Information and Inference: A Journal of the IMA (2024) **00**, iaae029
https://doi.org/10.1093/imaiai/iaae029

A new perspective on denoising based on optimal transport

Nicolás García Trillos*
Department of Statistics, University of Wisconsin-Madison, 1300 University Avenue, Madison, WI 53706, USA
*Corresponding author. Email: garciatrillo@wisc.edu

AND

Bodhisattva Sen
Department of Statistics, Columbia University, 1255 Amsterdam Avenue, New York, NY 10027, USA

[Received on 13 February 2024; revised on 11 July 2024; accepted on 22 September 2024]

In the standard formulation of the classical denoising problem, one is given a probabilistic model relating a latent variable $\Theta \in \Omega \subset \mathbb{R}^m$ ($m \geq 1$) and an observation $Z \in \mathbb{R}^d$ according to $Z \mid \Theta \sim p(\cdot \mid \Theta)$ and $\Theta \sim G^*$, and the goal is to construct a map to recover the latent variable from the observation. The posterior mean, a natural candidate for estimating Θ from Z, attains the minimum Bayes risk (under the squared error loss) but at the expense of over-shrinking the Z, and in general may fail to capture the geometric features of the prior distribution G^* (e.g. low dimensionality, discreteness, sparsity). To rectify these drawbacks, in this paper we take a new perspective on this denoising problem that is inspired by optimal transport (OT) theory and use it to study a different, OT-based, denoiser at the population level setting. We rigorously prove that, under general assumptions on the model, this OT-based denoiser is mathematically well-defined and unique, and is closely connected to the solution to a Monge OT problem. We then prove that, under appropriate identifiability assumptions on the model, the OT-based denoiser can be recovered solely from information of the marginal distribution of Z and the posterior mean of the model, after solving a linear relaxation problem over a suitable space of couplings that is reminiscent of standard multimarginal OT problems. In particular, due to Tweedie's formula, when the likelihood model $\{p(\cdot \mid \theta)\}_{\theta \in \Omega}$ is an exponential family of distributions, the OT-based denoiser can be recovered solely from the marginal distribution of Z. In general, our family of OT-like relaxations is of interest in its own right and for the denoising problem suggests alternative numerical methods inspired by the rich literature on computational OT.

Keywords: Bayes estimator; denoising estimands; optimal transport; empirical Bayes; latent variable model; multimarginal optimal transport; Tweedie's formula.

1. Introduction

Consider the following simple latent variable model:

$$Z \mid \Theta = \theta \sim p(\cdot \mid \theta) \quad \text{and} \quad \Theta \sim G^*, \tag{1.1}$$

where $\{p(\cdot \mid \theta)\}_{\theta \in \Omega}$ is a known parametric family of probability density functions (p.d.f.'s) on $\mathbb{R}^d$ ($d \geq 1$) with respect to (w.r.t.) the Lebesgue measure, and G^* is a probability distribution whose support is contained in the set Ω, a subset of $\mathbb{R}^m$ for $m \geq 1$. We only get to observe Z from the above model and Θ is the unobserved latent variable of interest. We denote by $P_{Z,\Theta}$ the joint distribution of (Z, Θ) on

$\mathbb{R}^d \times \Omega$. By defining a joint distribution over the observable Z and the latent variable Θ, the corresponding distribution of the observed variable is then obtained by marginalization; Z has marginal distribution μ with density (w.r.t. the Lebesgue measure)

$$f_{G^*}(z) := \int p(z \mid \theta)\, dG^*(\theta), \qquad \text{for } z \in \mathbb{R}^d. \tag{1.2}$$

Such latent variable models allow relatively complex marginal distributions to be expressed in terms of more tractable joint distributions over the expanded variable space and thus they provide an important tool for the analysis of multivariate data. Note that (1.1) captures a conceptual framework within which many disparate methods can be unified, including mixture models, factor models, etc; see e.g. [4]. In fact, (1.1) can be thought of as a simple Bayesian model where the prior distribution on Θ is G^*. A few important examples of such a setting are given below.

EXAMPLE 1 (Normal location mixture). Suppose that $p(z \mid \theta) = \varphi_\sigma(z - \theta)$, where $\varphi_\sigma(\cdot)$ is the p.d.f. of the multivariate normal distribution with mean 0 and variance $\sigma^2 I_d$ (σ^2 known), i.e. $\varphi_\sigma(z) := \frac{1}{(\sqrt{2\pi\sigma^2})^d} \exp(-\frac{z^\top z}{2\sigma^2})$, for $z \in \mathbb{R}^d$; here $m = d$. If G^* is a discrete distribution with finitely many atoms, then Z comes from a finite Gaussian mixture model. This model is ubiquitous in statistics and arises in many application domains including clustering; see e.g. [16, 57].

EXAMPLE 2 (Normal scale mixture). Suppose that $p(z \mid \theta) = \frac{1}{\theta}\varphi(\frac{z}{\theta})$, where $\varphi(\cdot)$ is the p.d.f. of the standard normal distribution on $\mathbb{R}$. Here G^* is a probability distribution on the positive real line $(0, \infty)$. This corresponds to the Gaussian scale mixture model; see [3]. This model has many applications including in Bayesian (linear) regression and multiple hypothesis testing, see e.g. [59, 68, 74].

EXAMPLE 3 (Uniform scale mixture). Suppose that G^* is a distribution on $(0, \infty)$ and $p(\cdot \mid \theta)$ corresponds to the uniform density on the interval $[0, \theta]$ (for $\theta > 0$). Thus, the marginal density of Z is given by $f_{G^*}(z) := \int \frac{1}{\theta}\mathbb{I}_{[0,\theta]}(z)\, dG^*(\theta) = \int_z^\infty \frac{1}{\theta}\, dG^*(\theta)$, for $z > 0$. It is well known that any (upper semicontinuous) non-increasing density on $(0, \infty)$ can be represented as f_{G^*} for a suitable G^* [32,p. 158]. This class of distributions arises naturally via connections with renewal theory (see e.g. [75]), multiple testing (see e.g. [51], [44]), etc.

We consider the goal of estimating the unobserved Θ in (1.1); we call this task *denoising* Z. Traditionally, this goal has been formulated as that of finding an estimator $\mathfrak{d}^*(\cdot)$ that minimizes the *Bayes risk* w.r.t. a *loss function* $\ell : \mathbb{R}^m \times \mathbb{R}^m \to [0, \infty)$, i.e.

$$\mathbb{E}\left[\ell(\mathfrak{d}(Z), \Theta)\right] \equiv \int_\Omega \int_{\mathbb{R}^d} \ell(\mathfrak{d}(z), \theta)\, p(z \mid \theta)\, dz\, dG^*(\theta) \tag{1.3}$$

over all measurable functions $\mathfrak{d} : \mathbb{R}^d \to \mathbb{R}^m$, where $(Z, \Theta) \sim P_{Z,\Theta}$ (i.e. $\Theta \sim G^*$ and $Z \mid \Theta = \theta \sim p(\cdot \mid \theta)$). The best estimator $\mathfrak{d}^*(Z)$ of Θ, in terms of minimizing (1.3), is called the *Bayes estimator* under the loss $\ell(\cdot, \cdot)$.

EXAMPLE 4 (Bayes estimator under squared error loss). When we use the loss function $\ell(a, \theta) := |a - \theta|^2$ (here $a, \theta \in \mathbb{R}^m$ and $|\cdot|$ denotes the usual Euclidean norm), the Bayes estimator $\overline{\theta}(\cdot)$ minimizing (1.3) turns out to be the *posterior mean*, i.e.

$$\overline{\theta}(Z) := \mathbb{E}[\Theta \mid Z]. \tag{1.4}$$

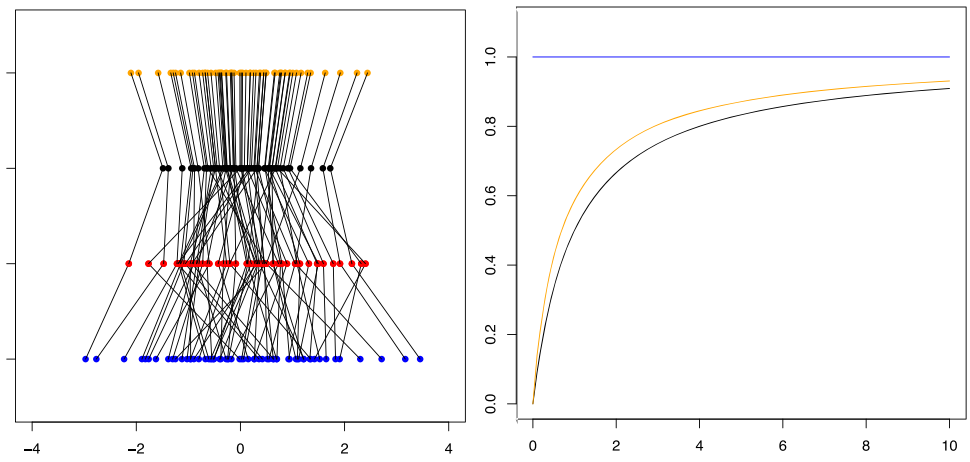

FIG. 1. Toy example with $n = 60$ in $d = 1$ where $p(\cdot \mid \theta)$ is the density of $N(\theta, 1)$ and $G^* = N(0, \tau^2)$. **Left:** observations $Z_1, \ldots, Z_n$ (in blue) obtained from model (1.1) with $\tau^2 = 1$ are connected to their true unobserved latent variables $\{\Theta_i\}_{i=1}^n$ (in red); the Bayes estimator $\overline{\theta}(Z_i)$ (in black) is connected to Θ_i (in red) and the corresponding OT-based denoiser $\delta^*(Z_i)$ (in orange). **Right:** plot of the risk curves of the three estimators of Θ—Z (in blue), $\overline{\theta}(Z)$ (in black) and $\delta^*(Z)$ (in orange)—as τ^2 varies from 0 to 10.

