McCullagh (2018). The saddlepoint equation is still defined via the gradient of the c.g.f. of $\boldsymbol{X} \sim \mu$ and the conjugate density is derived using the Legendre transform; see Kolassa (2006), p. 119 eq. (137). A particularly simple case is obtained when $\boldsymbol{X}$ has a bivariate Normal distribution $\mathcal{N}(\boldsymbol{m}, \boldsymbol{\Sigma})$, with $\boldsymbol{m} \in \mathbb{R}^2$ and $\boldsymbol{\Sigma}$ is a (2×2)-matrix. The c.g.f. is $K_{\boldsymbol{X}}(\boldsymbol{v}) = \boldsymbol{m}^\top \boldsymbol{v} + (1/2)\boldsymbol{v}^\top \boldsymbol{\Sigma} \boldsymbol{v}$, for $\boldsymbol{v} \in \mathbb{R}^2$. Thus, we find the saddlepoint $\boldsymbol{v}(\boldsymbol{t})$, for $\boldsymbol{t} := (t_1, t_2)^\top \in \mathbb{R}^2$, by solving $K'_{\boldsymbol{X}}(\boldsymbol{v}) = \boldsymbol{t}$. This yields $\boldsymbol{v}(\boldsymbol{t}) = \boldsymbol{\Sigma}^{-1}(\boldsymbol{t} - \boldsymbol{m})$ and $K''_{\boldsymbol{X}}(\boldsymbol{v}(\boldsymbol{t})) = \boldsymbol{\Sigma}$. The exponential tilting still depicts a change of variable $\boldsymbol{X} \mapsto \boldsymbol{Y}$, where $\boldsymbol{Y} \sim \nu_{\boldsymbol{t}}$, with $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{t}, \boldsymbol{\Sigma})$. The optimal transport mapping (a deterministic coupling) is $\boldsymbol{\mathcal{T}} : \boldsymbol{x} \mapsto \boldsymbol{x} + K'_{\boldsymbol{X}}(\boldsymbol{v}(\boldsymbol{t}))$; see e.g. Peyré and Cuturi (2019), p. 33. In Figure 7 we display the transformation for a bivariate Gaussian $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{\Sigma})$, with

$$\boldsymbol{m} = (0, 0)^\top \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \tag{45}$$

and target mean $\boldsymbol{t} = (-1, -1)^\top$. The interpretation of the plot remains the same as in Figure 3; in particular, the segments in the top panel illustrate the geodesics and the mass transport happens at constant speed. The bottom panels display the displacement interpolation, where each Gaussian at s has a different mean vector $\boldsymbol{m}$, but the same variance as the other Gaussians along the geodesic (13), which connects μ (at $s = 0$) and $\nu_{\boldsymbol{t}}$ (at $s = 1$).

One may think that the properties discussed in the previous example are limited to the Gaussian case, where the availability in closed form of the c.g.f. can be fully exploited. Nevertheless, from the literature about saddlepoint approximations we know that the knowledge