In this paper we take a different perspective on the denoising problem inspired by the theory of optimal transport (OT). To motivate our approach to estimating the unobserved Θ in (1.1), we first highlight a drawback of the Bayes estimator. Although the *posterior mean* $\overline{\theta}(Z) \equiv \mathbb{E}[\Theta \mid Z]$ in (1.4) attains the smallest Bayes risk (see (1.3)) among all estimators of Θ (under the squared error loss), its distribution is different from G^* (recall that $\Theta \sim G^*$). In fact, in some cases the Bayes estimator $\overline{\theta}(Z)$ yields a 'shrunken' estimate of Θ. The left panel of Fig. 1 illustrates this with $n = 60$ data points $Z_1, \ldots, Z_n$ (denoted by the blue dots) drawn from the model $Z_i \mid \Theta_i = \theta \sim N(\theta, 1)$, where $\Theta_i \overset{iid}{\sim} G^*$ with $G^* = N(0, \tau^2)$ and $\tau^2 = 1$. The latent Θ_i's are denoted by red dots, whereas the Bayes estimator $\overline{\theta}(Z_i)$ is depicted by black dots. We can see that the Bayes estimator (excessively) shrinks the observations in order to achieve optimal denoising (compare the distributions of the red and the black dots). The resulting distribution of the Bayes estimators $\overline{\theta}(Z)$ is $N(0, \frac{1}{2})$, which has a much smaller variance than $G^* \equiv N(0, 1)$.

In contrast, in this paper we consider the *OT-based denoiser* $\delta^*(Z)$ (see (2.9)), shown in the left plot of Fig. 1 by the orange dots, which corrects this drawback and produces estimates that have the distribution G^*; compare the distributions of the orange and the red dots. The plot of the risk functions for the three estimators—Z, $\overline{\theta}(Z)$ and $\delta^*(Z)$—as τ^2 varies shows that the proposed OT-based denoiser $\delta^*(Z)$ achieves the distributional stability (i.e. $\delta^*(Z) \sim G^*$) at very little cost; compare the risk functions for $\delta^*(Z)$ (in orange) and $\overline{\theta}(Z)$ (in black). See Remark 3 for the detailed computations.

This (over)-shrinkage by the Bayes estimator $\overline{\theta}(Z)$ is more acute when $d \geq 2$. In general, the Bayes estimator $\overline{\theta}(Z)$ is not necessarily guaranteed to lie 'close' to $\mathrm{spt}(G^*)$, the support[1] of G^* (recall that $\Theta \sim G^*$). To illustrate this, in Fig. 2 we consider another example, this time with $d = m = 2$. Here we take $n = 60$ data points $Z_1, \ldots, Z_n \in \mathbb{R}^2$ (depicted by the blue dots in the left panel of Fig. 2) drawn from

[1] The smallest closed set containing probability mass 1.

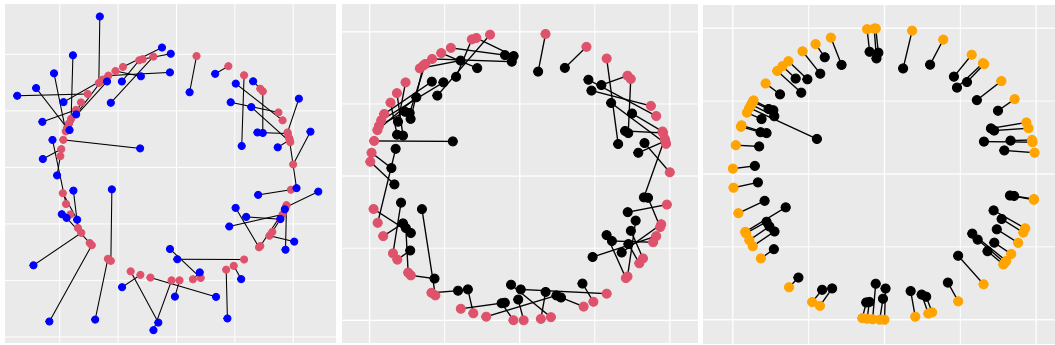

FIG. 2. Toy example with $n = 60$ in $d = 2$ where $p(\cdot \mid \theta)$ is the density of $N(\theta, (0.3)^2 \cdot I_2)$ and G^* is the uniform distribution on the unit circle. **Left:** observations $Z_1, \ldots, Z_n$ (in blue) obtained from model (1.1) are connected to the corresponding unobserved latent variables $\{\Theta_i\}_{i=1}^n$ (in red). **Centre:** the Bayes estimator $\overline{\theta}(Z_i)$ (in black) is connected to Θ_i (in red), for every $i = 1, \ldots, n$. **Right:** the Bayes estimator $\overline{\theta}(Z_i)$ (in black) is connected to its corresponding OT-based denoiser $\delta^*(Z_i)$ (in orange) lying on the circle.

the normal location mixture model (Example 1) with the latent variables $\Theta_1, \ldots, \Theta_n$ drawn uniformly on the circle of radius 1 (shown by the red dots in the left panel of Fig. 2). We connect Z_i with Θ_i by a black line, for each $i = 1, \ldots, n$, in the plot. The middle panel of Fig. 2 shows the Bayes estimator[2] at the observed data points $\overline{\theta}(Z_i)$ (depicted by the black dots) connected to the corresponding Θ_i's (in red). As can be easily seen from this plot, the Bayes estimator shrinks most of the observations towards $0 \in \mathbb{R}^2$. In contrast, our proposed OT-based denoiser corrects this drawback and maps the Bayes estimator $\overline{\theta}(Z_i)$ to $\delta^*(Z_i)$ (shown in the right panel of Fig. 2 by orange dots) which lies on the circle. Note that $\delta^*(Z_i)$, by definition, takes values in $\mathrm{spt}(G^*)$, the support of G^*.

In fact, if the goal is to estimate $\Theta \sim G^*$, it is reasonable to restrict $\mathfrak{d}(\cdot)$ in (1.3) to all estimators such that $\mathfrak{d}(Z)$ is distributed (approximately) as G^*. This type of requirement has been explored in previous works in the literature (see e.g. [34, 55]) and is particularly important when we believe that G^* is discrete with a few atoms (which corresponds to the *clustering problem*) or when we believe that G^* has 'structure' (e.g. supported on a lower dimensional manifold in $\mathbb{R}^m$). In light of this discussion, it is natural to seek solutions to

$$\inf_{\delta:\mathbb{R}^d \to \mathbb{R}^m} \mathbb{E}_{(Z,\Theta)\sim P_{Z,\Theta}} \left[|\delta(Z) - \Theta|^2 \right] \qquad \text{subject to} \quad \delta(Z) \sim G^*, \tag{1.5}$$

among all measurable functions $\delta : \mathbb{R}^d \to \mathbb{R}^m$, where we consider $\ell(\cdot, \cdot)$ to be the squared error loss for simplicity; see Appendix D for a discussion on more general loss functions. The constraint $\delta(Z) \sim G^*$ ensures that the 'estimator' $\delta(Z)$ of Θ has the same distribution as G^* (in particular, the same support as $\Theta \sim G^*$), thereby addressing the above drawbacks; cf. (1.3). Solutions of (1.5), if they exist, can be described as *distortion* minimizers under a perfect *perception* quality constraint; see [5, 34] for definitions and motivation for this terminology.

[2] Here, the Bayes estimator, defined in (1.4), is approximated by a fine discretization of G^*, i.e. $G^* \approx \frac{1}{M} \sum_{i=1}^M \delta_{a_i}$, where $M = 1200$ and the a_i's lie uniformly on the circle.

The above discussion leads to the following natural questions:

Q1. Under what conditions can it be guaranteed that there exist solutions to problem (1.5)? Are these solutions unique?

Q2. If there exists a solution, how can one characterize it, and what potential approaches can one follow to find it?

Q3. Is it possible to obtain a solution solely based on the marginal distribution of observations and knowledge of the likelihood model (without explicitly using G^*)?