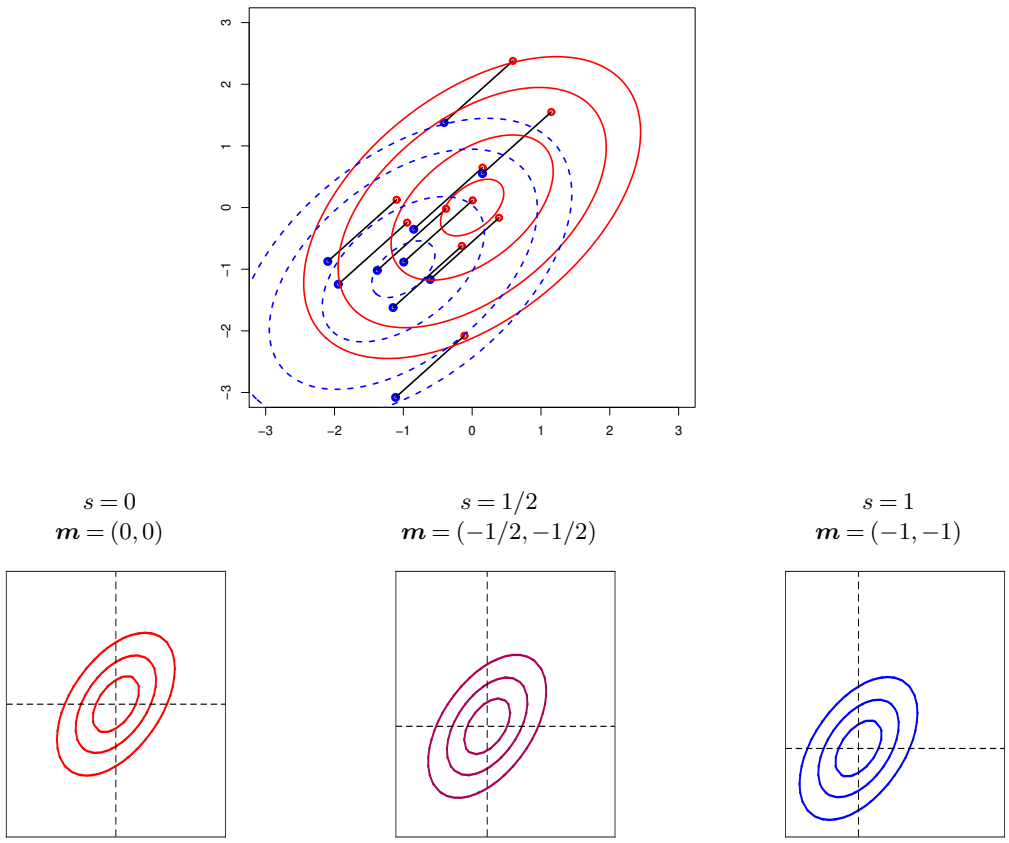

FIG 7. *Optimal transportation problem for a bivariate Gaussian random variables* $\boldsymbol{X} \sim \mu$ *and* $\boldsymbol{Y} \sim \nu_{\boldsymbol{t}}$. *Top panels: red dots represent 10 random observations drawn from an bivariate Gaussian with zero mean vector and variance matrix as in (45); the blue filled dots are for the observations of* $\boldsymbol{Y}$, *as obtained via the optimal coupling induced by the map* $\boldsymbol{\mathcal{T}}$, *where the transformed random variable* $\boldsymbol{Y}$ *has distribution* $\mathcal{N}(\boldsymbol{t}, \boldsymbol{\Sigma})$, *for* $\boldsymbol{t} = (-1, -1)^\top$ *The segments connecting red and blue dots represent the displacement* $\boldsymbol{\mathcal{T}}(\boldsymbol{x}) - \boldsymbol{x}$. *The dashed blue and continuous red ellipse correspond to some level curves of the two Gaussians. Bottom panels: displacement interpolation for* $s = 0, 1/2, 1$; *in each plot the dashed lines indicate the mean* $\boldsymbol{m}$ *of the distribution.*

of $K_{\boldsymbol{X}}$ is not a limitation. Indeed, one may think of approximating $K_{\boldsymbol{X}}$ using either the empirical c.g.f. (see Ronchetti and Welsh (1994) for the regression problem with multivariate parameter and Fasiolo et al. (2018) for multivariate random variables) or resorting on some approximations of the third- and/or forth-order cumulants (as in Easton and Ronchetti (1986)).

5.5 Change of variable and related equations

Measure transportation is essentially a change of variable. In §2.1 we illustrate that the conjugate density is obtained via the Jacobian formula in (14). This connection unveils a link between the theory on saddlepoint approximations and the theory of nonlinear (elliptical) partial differential equations (PDEs). Specifically, for $\mathcal{S} \subseteq \mathbb{R}$, under suitable (smoothness) conditions, the map $\mathcal{T}$ satisfies the PDE

$$f_Y(\mathcal{T}(x)) = f_X(x) \left(\left| \frac{\partial x}{\partial \mathcal{T}(x)} \right| \right), \tag{46}$$

which is the Jacobian formula (14) expressing the conjugate density h_t. This gives an additional insight. From §2.3, we know that there is a link between K'_X and the solution to (46). This creates a link to the PDEs theory. To our knowledge, this link is unexplored and it may yield interesting implications. Here we mention some of them.

For multivariate random variables $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^d$, with $d \geq 2$, the optimal transportation map for the quadratic cost is the solution to the Monge-Ampère equation; see Villani (2009), Santambrogio (2015) and De Philippis and Figalli (2014) for a recent review. We conjecture that this link between the saddlepoint equation and PDEs can be explored to relate $K'_{\boldsymbol{X}}$ to the notion of Gaussian curvature in Riemanian geometry—or more generally to the Ricci's curvature; see Villani (2009) Ch. 14, where the Jacobian determinant appearing in (46) plays a pivotal role. Similar theoretical developments can be conjectured for M-estimators of the multivariate parameter $\boldsymbol{\theta}$, as well. To elaborate further, the example about quantile regression, discussed in §5.1, illustrates that in some cases one may derive $h_{\boldsymbol{\psi},\boldsymbol{t}}$ in a closed form. For $c(x, y)$ satisfying the twist condition, ϕ is such that

$$\nabla\phi(x) - \nabla c(x, y) = 0. \tag{47}$$

Proceeding heuristically, we may differentiate w.r.t. x both sides of (47):

$$\nabla^2\phi(x) - \nabla^2_{xx}c(x, y) - \nabla^2_{xy}c(x, \mathcal{T}(x))\nabla\mathcal{T}(x) = 0.$$

Then, selecting the usual cost $c(x, y) = -xy$, we obtain $\nabla^2\phi(x) = \nabla\mathcal{T}(x)$. Taking the absolute value on both sides we have $|\nabla^2\phi(x)| = |\nabla\mathcal{T}(x)|$, and using $|\nabla\mathcal{T}(x)| = f(x)/h_{\boldsymbol{\psi},\boldsymbol{t}}(\mathcal{T}(x))$ we obtain the following PDE:

$$|\nabla^2\phi(x)| = f(x)/h_{\boldsymbol{\psi},\boldsymbol{t}}(\mathcal{T}(x)), \tag{48}$$

which is a special case of Monge-Ampére equation. For $h_{\boldsymbol{\psi},\boldsymbol{t}}$ is as in (33), we have

$$|\nabla^2\phi(x)|e^{\boldsymbol{\upsilon}(\boldsymbol{t})^\top\boldsymbol{\psi}[\nabla\phi(x);\boldsymbol{t}]-K_{\boldsymbol{\psi}}(\boldsymbol{\upsilon}(\boldsymbol{t});\boldsymbol{t})} - 1 = 0, \tag{49}$$

which defines a PDE for the Kantorovich potential and links $K'_{\boldsymbol{\psi}}$ to $\phi(x)$. Moreover, when $h_{\boldsymbol{\psi},\boldsymbol{t}}$ has the analytic expression (as in the regression case, see Field and Ronchetti (1990), p. 72), one may work on the Jacobian equation, which implies that $\mathcal{T}$ satisfies

$$h_{\boldsymbol{\psi},\boldsymbol{t}}[\mathcal{T}(x)] = f(x)e^{\boldsymbol{\upsilon}(\boldsymbol{t})^\top\boldsymbol{\psi}[\mathcal{T}(x);\boldsymbol{t}]-K_{\boldsymbol{\psi}}(\boldsymbol{\upsilon}(\boldsymbol{t});\boldsymbol{t})}. \tag{50}$$

Similarly to (46), we propose to interpret (50) as an equation in $\mathcal{T}$, in which we link $K'_{\boldsymbol{\psi}}$ to $\mathcal{T}$. We hope that (a) studying this link, (b) providing the suitable conditions under which our heuristic derivation holds and (c) investigating other connections between saddlepoint approximations and measure tranportation and/or inferential geometry can all represent interesting theoretical research topics. Moreover, the availability of analytic solution $\mathcal{T} = H^{-1}_{\boldsymbol{\psi},\boldsymbol{t}} \circ F_X$ can be of help to check the efficacy of numerical techniques applied in the literature on differential equations for solving (46)-(50).

APPENDIX

APPENDIX A: PROOF OF THE KULLBACK-LEIBLER OPTIMIZATION PROBLEMS

We show that for a fixed t, the conjugate density h_t is the solution to the following information theoretic problem:

$$\min_{g\in\mathcal{G}} \int_{\mathcal{X}} g(x)\log\frac{g(x)}{f_X(x)}dx, \text{ s.t. } g(x)\geq 0, \quad \int_{\mathcal{X}} g(x)dx = 1, \quad \int_{\mathcal{X}} xg(x)dx = t.$$