The purpose of this paper is to provide answers to the above questions in the population level setting, implicitly also assuming that G^* is known. In this process, we lay down some mathematical foundations and outline some strategies for future implementation of our ideas in finite data settings, with known or unknown G^*. Q3 is motivated by the fact that, in general, an attempt to recover G^* from observations can lead to a difficult deconvolution problem; see more discussion below.

Our first main result, **Theorem** 4, states that, under certain assumptions on the model, problem (1.5) indeed possesses a unique solution $\delta^*(\cdot)$; throughout the paper, by unique solution we mean unique μ-a.e. What is more, this solution can be found by solving an OT problem (defined precisely in (2.10)) between the distribution of the Bayes estimator $\overline{\theta}(Z)$ and G^*, and can be characterized as the composition of the gradient of a certain convex function and $\overline{\theta}(\cdot)$. We refer to this $\delta^*(\cdot)$ as the *OT-based denoiser* associated with the model $(\{p(\cdot \mid \theta)\}_{\theta\in\Omega}, G^*)$. Problem (1.5) can also be interpreted as an extreme case in a family of problems with a soft penalty defined according to

$$\inf_{\delta:\mathbb{R}^d\to\mathbb{R}^m} \mathbb{E}_{(Z,\Theta)\sim P_{Z,\Theta}}\left[|\delta(Z)-\Theta|^2\right] + \frac{1}{2\tau} W_2^2\left(\delta_\sharp\mu, G^*\right), \tag{1.6}$$

where $\tau > 0$ is a tuning parameter, $\delta_\sharp\mu$ is the pushforward of μ by δ (i.e. the distribution of $\delta(Z)$ if $Z \sim \mu$; see Definition 1) and $W_2(\cdot,\cdot)$ denotes the 2-Wasserstein distance between probability distributions (see Definition 2.3). Formally, when $\tau \to \infty$ we recover the standard unconstrained risk minimization problem, whose solution is the Bayes estimator, whereas we recover problem (1.5) when $\tau \to 0$. For any other value of τ in between these two extremes, **Theorem** 5 guarantees that the solution to (1.6) is unique and can be explicitly written as a simple linear interpolation of the OT-based denoiser $\delta^*(\cdot)$ and the Bayes estimator $\overline{\theta}(\cdot)$, a result very closely related to the characterization of the so-called *distortion-perception trade-off* in Wasserstein space established in [34]. We will refer to (1.6) as a *latent space penalization* approach to denoising, given that the penalty term $W_2^2(\delta_\sharp\mu, G^*)$ involves an explicit comparison of distributions in the latent space $\Omega \subset \mathbb{R}^m$. It is worth highlighting that other optimization problems similar to (1.6) have been considered in papers such as [5, 72] (see also references therein), where the W_2 distance between measures is substituted by other metrics over probability measures, including the 1-OT distance W_1 and other loss functions as used in generative adversarial networks. As mentioned earlier, and in contrast to the aforementioned papers, in this paper we pursue an in-depth analysis of the properties of solutions to problems like (1.6) (or (1.7) below) and suggest novel strategies to find them.

Although the characterization of the OT-based denoiser $\delta^*(\cdot)$ as a solution to an OT problem is appealing, in many real applications G^* may be unknown, making this characterization difficult to implement. One possible approach to go around this issue is to estimate G^* using i.i.d. data from (1.1) using tools from what is usually referred to in statistics as deconvolution (see e.g. [9, 29, 58, 77]). This approach is also taken in the *empirical Bayes* literature; see, e.g. [23, 26, 45, 63, 67], as well as the brief discussion on this topic that we present in Appendix E. In this paper, however, we offer an alternative approach and study yet another formulation for the denoising problem that closely resembles (1.6) but

where we directly work with μ, the (marginal) distribution of the observed data (see (1.2)). Indeed, we consider the optimization problem:

$$\inf_{\delta:\mathbb{R}^d\to\mathbb{R}^m} \mathscr{E}_\tau(\delta) := \mathbb{E}_{(Z,\Theta)\sim P_{Z,\Theta}}\left[|\delta(Z)-\Theta|^2\right] + \frac{1}{2\tau}W_2^2(\mu_\delta,\mu); \tag{1.7}$$

here, for a given map δ we define μ_δ as the probability measure over $\mathbb{R}^d$ defined as

$$\mu_\delta(A) := \int_A \int_{\mathbb{R}^d} p(z' \mid \delta(z))\, d\mu(z)\, dz', \quad \forall A \subseteq \mathbb{R}^d \text{ Borel measurable.} \tag{1.8}$$

In words, μ_δ is the marginal distribution of the variable z assuming that the underlying distribution of the latent variable θ is given by $G = \delta_\sharp \mu$. We will refer to (1.7) as an *observable space penalization* approach to denoising, given that the penalty term $W_2^2(\mu_\delta,\mu)$ involves an explicit comparison of distributions in the observable space $\mathbb{R}^d$. In **Proposition** 7 we show that, under suitable assumptions, the objective function $\mathscr{E}_\tau$ (in (1.7)) is Gateaux differentiable w.r.t. the target $\delta \in \mathbb{L}^2(\mathbb{R}^d : \mathbb{R}^m;\mu)$[3] and provide an explicit formula for its gradient (see (3.1)). The formula for the gradient, which can be easily adapted to the empirical setting, can in principle be used to implement a first-order optimization method seeking a solution for (1.7). Unfortunately, problem (1.7) is non-convex in δ and one cannot guarantee the convergence of a steepest descent scheme towards a global minimizer of $\mathscr{E}_\tau(\cdot)$. In fact, even the existence of global solutions to (1.7) is not guaranteed by straightforward arguments in the calculus of variations. The main technical difficulty for this is the lack of lower semicontinuity of the functional $\delta \mapsto W_2^2(\mu_\delta,\mu)$ w.r.t. the *weak* topology in the Hilbert space $\mathbb{L}^2(\mathbb{R}^d : \mathbb{R}^m;\mu)$ (see Definition C1), a natural topology where one can guarantee pre-compactness of minimizing sequences.

Despite the above discussion, we can prove that indeed there exist solutions to (1.7); see **Theorem** 11. This is achieved by considering a suitable relaxation argument where we 'lift' the original problem (1.7) to a problem over couplings (see (3.4) for details) that, while not of a standard type in OT theory, does resemble multimarginal optimal transport (MOT) problems. Like MOT problems, our relaxation is linear, and its search space enjoys better compactness properties than the original problem (1.7) that in particular can be used to prove existence of solutions (see **Theorem** 10). This relaxation, which we show is exact under suitable assumptions, also motivates the use of computational tools in OT for constructing solutions of (1.7); this will be explored in future work. Finally, we highlight that this relaxation is the key mathematical construction that allows us to prove **Theorem** 13, which states that, under the identifiability assumptions on the probabilistic model that are written down precisely in Assumption 12, the solutions δ_τ^* of (1.7) converge, as $\tau \to 0$, to the OT-based denoiser δ^*; in Remark 12 we discuss the non-identifiable case.

As we discuss in Section 6, in order to use the relaxation problem (3.4) to approximate δ^* from finitely many observations, one would first need to estimate $\overline{\theta}(\cdot)$ from the available data. This is where Tweedie's formula (see (B.5) in Appendix B) can be very useful. This formula expresses the posterior mean $\overline{\theta}(\cdot)$ in an exponential family model (see Appendix A) in terms of the marginal density f_{G^*} of the observations (and its gradient) only, and can thus be estimated (non-parametrically) directly from observations $Z_1,\dots,Z_n$, say via kernel density estimation. We thus anticipate to be able to construct

[3] $\mathbb{L}^2(\mathbb{R}^d : \mathbb{R}^m;\mu)$ is the space of vector valued (equivalence classes of) measurable functions from $\mathbb{R}^d$ into $\mathbb{R}^m$ that are square-integrable w.r.t. μ.

consistent estimators for δ^* without knowing G^* explicitly or having to directly estimate it, at least in the case when the likelihood model is an exponential family of distributions.

1.1 *Previous works*

The main motivation of our work comes from the theory of empirical Bayes (Robbins, [63]) and its recent revisitations (see e.g. Efron, [20–22, 26]) which consider large data sets that arise from parallel and similar experiments. In the classical empirical Bayes set-up the unknown parameters arising from the parallel experiments are assumed to be i.i.d. random variables with an unknown common prior distribution G^*.

Typically, empirical Bayes methodologies (see e.g. [6,23–25,39,40,45,47–50,67,79] and the references therein) provide statistical procedures which approximate the Bayes rule for the true model (without specifying a prior). In this paper we question this very premise and illustrate (cf. Figs 1 and 2) that the Bayes estimator, which is optimal in terms of squared error risk, deforms the underlying true distribution of the latent variables (i.e. Θ_i's) and may not be ideal in large-scale denoising problems. This naturally leads us to the field of OT in search of strategies to correct for this deformation.