The Lagrangian of this problem is

$$L[g(x)] = g(x)\log\tfrac{g(x)}{f_X(x)} - \lambda_1 g(x) - \lambda_2(x-t)g(x),$$

and the Euler-Lagrange equation

$$\frac{\partial L}{\partial g} - \frac{d}{dx}\frac{\partial L}{\partial g'} = 0.$$

Since $\frac{\partial L}{\partial g'} = 0$ we have

$$0 = \tfrac{\partial L}{\partial g} = \tfrac{f_X(x)}{g(x)}\tfrac{1}{f_X(x)}g(x) + \log\tfrac{g(x)}{f_X(x)} - \lambda_1 - \lambda_2(x-t),$$

i.e.

$$\log\tfrac{g(x)}{f_X(x)} = \lambda_1 - 1 + \lambda_2(x-t).$$

If we let $c_1 = e^{\lambda_1 - 1}$, the last expression implies that

$$g(x) = c_1 e^{\lambda_2(x-t)} f_X(x).$$

Notice a slight abuse of notation: in fact, the function g depends on x and t, but we drop the last argument and we write $g(x)$.

We are now left to find the values of c_1 and λ_2, based on the initial conditions. Namely,

$$0 = E_g[(X-t)] = \int (x-t)c_1 e^{\lambda_2(x-t)} f(x)dx$$

which implies that

$$t = \frac{\int xe^{\lambda_2 x} f(x)dx}{\int e^{\lambda_2 x} f_X(x)dx} = K'(\lambda_2).$$

Hence, $\lambda_2 = \upsilon(t)$. In addition

$$1 = \int g(x)dx = c_1 \int e^{\lambda_2(x-t)} f(x)dx$$

implies that

$$1/c_1 = e^{-t\upsilon(t)} \int e^{\upsilon(t)x} f_X(x)dx = e^{-t\upsilon(t)+K(\upsilon(t))} = 1/C(t).$$

Hence

$$g(x) = C(t)\exp\{\lambda_2(x-t)\}f(x) = C(t)\exp\{\upsilon(t)(x-t)\}f(x).$$

Since $C(t)\geq, 0$ the condition $g(x)\geq 0$ is indeed satisfied and it follows that $g(x)\equiv h_t(x)$. Thus, the conjugate density h_t is the solution to the Kullback-Leibler optimization problem.

The same arguments apply to prove the more general optimization problem (37), where the solution is given by (33).

APPENDIX B: SKETCH OF DERIVATION OF KANTOROVICH'S DUAL PROBLEM

Let us introduce two integrable functions $\varphi \in L_1(\nu_t)$ and $\phi \in L_1(\nu_t)$ and express the constraint $\gamma \in \Gamma(\mu, \nu_t)$ in the following way:

$$\ell_\Gamma(\gamma) := \sup_{\phi,\varphi} \left(\int_{\mathcal{X}} \varphi(y) d\nu_t(y) - \int_{\mathcal{X}} \phi(x) d\mu(x) - \int_{\mathcal{X}\times\mathcal{X}} (\varphi(y) - \phi(x)) d\gamma(x,y) \right),$$

which is equal to 0 if $\gamma \in \Gamma(\mu, \nu_t)$ (the coupling has the right marginals) and it is equal to $+\infty$ otherwise. For a cost function $c(x,y)$ we consider the problem

$$\inf_{\gamma \in \mathcal{P}_+(\mathcal{X}\times\mathcal{X})} \int_{\mathcal{X}\times\mathcal{X}} c(x,y) d\gamma(x,y) + \ell_\Gamma(\gamma),$$

where the infimum is taken over all joint positive measures, not necessarily with marginals μ and ν_t. Then, omitting the arguments for the ease-of-notation, we rewrite

$$\inf_{\gamma \in \mathcal{P}_+(\mathcal{X}\times\mathcal{X})} \sup_{\phi,\varphi} \int_{\mathcal{X}\times\mathcal{X}} c d\gamma + \int_{\mathcal{X}} \varphi d\nu_t - \int_{\mathcal{X}} \phi d\mu - \int_{\mathcal{X}\times\mathcal{X}} (\varphi - \phi) d\gamma$$

$$= \inf_{\gamma \in \mathcal{P}_+(\mathcal{X}\times\mathcal{X})} \sup_{\phi,\varphi} \int_{\mathcal{X}\times\mathcal{X}} c d\gamma - \int_{\mathcal{X}\times\mathcal{X}} (\varphi - \phi) d\gamma + \int_{\mathcal{X}} \varphi d\nu_t - \int_{\mathcal{X}} \phi d\mu.$$

Under suitable assumptions (see e.g. Santambrogio (2015), p. 9, and Rockafellar (2015)), we may invert the infimum and supremum, so the problem becomes

$$\sup_{\phi,\varphi} \inf_{\gamma \in \mathcal{P}_+(\mathcal{X}\times\mathcal{X})} \int_{\mathcal{X}\times\mathcal{X}} (c - (\varphi - \phi)) d\gamma + \int_{\mathcal{X}} \varphi d\nu_t - \int_{\mathcal{X}} \phi d\mu.$$

The

$$\inf_{\gamma \in \mathcal{P}_+(\mathcal{X}\times\mathcal{X})} \int_{\mathcal{X}\times\mathcal{X}} (c - (\varphi - \phi)) d\gamma$$

is zero if $c - (\varphi - \phi) \geq 0$ and minus infinity otherwise. Therefore, the Kantorovich's dual problem is

$$\begin{aligned} \mathrm{KD}(\mu, \nu_t) \;=\; & \sup_{\phi,\varphi} \left(\int_{\mathcal{X}} \varphi(y) d\nu_t(y) - \int_{\mathcal{X}} \phi(x) d\mu(x) \right) \\ & \text{s.t. } \varphi(y) - \phi(x) \leq c(x,y), \quad \forall (x,y), \end{aligned}$$

which is (8).

APPENDIX C: PROOF OF PROPOSITION 2.1

PROOF. Existence and uniqueness of the optimal transport plan γ follows from Th. 2.9 in Santambrogio (2015), which guarantees also that the optimal plan is induced by the monotone map $\mathcal{T} = H_t^{-1} \circ F_X$. This transportation map yields a deterministic coupling (X, Y) of (μ, ν_t). □

APPENDIX D: PROOF OF PROPOSITION 5.1

PROOF. Part (i). We have:

$$W_1(P_n, \Phi) = \int_0^1 |P_n^{-1}(\alpha) - \Phi^{-1}(\alpha)| \, d\alpha. \tag{51}$$

For $\alpha \in (0,1)$, let us denote by s_α the α-th quantile of S_n, which is such that $P_n(s_\alpha) = \alpha$. Similarly, we denote by u_α, the α-th quantile of Z, where $Z \sim \mathcal{N}(0,1)$. Fron the CLT it follows that $S_n \xrightarrow{\mathrm{D}} Z$ and, for $\rho = E[|X_1|^3] < \infty$, the Berry-Esseen theorem implies that $|P_n - \Phi| \leq 3\rho/\sqrt{n}$. Thus, $s_\alpha - u_\alpha = O(n^{-1/2})$. Making use of this result in (51), we have $W_1(P_n, \Phi) = O(n^{-1/2})$.

Part (ii). By the alternative form of the Wasserstein distance (see e.g. Bobkov and Ledoux (2019), Th. 2.10) we have:

$$W_1(P_n, P_{\text{sad}}) = \int_{-\infty}^{\infty} |P_n(x) - P_{\text{sad}}(x)|\, dx. \tag{52}$$

Let f_n denote the exact pdf of $\bar{X}_n$. The Jacobian formula for the transformation of random variables yields the relation between the pdf of $\bar{X}_n$ and the pdf of S_n, namely $p_n(x) = (1/\sqrt{n}) f_n(x/\sqrt{n})$. Similarly, for the saddlepoint density approximations we have

$$p_{\text{sad}}(x) = (1/\sqrt{n}) f_{\text{sad}}(x/\sqrt{n}), \tag{53}$$

where f_{sad} represents the saddlepoint density approximation for $\bar{X}_n$.

Under the assumptions in Daniels (1954, see Th. 7.1 and Th. 7.2), one can prove that

$$\left| \frac{f_{\text{sad}}(x)}{f_n(x)} - 1 \right| \leq O(n^{-1}), \tag{54}$$

where the bound is valid *uniformly* over the whole support, namely also for those values of x close to the boundaries a and/or b. Now, using (53) we have

$$\left| \frac{p_{\text{sad}}(x)}{p_n(x)} - 1 \right| = \left| \frac{f_{\text{sad}}(xn^{-1/2})}{f_n(xn^{-1/2})} - 1 \right| \leq O(n^{-1}), \tag{55}$$