The area of OT has seen rapid growth in the past years with various applications in statistics and machine learning. In statistical theory, OT appears in at least two general settings: (i) as an interesting estimation problem in its own right, where one uses observations to either approximate the Wasserstein distance between two ground truth distributions (see e.g. [14, 33, 73]) or to estimate the actual OT map between them (see e.g. [12,18,30,43,56]), and (ii) as a tool to propose and/or analyse statistical models in classical Euclidean settings (e.g. as in [10,13,38,41,42,65]) or in more abstract settings where data sets consist of, for example, probability distributions (e.g. as in the regression setting for distribution-on-distribution data explored in works like [36, 37, 76]). This paper better fits the second class of works, but the adaptation of our ideas to the finite data setting will require the exploration of questions that fall in the first category mentioned above.

Connections between the denoising problem (understood in a general sense) and ideas from computational OT have been explored before in applications to image and signal denoising in works like [5, 17, 34, 72]; a hard constraint version of (1.6) has been considered in [34], where a quantitative form of the so-called distortion-perception trade-off is established. Modern approaches for noise removal with additional good perception quality constraints have been proposed in [15]. These approaches take advantage of the gradient structure that denoisers often have. In this paper, we pursue a deeper mathematical analysis than previous works in the literature and explore new approaches, motivated by ideas from the theory of OT, for recovering the OT-based denoiser. One of the key tools that we use in our paper from the literature of OT is the concept of MOT (see [60]), which has been explored in the past in a variety of fields including density functional theory in physics and chemistry [7, 11], economics [8, 28] and image processing [62], among others. This paper introduces new applications of closely related OT problems.

1.2 *Outline*

The rest of the paper is organized as follows. In Section 2 we present our main results on problems (1.5) and (1.6). First, in Section 2.1 we introduce some necessary notation and background that we use in the rest of the paper. Then, in Section 2.2 we state our first main result, Theorem 4, which establishes the existence and uniqueness of solutions of (1.5). In Section 2.3 we state Theorem 5, which characterizes the solution of the soft penalty problem (1.6). Section 3 is devoted to the observable space penalization problem (1.7). First, we provide a characterization of the Frechét derivative of its objective under suitable

differentiability assumptions on the likelihood model. Then we present Theorems 11 and 13, where, respectively, we state the existence of solutions to (1.7) and characterize the behaviour of solutions to (1.7) as the parameter τ goes to zero. In particular, Theorem 13 states that, under suitable identifiability assumptions, the OT-based denoiser δ^* can be recovered as a limit of solutions to (1.7). Sections 4 and 5 are devoted to the proofs of our main results from Sections 2 and 3, respectively. In Section 6, we discuss some future directions for research stemming from this work.

In Appendices A–E we provide various discussions connected to our main results. In particular, in Appendices A–B we introduce exponential families of distributions and describe Tweedie's formula. In Appendix C we state and prove a few results from measure theory and functional analysis that are relevant to the proof of Theorem 13. Appendix D briefly describes OT formulations of (1.5) when using more general loss functions (beyond the squared error loss). In Appendix E we briefly review the (non-parametric) maximum likelihood estimator of G^*, which could potentially be used to implement (1.5) in practical settings.

2. Denoising with latent space penalization

2.1 *Preliminaries*

We first introduce some definitions and notation from the theory of OT (see e.g. [69, 70]). For any metric space $\mathscr{X}$, let $\mathfrak{B}(\mathscr{X})$ denote the set of all Borel measurable subsets of $\mathscr{X}$, and let $\mathscr{P}(\mathscr{X})$ be the set of all Borel probability measures over $\mathscr{X}$. It will be convenient to first introduce the notion of *pushforward* of a measure by a map and rewrite the constraint in (1.5) in terms of pushforwards.

Definition 1 (Pushforward of a measure). Given a measurable map $\delta : \mathscr{X} \to \mathscr{Y}$ and a probability measure ν over $\mathscr{X}$, the measure $\delta_\sharp \nu$, the pushforward of ν by δ, is the measure defined according to $\delta_\sharp \nu(A) := \nu(\delta^{-1}(A))$ for every Borel subset A of $\mathscr{Y}$. In other words, if $X \sim \nu$, then $\delta(X) \sim \delta_\sharp \nu$.

Remark 1. The constraint $\delta(Z) \sim G^*$ in (1.5) can be rewritten as $\delta_\sharp \mu = G^*$.

Let two probability measures $\nu, \widetilde{\nu}$ be defined over two Polish spaces $\mathscr{X}$ and $\mathscr{Y}$, and consider a lower semicontinuous cost function $c : \mathscr{X} \times \mathscr{Y} \to [0, \infty]$. The dual of the Kantorovich OT problem (see e.g. [69, 70])

$$C(\nu, \widetilde{\nu}) := \min_{\pi \in \Gamma(\nu, \widetilde{\nu})} \int \int c(x, y)\, d\pi(x, y), \tag{2.1}$$

where $\Gamma(\nu, \widetilde{\nu})$ denotes the set of all Borel probability measures on the product $\mathscr{X} \times \mathscr{Y}$ with marginals ν and $\widetilde{\nu}$ (a.k.a. couplings between ν and $\widetilde{\nu}$), is the problem

$$\sup_{\phi, \psi} \int \phi(x)\, d\nu(x) + \int \psi(y)\, d\widetilde{\nu}(y), \quad \text{s.t.} \phi(x) + \psi(y) \leq c(x, y), \quad \nu\text{-a.e. } x \in \mathscr{X},\ \widetilde{\nu}\text{-a.e. } y \in \mathscr{Y}, \tag{2.2}$$

where ϕ and ψ are, respectively, in $\mathbb{L}^1(\mathscr{X}, \nu)$ and $\mathbb{L}^1(\mathscr{Y}, \widetilde{\nu})$. Theorem 5.10 in [70] guarantees that primal and dual problems are equivalent. Any solution pair (ϕ, ψ) of (2.2), if it exists, will be referred to as *optimal dual potentials* for the OT problem (2.1).

We will often consider the setting where the space $\mathscr{X}$ is a subset of some Euclidean space, $\mathscr{X} = \mathscr{Y}$, and $c(x, y) = |x - y|^2$. When in this setting, we will refer to (2.1) as the 2-OT problem between ν and

$\widetilde{\nu}$ and denote by $W_2^2(\nu, \widetilde{\nu})$ the minimum value in (2.1), which is nothing but the square of the so-called Wasserstein distance between ν and $\widetilde{\nu}$.

DEFINITION 2 (2-Wasserstein distance). Given two probability measures $\nu, \widetilde{\nu}$ over $\mathbb{R}^p$ with finite second moments, we define their Wasserstein distance $W_2(\nu, \widetilde{\nu})$ as

$$W_2^2(\nu, \widetilde{\nu}) := \min_{\pi \in \Gamma(\nu, \widetilde{\nu})} \int |x - y|^2 \, d\pi(x, y). \tag{2.3}$$

A landmark result in the theory of OT due to Brenier characterizes the optimal coupling π for the 2-OT problem between two measures ν and $\widetilde{\nu}$ when ν is absolutely continuous w.r.t. the Lebesgue measure; see e.g. Villani [70, Theorem 3.15].

THEOREM 1 (Brenier). Let ν and $\widetilde{\nu}$ be two Borel probability measures over $\mathbb{R}^p$ such that $\int |x|^2 \, d\nu(x) < \infty$ and $\int |y|^2 \, d\widetilde{\nu}(y) < \infty$. Suppose further that ν has a Lebesgue density. Then there exists a convex function $\psi : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ whose gradient $T = \nabla \psi$ pushes ν forward to $\widetilde{\nu}$. In fact, there exists only one such T that arises as the gradient of a convex function, i.e. T is unique ν-a.e. Moreover, T uniquely minimizes Monge's problem:

$$\inf_{T \,:\, T_\sharp \nu = \widetilde{\nu}} \int |x - T(x)|^2 d\nu(x)$$

and the coupling $(\mathrm{Id} \times T)_\sharp \nu$ uniquely minimizes (2.3). In the above and in the remainder of the paper, the map $(\mathrm{Id} \times T) : \mathbb{R}^p \to \mathbb{R}^p \times \mathbb{R}^p$ is defined as $(\mathrm{Id} \times T)(x) = (x, T(x))$.