where (54) implies that the bound in (55) holds uniformly. Therefore, we obtain:

$$\begin{aligned} W_1(P_n, P_{\text{sad}}) &= \int_{-\infty}^{0} |P_n(x) - P_{\text{sad}}(x)|\, dx + \int_{0}^{\infty} |P_n(x) - P_{\text{sad}}(x)|\, dx \\ &= \int_{-\infty}^{0} |rel(P_n(x), P_{\text{sad}}(x))| P_n(x)\, dx \\ &+ \int_{0}^{\infty} |rel(1 - P_n(x), 1 - P_{\text{sad}}(x))| (1 - P_n(x))\, dx, \end{aligned} \tag{56}$$

where

$$rel(Q_1(x), Q_2(x)) = \frac{Q_1(x) - Q_2(x)}{Q_1(x)}$$

is the relative error of the approximation $Q_2(x)$ with respect to the exact $Q_1(x)$. Since the relative error in (56) is $O(n^{-1})$ uniformly, we get

$$\begin{aligned} W_1(P_n, P_{\text{sad}}) &= \left\{ \int_{-\infty}^{0} P_n(x)\, dx + \int_{0}^{\infty} (1 - P_n(x))\, dx \right\} O(n^{-1}) \\ &= E[|S_n|]\, O(n^{-1}) = O(n^{-1}). \end{aligned}$$

The same arguments hold for f satisfying the assumptions in Th. 7.3 in Daniels (1954) □

As pointed out by an anonymous referee, the proof holds for any density approximation which has a relative error of order $O(n^{-1})$ on the *whole support* of the random variable. However, we are not aware of any approximation other than the saddlepoint that satisfies this accuracy criterion, at least in the frequentist context.

ACKNOWLEDGEMENT

The authors would like to thank the Editor, the Associate Editor, and four referees for helpful and stimulating comments on the original manuscript. Andrej Ilievski (visiting student at University of Geneva in the Summer 2019) thanks the Office of Science, Technology and Higher Education at the Embassy of Switzerland in Washington, D.C. for the financial support through the ThinkSwiss program. Davide La Vecchia and Elvezio Ronchetti thank Roger Koenker, Marc Hallin, Cesare Miglioli and Alban Moor for their comments on the manuscript. Davide La Vecchia is particularly thankful to Alan Welsh for very helpful and stimulating discussions on saddlepoint approximations and measure transportation theory.

REFERENCES

Aeberhard, W. H., Cantoni, E., and Heritier, S. (2017). Saddlepoint tests for accurate and robust inference on overdispersed count data. *Computational Statistics & Data Analysis*, 107:162–175.

Amari, S. (1989). The geometry of asymptotic inference: Comment. *Statistical Science*, 4(3):220–222.

Amari, S. (2016). *Information geometry and its applications*, volume 194. Springer.

Amari, S., Karakida, R., and Oizumi, M. (2018). Information geometry connecting Wasserstein distance and Kullback-Leibler divergence via the entropy-relaxed transportation problem. *Information Geometry*, 1(1):13–37.

Arcones, M. A. (2006). Large deviations for M-estimators. *Annals of the Institute of Statistical Mathematics*, 58(1):21–52.

Bahadur, R. (1971). *Some Limit Theorems in Statistics*. Soc. Ind. Appl. Math., Philadelphia.

Barndorff-Nielsen, O. and Cox, D. (1989). *Asymptotic Techniques for Use in Statistics*. Chapman and Hall London.

Barndorff-Nielsen, O. and Cox, D. R. (1979). Edgeworth and saddle-point approximations with statistical applications. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(3):279–299.

Bhattacharya, R. N. and Rao, R. R. (1986). *Normal approximation and asymptotic expansions*, volume 64. Siam.

Bobkov, S. and Ledoux, M. (2019). *One-dimensional empirical measures, order statistics, and Kantorovich transport distances*, volume 261. American Mathematical Society.

Brazzale, A., Davison, A. C., and Reid, N. (2007). *Applied Asymptotics: Case Studies in Small-Sample Statistics*. Cambridge University Press.

Carter, K. M., Raich, R., and Hero, A. (2007). Learning on statistical manifolds for clustering and visualization. In *45th Allerton Conference on Communication, Control, and Computing, Monticello, Illinois*.

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on sums of observations. *Annals of Mathematical Statistics*, 23:493–507.

Chernozhukov, V., Galichon, A., Hallin, M., and Henry, M. (2017). Monge–Kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, 45(1):223–256.

Costa, S. I., Santos, S. A., and Strapasson, J. E. (2015). Fisher information distance: a geometrical reading. *Discrete Applied Mathematics*, 197:59–69.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300.

Daniels, H. E. (1954). Saddlepoint approximations in statistics. *Annals of Mathematical Statistics*, 25:631–650.

Davison, A. C. and Hinkley, D. V. (1988). Saddlepoint approximations in resampling methods. *Biometrika*, 75(3):417–431.

De Philippis, G. and Figalli, A. (2014). The Monge–Ampère equation and its link to optimal transportation. *Bulletin of the American Mathematical Society*, 51(4):527–580.

Del Barrio, E., Cuesta-Albertos, J. A., Matrán, C., and Rodríguez-Rodríguez, J. M. (1999). Tests of goodness of fit based on the L2-Wasserstein distance. *The Annals of Statistics*, 27:1230–1239.

del Barrio, E., González-Sanz, A., and Hallin, M. (2020). A note on the regularity of optimal-transport-based center-outward distribution and quantile functions. *Journal of Multivariate Analysis*, 180:104671.

Easton, G. S. and Ronchetti, E. (1986). General saddlepoint approximations with applications to L statistics. *Journal of the American Statistical Association*, 81(394):420–430.

Esscher, F. (1932). On the probability function in the collective theory of risk. *Skand. Aktuarie Tidskr.*, 15:175–195.

Fasiolo, M., Wood, S. N., Hartig, F., Bravington, M. V., et al. (2018). An extended empirical saddlepoint approximation for intractable likelihoods. *Electronic Journal of Statistics*, 12(1):1544–1578.

Field, C. (1982). Small sample asymptotic expansions for multivariate M-estimates. *The Annals of Statistics*, 10:672–689.

Field, C. A. and Ronchetti, E. (1990). *Small Sample Asymptotics*. IMS, Lecture notes-monograph series.

Galichon, A. (2016). *Optimal Transport Methods in Economics*. Princeton University Press.

Galichon, A. (2017). A survey of some recent applications of optimal transport methods to econometrics. *Econometrics Journal*, 20:C1–C11.

Gatto, R. (2017). Multivariate saddlepoint tests on the mean direction of the von Mises–Fisher distribution. *Metrika*, 80(6-8):733–747.

Goutis, C. and Casella, G. (1999). Explaining the saddlepoint approximation. *The American Statistician*, 53(3):216–224.

Grenander, U., Miller, M. I., Miller, M., et al. (2007). *Pattern Theory: From Representation to Inference*. Oxford University Press.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer Science & Business Media.

Hallin, M. (2017). On distribution and quantile functions, ranks and signs in $\mathbb{R}^d$: a measure transportation approach. Available at ideas.repec.org/p/eca/wpaper/2013-258262.html.

Hallin, M., del Barrio, E., Cuesta-Albertos, J., and Matran, C. (2020a). Center-outward distribution and quantile functions, ranks, and signs in dimension d: a measure transportation approach. *The Annals of Statistics*, In press.

Hallin, M., La Vecchia, D., and Liu, H. (2020b). Center-outward R-estimation for semiparametric VARMA models. *Journal of the American Statistical Association*, DOI: 10.1080/01621459.2020.1832501.

Hallin, M., La Vecchia, D., and Liu, H. (2020c). Rank-based testing for semiparametric VAR models: a measure transportation approach. arXiv:2011.06062.

Hampel, F., Ronchetti, E., Rousseeuw, P., and Stahel, W. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69:383–393.

Holcblat, B. and Sowell, F. (2019). The empirical saddlepoint estimator. *arXiv preprint arXiv:1905.06977*.

Huber, P. J. (1981). *Robust Statistics.* Wiley.

Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics.* Wiley, 2nd edition.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning*, volume 112. Springer.

Jensen, J. L. (1995). *Saddlepoint Approximations*. Oxford University Press.

Jiang, C., La Vecchia, D., Ronchetti, E., and Scaillet, O. (2021). Saddlepoint approximations for spatial panel data models. arxiv.org/abs/2001.10377v3.

Kantorovich, L. V. (1942). On the translocation of masses. *(Dokl.) Acad. Sci. URSS*, 37(3):199–201.