2.2 *Rewriting* (1.5) *as an OT problem*

In this subsection we study problem (1.5) and develop its connection with standard Monge and Kantorovich OT problems with a suitable cost function. Due to Remark 1, problem (1.5) can be written as

$$\min_{\delta \,:\, \delta_\sharp \mu = G^*} \mathbb{E}_{(Z,\Theta) \sim P_{Z,\Theta}} \left[|\delta(Z) - \Theta|^2 \right]. \tag{2.4}$$

In turn, problem (2.4) is equivalent to

$$\min_{\delta \,:\, \delta_\sharp \mu = G^*} \mathbb{E}_{Z \sim \mu} \left[|\delta(Z) - \overline{\theta}(Z)|^2 \right], \tag{2.5}$$

where

$$\overline{\theta}(z) := \mathbb{E}_{(Z,\Theta) \sim P_{Z,\Theta}} [\Theta \mid Z = z] \tag{2.6}$$

is the *posterior mean* (and the Bayes estimator under the quadratic loss). This equivalence follows from the well-known bias-variance decomposition for the squared error loss:

$$\mathbb{E}[|\delta(Z) - \Theta|^2 \mid Z] = \mathbb{E}[|\delta(Z) - \overline{\theta}(Z)|^2 \mid Z] + \mathbb{E}[|\overline{\theta}(Z) - \Theta|^2 \mid Z],$$

which implies that for any arbitrary $\delta : \mathbb{R}^d \to \mathbb{R}^m$ we have

$$\mathbb{E}_{(Z,\Theta)\sim P_{Z,\Theta}}\left[|\delta(Z)-\Theta|^2\right] = \mathbb{E}_{Z\sim\mu}\left[|\delta(Z)-\overline{\theta}(Z)|^2\right] + \mathbb{E}_{(Z,\Theta)\sim P_{Z,\Theta}}\left[|\overline{\theta}(Z)-\Theta|^2\right],$$

from where it follows that the objective in (2.4) is equal to the objective function in (2.5) up to the constant $R_{\text{Bayes}} := \mathbb{E}_{(Z,\Theta)\sim P_{Z,\Theta}}[|\overline{\theta}(Z)-\Theta|^2]$, i.e. the Bayes risk.

The advantage of problem (2.5) is that, as discussed below, it is amenable to the type of relaxation methods that have been studied in OT theory. Indeed, in order to construct a solution to (2.5) (and thus also to (2.4)), at least for certain families of models $(\{p(\cdot \mid \theta)\}_{\theta\in\Omega}, G^*)$ satisfying suitable assumptions, we will first consider a Kantorovich relaxation of (2.5) given by

$$\min_{\pi\in\Gamma(\mu,G^*)} \iint c_{G^*}(z,\vartheta)\, d\pi(z,\vartheta), \tag{2.7}$$

where the cost function $c_{G^*}(\cdot,\cdot)$ is defined as

$$c_{G^*}(z,\vartheta) := |\overline{\theta}(z)-\vartheta|^2, \quad (z,\vartheta)\in\mathbb{R}^d\times\Omega; \tag{2.8}$$

note the dependence of G^* on the cost function $c_{G^*}(z,\vartheta)$ in (2.8) via the Bayes estimator $\overline{\theta}(z)$, which depends on G^*.

We make the following assumptions.

Assumption 2. The distribution G^* is such that $\int_\Omega |\vartheta|^2 dG^*(\vartheta) < \infty$, i.e. G^* has finite second moments.

Assumption 3. The measure $\overline{\theta}_\sharp\mu$ is absolutely continuous w.r.t. the Lebesgue measure in $\mathbb{R}^m$.

Remark 2 (On our assumptions). Since μ is absolutely continuous w.r.t. the Lebesgue measure (recall (1.2)), note that Assumption 3 holds if we assume that the map $\overline{\theta} : \mathbb{R}^d \to \mathbb{R}^m$ is locally Lipschitz (and thus differentiable Lebesgue a.e.) and that the Jacobian matrix $D\overline{\theta}(z) \in \mathbb{R}^{d\times m}$ has full rank μ-a.e. z; indeed, this implication follows from the so-called coarea formula (see e.g. the theorem in section 3.1 in [31]). Thus, implicitly, we would be assuming that $d \geq m$. In particular, the above is satisfied for the following scenario. If $\{p(\cdot \mid \theta)\}_{\theta\in\Omega}$ is a regular k-parameter exponential family in canonical form, then $\overline{\theta}(\cdot)$ is the gradient of a convex function $\kappa(\cdot)$ (which happens to be the log-partition function of the family); see Section A and Section B. Moreover, $\kappa(z)$ is a strictly convex function of z on its domain if the representation is *minimal*; see e.g. Wainwright [71, Proposition 3.1]. As convex functions are a.e. twice continuously differentiable, the Jacobian matrix $D\overline{\theta}(z) \in \mathbb{R}^{d\times m}$ exists a.e., and thus Assumption 3 is automatically satisfied. Assumption 2 just assumes a finite second moment condition on G^*, which is quite mild.

We are ready to state our first main result.

Theorem 4. Under Assumptions 2 and 3, there exists a unique solution π^* to problem (2.7) with cost function (2.8), which takes the form

$$\pi^* = (\mathrm{Id}\times\delta^*)_\sharp\mu$$

for a map $\delta^*(\cdot)$ that is the μ-a.e. unique solution to problem (2.4), i.e. it is the OT-based denoiser. Furthermore, $\delta^*(\cdot)$ can be written as

$$\delta^*(z) = \nabla\varphi^*(\overline{\theta}(z)), \qquad \text{for } z \in \mathbb{R}^d, \tag{2.9}$$

to be read: 'the gradient of the function φ^* evaluated at $\overline{\theta}(z)$', where $\varphi^* : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ is a convex function. In fact, $T^* := \nabla\varphi^*$ is the solution to the standard quadratic cost Monge OT problem

$$\min_{T:T_\sharp(\overline{\theta}_\sharp\mu)=G^*} \int |\theta - T(\theta)|^2 \, d\overline{\theta}_\sharp\mu(\theta) \tag{2.10}$$

between the measures $\overline{\theta}_\sharp\mu$ and G^*.

The proof of Theorem 4 is presented in Section 4. It builds upon Brenier's theorem (Theorem 1). The first part of Theorem 4 implies that, under Assumptions 2 and 3, the value of Kantorovich's relaxation problem in (2.7) is indeed the same as that of Monge's problem (2.4). Further, the optimal coupling in (2.7) yields the solution to (2.4). Theorem 4 further says that the optimal solution $\delta^*(\cdot)$ of (2.4) is related to the Bayes estimator (2.6); in fact, $\delta^*(\cdot)$ pushes the Bayes estimator $\overline{\theta}(\cdot)$ to satisfy the distributional constraint $\delta^*(Z) \sim G^*$. The fact that $\delta^*(\cdot)$ has such a simple form is not immediately obvious from the original formulation of the problem in (1.5).

Remark 3 (Normal–normal location model). Suppose that $d = m$ and $\Theta \sim N_m(\theta^*, \Sigma^*)$ and $Z \mid \Theta = \theta \sim N_d(\theta, \Sigma)$, where $\theta^* \in \mathbb{R}^m$ is known, and $\Sigma^* \in \mathbb{R}^{m\times m}$ and $\Sigma \in \mathbb{R}^{d\times d}$ are symmetric positive definite (fixed) matrices. It is then well known that

$$\overline{\theta}(Z) = \Sigma^*(\Sigma^* + \Sigma)^{-1}Z + \Sigma(\Sigma^* + \Sigma)^{-1}\theta^*,$$

which shows that

$$\overline{\theta}(Z) \sim N_m\Big(\theta^*, A\Big), \qquad \text{where} \quad A := \Sigma^*(\Sigma^* + \Sigma)^{-1}\Sigma^*, \tag{2.11}$$

as unconditionally, $Z \sim N_d(\theta^*, \Sigma^* + \Sigma)$. Therefore, by Theorem 4, to find the OT-based denoiser δ^* we need to find the OT map T^* between the distributions $N_m\Big(\theta^*, A\Big)$ and $N_m\Big(\theta^*, \Sigma^*\Big)$, which is given by

$$T^* : y \mapsto \theta^* + B(y - \theta^*), \qquad \text{where} \quad B := A^{-1/2}(A^{1/2}\Sigma^*A^{1/2})^{1/2}A^{-1/2}.$$

Thus, the OT-based denoiser δ^* has the form $\delta^*(Z) = T^*(\overline{\theta}(Z))$.

To get a better feel for the estimators—$\overline{\theta}(Z)$ and $\delta^*(Z)$—in this problem, let us consider the special case $d = m = 1$ with $\Sigma = 1$ and $\Sigma^* = \tau^2$ and $\theta^* = 0$. Here we can see that the Bayes estimator satisfies

$$\overline{\theta}(Z) := \tau^2(1 + \tau^2)^{-1}Z, \qquad \text{and thus,} \qquad \overline{\theta}(Z) \sim N(0, \tau^4/(1 + \tau^2)).$$

Note that $\tau^4/(1+\tau^2) < \tau^2$ and thus the Bayes estimator has lower variance than $G^* \equiv N(0, \tau^2)$ (see Fig. 1 for an illustration of this phenomenon via a simple simulation). However, the OT-based denoiser δ^*

has the form $T^*(\overline{\theta}(Z))$, where $T^*(y) := \tau(\tau^4/(1+\tau^2))^{-1/2}y$. Here the Bayes risk (i.e. $\mathbb{E}[(\overline{\theta}(Z)-\Theta)^2]$) is $\tau^2/(1+\tau^2) < 1$ and the risk of δ^* is $2\tau^2\left(1-\frac{\tau}{1+\tau^2}\right)$ (see the black and orange curves in the right panel of Fig. 1).

REMARK 4 (When $m = 1$). In the special case when $m = 1$, the OT-based denoiser $\delta^*(\cdot)$ in (2.9) can be explicitly expressed as $\delta^*(z) = F^{-1}_{G^*}(F_{\overline{\theta}}(\overline{\theta}(z)))$, for $z \in \mathbb{R}$, where $F^{-1}_{G^*}$ is the quantile function corresponding to the distribution G^* (i.e. $F^{-1}_{G^*}(p) := \inf\{x \in \mathbb{R} : p \le F_{G^*}(x)\}$, for $p \in (0,1)$) and $F_{\overline{\theta}}$ is the distribution function of the random variable $\overline{\theta}(Z)$. This follows easily from the fact that, in one dimension, Brenier maps have explicit solutions in terms of distribution/quantile functions.

2.3 *Soft penalty versions of* (2.4)

We now consider problem (1.6), which is a type of relaxation of problem (2.4) where we use a soft penalty on δ to enforce $\delta_\sharp\mu$ to be sufficiently close to G^* as opposed to enforcing a hard constraint as in (2.4). The strength of the penalization is determined by the parameter τ, and, intuitively, we should expect to recover the classical Bayes estimator $\overline{\theta}(Z)$ when $\tau \to \infty$, and the OT-based denoiser $\delta^*(Z)$ when $\tau \to 0$. As we show in the result below (see Section 4.2 for its proof), the estimators recovered by solving (1.6) are simple linear interpolators of the Bayes estimator $\overline{\theta}(Z)$ and $\delta^*(Z)$.

THEOREM 5. Under the same assumptions as in Theorem 4, there exists a unique solution δ^*_τ to (1.6). Furthermore, the map $\delta^*_\tau(\cdot)$ can be written as

$$\delta^*_\tau(z) = \frac{2\tau}{1+2\tau}\overline{\theta}(z) + \frac{1}{1+2\tau}\delta^*(z), \qquad \text{for} \quad z \in \mathbb{R}^d, \tag{2.12}$$

where $\delta^*(\cdot)$ is the map from Theorem 4.

REMARK 5 (On the proof of Theorem 5). The proof of Theorem 5 is based on a simple relaxation argument that mimics the relaxation in [1] used to reformulate Wasserstein barycentre problems as MOT problems.

3. Denoising with observable space penalization

Although the characterization of the OT-based denoiser $\delta^*(\cdot)$ as a solution to an OT problem is appealing, in most real applications G^* is unknown. As discussed in the Introduction right before (1.7) (see also Appendix E), one possible approach to go around this issue is to estimate G^* using i.i.d. data from (1.1) using deconvolution techniques that are also used in the empirical Bayes literature; the resulting approach is in line with the concept of g-modelling discussed in [24]. In this section, and in the spirit of the f-modelling discussed in [24], we take a different approach and study yet another formulation for the denoising problem that closely resembles (1.6) but where we directly work with μ, the (marginal) distribution of the observed data (see (1.2)). In particular, we consider the optimization problem (1.7) with objective $\mathscr{E}_\tau(\delta)$.

First, we provide an explicit formula for the Gateaux derivative of $\mathscr{E}_\tau$ w.r.t. δ, when the likelihood model is sufficiently regular. In principle, this Gateaux derivative can be used to implement a first-order optimization method to find solutions of (1.7), but as discussed in the Introduction, the convergence to global optimizers of this scheme cannot be guaranteed due to the non-convexity of $\mathscr{E}_\tau$. For this reason we

consider an alternative methodology which holds under milder assumptions and which will allow us to: (1) prove the existence of solutions of (1.7), (2) suggest a linear optimization problem for solving (1.7) and (3) recover δ^*, the OT-based denoiser, without explicit knowledge of G^*. Throughout this section we make the following assumption, which is used to guarantee that problem (1.7) is non-trivial.

ASSUMPTION 6. The marginal distribution μ with density as in (1.2) has finite second moments.

PROPOSITION 7. Suppose that Assumptions 2 and 6 hold. Suppose also that the likelihood model is such that $p(z \mid \theta)$ is continuously differentiable in θ (for every $z \in \mathbb{R}^d$). Let $\delta \in \mathbb{L}^2(\mathbb{R}^d : \mathbb{R}^m; \mu)$ and suppose that μ, μ_δ (recall μ_δ was defined in (1.8)) are such that they admit a unique (up to constant shifts) solution $(\widetilde{\phi}, \widetilde{\psi})$ to the dual of the 2-OT problem between μ and μ_δ ; in particular,

$$\int \widetilde{\phi}\, d\mu + \int \widetilde{\psi}\, d\mu_\delta = W_2^2(\mu, \mu_\delta).$$

Finally, suppose that the function

$$z \in \mathbb{R}^d \mapsto \int_{\mathbb{R}^d} \widetilde{\psi}(z') \nabla_\theta p(z' \mid \delta(z))\, dz'$$

belongs to $L^2(\mathbb{R}^d : \mathbb{R}^m; \mu)$.

Then the objective function $\mathscr{E}_\tau : \mathbb{L}^2(\mathbb{R}^d : \mathbb{R}^m; \mu) \to \mathbb{R}$ defined in (1.7) is Gateaux differentiable at δ, and its gradient at that point takes the form

$$\nabla \mathscr{E}_\tau(\delta) = 2(\delta(\cdot) - \overline{\theta}(\cdot)) + \frac{1}{2\tau} \int_{\mathbb{R}^d} \widetilde{\psi}(z') \nabla_\theta p(z' \mid \delta(\cdot))\, dz'. \tag{3.1}$$

Proof. Given the form of $\mathscr{E}_\tau$, it suffices to compute the Gateaux derivative of $W_2^2(\mu, \mu_\delta)$ at δ. Let $\eta \in \mathbb{L}^2(\mathbb{R}^d : \mathbb{R}^m; \mu)$ be arbitrary. Taking the derivative of $W_2^2(\mu_{\delta+\epsilon\eta}, \mu)$ w.r.t. ϵ at $\epsilon = 0$, we obtain

$$\begin{aligned}
\frac{d}{d\epsilon}\Big|_{\epsilon=0} W_2^2(\mu_{\delta+\epsilon\eta}, \mu) &= \frac{d}{d\epsilon}\Big|_{\epsilon=0} \sup_{(\phi,\psi)\text{s.t.}\phi(x)+\psi(y)\le|x-y|^2} \int \phi\, d\mu + \int \psi\, d\mu_{\delta+\epsilon\eta} \\
&= \frac{d}{d\epsilon}\Big|_{\epsilon=0} \int \widetilde{\psi}\, d\mu_{\delta+\epsilon\eta} \\
&= \frac{d}{d\epsilon}\Big|_{\epsilon=0} \int \int \widetilde{\psi}(z') p(z' \mid \delta(z) + \epsilon\eta(z))\, dz'\, d\mu(z) \\
&= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \widetilde{\psi}(z') \frac{d}{d\epsilon}\Big|_{\epsilon=0} (p(z' \mid \delta(z) + \epsilon\eta(z)))\, dz'\, d\mu(z) \\
&= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \widetilde{\psi}(z') \langle \nabla_\theta p(z' \mid \delta(z)), \eta(z) \rangle_{\mathbb{R}^m}\, dz'\, d\mu(z) \\
&= \int_{\mathbb{R}^d} \left\langle \eta(z), \int_{\mathbb{R}^d} \widetilde{\psi}(z') \nabla_\theta p(z' \mid \delta(z))\, dz' \right\rangle_{\mathbb{R}^m} d\mu(z).
\end{aligned} \tag{3.2}$$

The second equality follows as in Proposition 7.17 (and Proposition 7.18) in [64] and the third equality just uses the definition of $\mu_{\delta+\epsilon\eta}$. Since η was arbitrary, we deduce (3.1). □

REMARK 6. When given finitely many observations $Z_1, \dots, Z_n$ sampled from μ, the formula in (3.1) suggests the following algorithm to construct a (finite sample) denoising estimator from the observations. In what follows we let

$$\mu_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{Z_i}$$

be the empirical measure of the observations.

Set $k = 0$, and initialize $\delta_k(Z_1), \dots, \delta_k(Z_n) \in \mathbb{R}^m$.