Kass, R. (1989). The geometry of asymptotic inference. *Statistical Science*, 4(3):188–219.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46:33–50.

Kolassa, J. (2006). *Series Approximation Methods in Statistics*, volume 88. Springer.

Kolassa, J. and Li, J. (2010). Multivariate saddlepoint approximations in tail probability and conditional inference. *Bernoulli*, 16(4):1191–1207.

Kolouri, S., Park, S. R., Thorpe, M., Slepcev, D., and Rohde, G. K. (2017). Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34(4):43–59.

Kremer, E. (1982). A characterization of the Esscher-transformation. *ASTIN Bulletin: The Journal of the IAA*, 13(1):57–59.

Kullback, S. (1997). *Information Theory and Statistics*. Dover Publications.

La Vecchia, D. (2016). Stable asymptotics for M-estimators. *International Statistical Review*, 84(2):267–290.

La Vecchia, D. and Ronchetti, E. (2019). Saddlepoint approximations for short and long memory time series: A frequency domain approach. *Journal of Econometrics*, 213(2):578–592.

La Vecchia, D., Ronchetti, E., and Trojani, F. (2012). Higher-order infinitesimal robustness. *Journal of the American Statistical Association*, 107(500):1546–1557.

Léonard, C. (2007). A large deviation approach to optimal transport. Working paper.

McCann, R. J. and Guillen, N. (2011). Five lectures on optimal transportation: geometry, regularity and applications. *Analysis and Geometry of Metric Measure Spaces: Lecture Notes of the Séminaire de Mathématiques Supérieure (SMS) Montréal*, pages 145–180.

McCullagh, P. (2018). *Tensor Methods in Statistics*. Dover Publications.

Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*.

Monti, A. C. and Ronchetti, E. (1993). On the relationship between empirical likelihood and empirical saddlepoint approximation for multivariate M-estimators. *Biometrika*, 80(2):329–338.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.

Panaretos, V. M. and Zemel, Y. (2019). Statistical aspects of Wasserstein distances. *Annual Review of Statistics and its Application*, 6:405–431.

Panaretos, V. M. and Zemel, Y. (2020). *An Invitation to Statistics in Wasserstein Space*. Springer Nature.

Peyré, G. and Cuturi, M. (2019). Computational optimal transport: With applications to Data Science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Pitié, F., Kokaram, A. C., and Dahyot, R. (2007). Automated colour grading using colour distribution transfer. *Computer Vision and Image Understanding*, 107(1-2):123–137.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reid, N. (1988). Saddlepoint methods and statistical inference. *Statistical Science*, 3:213–227.

Reid, N. and Fraser, D. A. S. (1989). The geometry of asymptotic inference: Comment. *Statistical Science*, 4(3):231–233.

Rio, E. (2009). Upper bounds for minimal distances in the central limit theorem. In *Annales de l'IHP Probabilités et statistiques*, volume 45, pages 802–817.

Robinson, J., Ronchetti, E., and Young, G. (2003). Saddlepoint approximations and tests based on multivariate M-estimates. *The Annals of Statistics*, 31:1154–1169.

Rockafellar, R. T. (2015). *Convex Analysis*. Princeton University Press.

Ronchetti, E. and Sabolová, R. (2016). Saddlepoint tests for quantile regression. *Canadian Journal of Statistics*, 44(3):271–299.

Ronchetti, E. and Welsh, A. (1994). Empirical saddlepoint approximations for multivariate M-estimators. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56:313–326.

Santambrogio, F. (2015). Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94.

Serfling, R. J. (2009). *Approximation Theorems of Mathematical Statistics*, volume 162. Wiley.

Small, C. G. (2010). *Expansions and Asymptotics for Statistics*. Chapman and Hall/CRC.

Toma, A. and Broniatowski, M. (2011). Dual divergence estimators and tests: robustness results. *Journal of Multivariate Analysis*, 102:20–36.

Toma, A. and Leoni-Aubin, S. (2010). Robust tests based on dual divergence estimators and saddlepoint approximations. *Journal of Multivariate Analysis*, 101:1143–1155.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics.

Villani, C. (2009). *Optimal Transport: Old and New*, volume 338. Springer Science & Business Media.

Young, G. A. (2009). Routes to higher-order accuracy in parametric inference. *Australian & New Zealand Journal of Statistics*, 51(2):115–126.