Then do until a stopping criterion is satisfied:

1. Find $\widetilde{\psi}$, optimal dual potential for the 2-OT problem between μ_n and the measure with density $\frac{1}{n}\sum_{i=1}^{n} p(\cdot \mid \delta_k(Z_i))$.
2. Set, for $i = 1, \dots, n$,

$$\delta_{k+1}(Z_i) := \delta_k(Z_i) - \lambda\left(2(\delta_k(Z_i) - \overline{\theta}(Z_i)) + \frac{1}{2\tau}\int \widetilde{\psi}(z')\frac{\nabla_\theta p(z' \mid \delta_k(Z_i))}{p(z' \mid \delta_k(Z_i))} p(z' \mid \delta_k(Z_i))\, dz'\right).$$

3. Set $k = k + 1$.

In the above, $\lambda > 0$ is a time step parameter. Note that when the likelihood model is an exponential family of distributions, we can use Tweedie's formula (see Appendix B) to estimate $\overline{\theta}(Z_i)$. The computation of $\widetilde{\psi}$ can be carried out with an OT solver. We leave it for future work to explore the use of different solvers for computing the gradient of $\mathscr{E}_\tau$ in practical finite data settings.

3.1 *A Kantorovich relaxation of* (1.7) *and recovery of* δ^*

We now turn our attention to studying the existence of solutions to problem (1.7). To achieve this, we first introduce a suitable Kantorovich relaxation of (1.7) for which we can prove existence of solutions using the direct method of the calculus of variations. Under Assumption 8 stated below, we will further characterize the structure of solutions of this relaxation and in particular show that any solution to (3.4) (see below) naturally induces a solution to the original problem (1.7). To define the desired Kantorovich relaxation, let us first introduce the set of admissible couplings

$$\mathscr{A} := \left\{\gamma \in \mathscr{P}(\mathbb{R}^d \times \Omega \times \mathbb{R}^d \times \mathbb{R}^d) : \gamma_1 = \mu, \gamma_4 = \mu, \text{ and } \int p(\cdot \mid \theta)\, d\gamma_2(\theta) = \gamma_3(\cdot)\right\}; \tag{3.3}$$

in the above display by γ_k, for $k = 1, 2, 3, 4$, we mean the kth marginal of γ. We observe that the set $\mathscr{A}$ is determined by μ (the marginal distribution of observed variables) and $\{p(\cdot \mid \theta)\}_{\theta\in\Omega}$ (the likelihood model). We now introduce the problem

$$\inf_{\gamma\in\mathscr{A}} \int \mathbf{c}_\tau(z_1, \theta, z_3, z_4)\, d\gamma(z_1, \theta, z_3, z_4), \tag{3.4}$$

where the cost function $\mathbf{c}_\tau$ is defined as

$$\mathbf{c}_\tau(z_1, \theta, z_3, z_4) := |\theta - \overline{\theta}(z_1)|^2 + \frac{1}{2\tau}|z_3 - z_4|^2. \tag{3.5}$$

REMARK 7 (Comparison with MOT). Problem (3.4) resembles an MOT problem (e.g. see [60]) with four marginals, but differs from a standard MOT in the type of constraint that we put on the second and third marginals of the coupling γ.

We will make the following assumptions on our probabilistic model.

ASSUMPTION 8. We assume that the set Ω is a closed subset of $\mathbb{R}^m$. In addition, we assume that the family of probability measures $\{p(\cdot \mid \theta)\}_{\theta\in\Omega}$ is continuous in θ in the weak sense, i.e. if $\{\theta_n\}_{n\in\mathbb{N}}$ is a sequence in Ω converging to some $\theta \in \Omega$, then $p(\cdot \mid \theta_n)$ converges weakly to $p(\cdot \mid \theta)$.

ASSUMPTION 9. We assume that the posterior mean $\overline{\theta}(z)$ is continuous for μ-a.e. $z \in \mathbb{R}^d$.

REMARK 8 (On the first part of Assumption 8). In order to prove the existence of solutions to problem (3.4) we assume that Ω is a closed subset of $\mathbb{R}^m$ for simplicity. In case Ω is not closed, one can consider modifying the definition of problem (3.4) by changing all appearances of Ω with $\overline{\Omega}$, the closure of Ω. This can be done if we assume that the family $\{p(\cdot \mid \theta)\}_{\theta\in\Omega}$ can be extended to a family of distributions $\{p(\cdot \mid \theta)\}_{\theta\in\overline{\Omega}}$ (not necessarily with densities w.r.t. the Lebesgue measure) for which we still have the weak continuity property: if $\{\theta_n\}_n \subseteq \overline{\Omega}$ and $\theta_n \to \theta$, then $p(\cdot \mid \theta_n)$ converges weakly to $p(\cdot \mid \theta)$. For instance, this can be done in the normal scale mixture problem in Example 2.

REMARK 9 (On Assumption 9). We will also impose Assumption 9 to guarantee the existence of solutions to the relaxation problem (3.4). This assumption is mild and for example is satisfied when $\{p(\cdot \mid \theta)\}_{\theta\in\Omega}$ is an exponential family of distributions under suitable assumptions (see Lemma B2 in Appendix B). Indeed, in this case $\overline{\theta}(\cdot)$ coincides with the gradient of a real-valued convex function. As, by Alexandrov's theorem, a convex function is (Lebesgue) a.e. twice differentiable, its gradient is (Lebesgue) a.e. continuous. Since μ has a density w.r.t. the Lebesgue measure, it then follows that $\overline{\theta}(z)$ is indeed continuous for μ-a.e. $z \in \mathbb{R}^d$.

As stated in the next theorem, problem (3.4) admits minimizers. More importantly, all minimizers of this problem possess a convenient structure that we later use to prove existence of solutions to problem (1.7).

THEOREM 10. Suppose Assumptions 2, 6, 8 and 9 hold. Then there exist solutions to (3.4). Moreover, if γ^* is a solution of (3.4), then γ^*_{12}, the projection of γ onto the first two coordinates, is a solution to the problem

$$\inf_{\pi\in\Gamma(\gamma_1^*,\gamma_2^*)} \int |\theta - \overline{\theta}(z)|^2 \, d\pi(z, \theta).$$

In turn, under Assumption 3, γ^*_{12} must have the form $\gamma^*_{12} = (\mathrm{Id} \times \delta_{\gamma^*})_\sharp \mu$ for $\delta_{\gamma^*} \in \mathbb{L}^2(\mathbb{R}^d : \mathbb{R}^m; \mu)$ the unique solution to the problem:

$$\inf_{\delta:\delta_\sharp\mu=\gamma_2^*} \int |\delta(z) - \overline{\theta}(z)|^2 \, d\mu(z). \tag{3.6}$$

Theorem 10 is proved in Section 5.1, and, as stated earlier, will be used to deduce the existence of solutions of (1.7). Precisely, as we state in Theorem 11 below, the existence of solutions of (1.7) follows from the equivalence between problems (1.7) and (3.4). To describe this equivalence, we introduce some notation first.

Given $\delta \in \mathbb{L}^2(\mathbb{R}^d : \mathbb{R}^m; \mu)$, let π_{34}^{δ} be a 2-OT plan between μ_δ (as defined in (1.8)) and μ. Using π_{34}^{δ}, we define γ_δ as the measure which acts on an arbitrary test function $\phi : \mathbb{R}^d \times \Omega \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ according to

$$\int \phi(z_1, \theta, z_3, z_4)\, d\gamma_\delta(z_1, \theta, z_3, z_4) = \int \phi(z_4, \delta(z_4), z_3, z_4)\, d\pi_{34}^{\delta}(z_3, z_4). \tag{3.7}$$

In simple terms, to sample from γ_δ it is sufficient to sample $(z_3, z_4) \sim \pi_{34}^{\delta}$ and then set $z_1 = z_4$ and $\theta = \delta(z_4)$. Notice that $\gamma_\delta \in \mathscr{A}$. The proof of Theorem 11 below can be found in Section 5.2.

THEOREM 11. Under Assumptions 2, 3, 6,8 and 9 the following properties hold:

1. Let γ^* be any solution to (3.4). Then the map δ_{γ^*} for which $\gamma_{12}^* = (\mathrm{Id} \times \delta_{\gamma^*})_\sharp \mu$ is a solution to (1.7). In particular, due to Theorem 10, there exist solutions to (1.7).
2. Conversely, if $\widetilde{\delta}$ is a solution to (1.7), then $\gamma_{\widetilde{\delta}} \in \mathscr{P}(\mathbb{R}^d \times \Omega \times \mathbb{R}^d \times \mathbb{R}^d)$ defined as in (3.7) for $\delta = \widetilde{\delta}$ is a solution to (3.4).

REMARK 10 (Equivalence between (3.4) and (1.7)). Theorem 11 captures the equivalence between problems (3.4) and (1.7): from a solution γ^* to (3.4) (which exists by the first part of Theorem 10) we can obtain a map δ^* that is a solution to (1.7). Conversely, from a solution to (1.7) we can construct a solution to (3.4). Interestingly, the relaxation (3.4) provides an avenue for designing alternative numerical methods for optimizing (1.7) that do not rely on the gradient descent strategy in the $\mathbb{L}^2(\mathbb{R}^d : \mathbb{R}^m; \mu)$ space suggested at the beginning of Section 3. Notice that (3.4) is a linear optimization problem, which, as discussed in Remark 7, resembles an MOT problem. For this reason we expect to be able to use computational OT techniques to solve (1.7).

REMARK 11. We do not claim uniqueness of solutions of (1.7). This non-uniqueness may not be surprising, since problem (1.7) is in general non-convex in δ.

Next, we discuss the behaviour of solutions to problem (1.7) as the parameter $\tau \to 0$. We show in Theorem 13 (see Section 5.3 for its proof) that under the identifiability assumption stated below, we can recover δ^*, the OT-based denoiser, from the solutions of (1.7).

ASSUMPTION 12. The following identifiability condition on $\{p(\cdot \mid \theta)\}_{\theta \in \Omega}$ holds: If $\int_\Omega p(\cdot \mid \theta)\, dG(\theta) = \int_\Omega p(\cdot \mid \theta)\, dG'(\theta)$ for two probability measures G and G' over Ω, then $G = G'$.

THEOREM 13. Let $\{\tau_n\}_{n \geq 1}$ be a sequence of positive numbers converging to 0. Let δ_n^* be a solution to problem (1.7) with $\tau = \tau_n$ (we know solutions exist due to Theorem 11). Then, under the same assumptions as in Theorem 11 and the additional Assumption 12, δ_n^* converges in $\mathbb{L}^2(\mathbb{R}^d : \mathbb{R}^m; \mu)$ to δ^* as defined in Theorem 4. In other words,

$$\lim_{n \to \infty} \int |\delta_n^*(z) - \delta^*(z)|^2\, d\mu(z) = 0.$$

REMARK 12 (Non-identifiable version of Theorem 13). An inspection of the proof of Theorem 13 reveals that if we drop Assumption 12, then we can conclude that the set of accumulation points of $\{\delta_n^*\}_{n\in\mathbb{N}}$ in the (strong) $\mathbb{L}^2(\mathbb{R}^d : \mathbb{R}^m; \mu)$ topology is contained in the set of minimizers of the problem

$$\min_{\delta\,:\,\mu_\delta=\mu} \mathbb{E}_{(Z,\Theta)\sim P_{Z,\Theta}}\left[|\delta(Z)-\Theta|^2\right].$$

In other words, from the family of problems (1.7) we can find a map δ with the smallest risk attainable within the set of maps that consistently reproduce the distribution of observations μ.

REMARK 13. Theorem 13 suggests taking small values of τ in (1.7) (or in its equivalent formulation (3.4)) to recover the OT-based denoiser. However, we anticipate a certain computational hardness for the optimization problem (1.7) when τ is small. To better appreciate this, observe that small values of τ in the equivalent formulation (3.4) essentially enforce the hard constraint

$$\int p(\cdot|\theta)d\gamma_2(\theta) = \mu,$$

which is equivalent to solving a deconvolution problem.

4. Proofs of main results from Section 2

4.1 *Proof of Theorem* 4

In order to prove Theorem 4, we first present some preliminary results relating solutions of problem (2.7) (with the cost function as in (2.8)) and its dual with solutions of the problem

$$\min_{\pi\in\Gamma(\overline{\theta}_\sharp\mu,G^*)} \int\int |\theta-\vartheta|^2 d\pi(\theta,\vartheta) \tag{4.1}$$

and its dual.

PROPOSITION 14. Let $\widetilde{\pi}$ and $(\widetilde{\phi},\widetilde{\psi})$ be solutions to (4.1) and its dual, respectively. Suppose that Assumption 2 holds. Then (4.1) = (2.7). Furthermore, the functions $(\widetilde{\phi}\circ\overline{\theta},\widetilde{\psi})$ form a solution pair for the dual of (2.7). In addition, the coupling π defined according to

$$d\pi(z,\theta) := d\widetilde{\pi}(\theta \mid \overline{\theta}(z))\, d\mu(z) \tag{4.2}$$

is a solution for (2.7); here, by $\widetilde{\pi}(\cdot \mid \vartheta)$ we mean the conditional distribution of θ given ϑ when $(\theta,\vartheta)\sim\widetilde{\pi}$.

Proof. Using the Kantorovich duality theorem (see Theorem 1.3 in [70]), it follows that

$$\int \widetilde{\phi}(\vartheta)d\overline{\theta}_\sharp\mu(\vartheta) + \int \widetilde{\psi}(\theta)dG^*(\theta) = \int |\theta-\vartheta|^2 d\widetilde{\pi}(\vartheta,\theta).$$

Now, the left-hand side of the above display can be written as

$$\int \widetilde{\phi}(\overline{\theta}(z))d\mu(z) + \int \widetilde{\psi}(\theta)dG^*(\theta),$$

while the right-hand side can be written, using the disintegration theorem, as

$$\int \left(\int |\theta - \vartheta|^2 d\widetilde{\pi}(\theta|\vartheta) \right) d\overline{\theta}_\sharp \mu(\vartheta) = \int \left(\int |\overline{\theta}(z) - \theta|^2 d\widetilde{\pi}(\theta|\overline{\theta}(z)) \right) d\mu(z)$$
$$= \int \int |\overline{\theta}(z) - \theta|^2 d\pi(z, \theta).$$

It follows that

$$\int \widetilde{\phi} \circ \overline{\theta}(z) d\mu(z) + \int \widetilde{\psi}(\theta) dG^*(\theta) = \int |\overline{\theta}(z) - \theta|^2 d\pi(z, \theta),$$

implying that π and $(\widetilde{\phi} \circ \overline{\theta}, \widetilde{\psi})$ are solutions of (2.7) and its dual, respectively. This computation also shows that (4.1) = (2.7), as claimed. □

For the uniqueness statement in Theorem 4 we will establish a converse statement to Proposition 14. Namely, we will prove that any solution to (2.7) must have the form (4.2). We notice that without the additional Assumption 3 this converse statement may fail, as the next remark illustrates.

REMARK 14. In general, a converse statement to Proposition 14 may not be true if Assumption 3 does not hold (i.e. if $\overline{\theta}_\sharp \mu$ is not absolutely continuous w.r.t. the Lebesgue measure), as the following example illustrates. Let G^* be the uniform measure over the set $\Omega := \{0, 1, 3, 4\}$, and for every $\theta \in \Omega$, let $p(\cdot \mid \theta)$ be the uniform distribution on the interval $[0, 1]$. Then, we can see that μ is the uniform distribution on $[0, 1]$ and $\overline{\theta}(\cdot) \equiv 2$, which implies that $\overline{\theta}_\sharp \mu = \delta_2$, the Dirac delta measure at the point 2. Since $\overline{\theta}_\sharp \mu$ is concentrated at a point, there is a unique solution $\widehat{\pi}$ to problem (4.1) (in fact, there is only one coupling between a Dirac delta measure and an arbitrary probability measure). However, as $\overline{\theta}(\cdot)$ is a constant, any coupling between μ =Uniform[0,1] and the uniform distribution on Ω would have the same cost; hence there are actually multiple solutions to problem (2.7).

In what follows we let $\nu \in \mathscr{P}(\mathbb{R}^d \times \mathbb{R}^m)$ be the joint distribution of $(Z, \overline{\theta}(Z))$ where $Z \sim \mu$. We use the disintegration theorem to write ν as

$$d\nu(z, \theta) = d\nu(z \mid \theta)\, d(\overline{\theta}_\sharp \mu)(\theta), \qquad \text{for } \theta \in \mathbb{R}^m, z \in \mathbb{R}^d. \tag{4.3}$$

Notice that the support of $\nu(\cdot \mid \theta)$ can be assumed to be contained in $\{z \in \mathbb{R}^d : \overline{\theta}(z) = \theta\}$.

LEMMA 1. Let $\pi_0 \in \Gamma(\mu, G^*)$ and let $\widehat{\pi} := (\overline{\theta}, \mathrm{Id})_\sharp \pi_0 \in \Gamma(\overline{\theta}_\sharp \mu, G^*)$, where $(\overline{\theta}, \mathrm{Id}) : (z, \vartheta) \mapsto (\overline{\theta}(z), \vartheta)$. Suppose, in addition, that $\widehat{\pi}$ is known to have the form $\widehat{\pi} = (\mathrm{Id} \times T)_\sharp (\overline{\theta}_\sharp \mu)$ for some map T. Then

$$\pi_0(\cdot \mid z) = \delta_{T \circ \overline{\theta}(z)}(\cdot)$$

for μ-a.e. z. In the above, $\pi_0(\cdot \mid z)$ stands for the conditional distribution of θ given z when $(z, \theta) \sim \pi_0$. In particular,

$$\pi_0 = (\mathrm{Id} \times T \circ \overline{\theta})_\sharp \mu.$$

Proof. For $(\theta, \widetilde{\theta}) \sim \widehat{\pi}$ of the form $\widehat{\pi} = (\mathrm{Id}, T)_\sharp(\overline{\theta}_\sharp \mu) \in \Gamma(\overline{\theta}_\sharp \mu, G^*)$ it is clear that

$$\widehat{\pi}(\cdot \mid \theta) = \delta_{T(\theta)} \tag{4.4}$$

for $\overline{\theta}_\sharp \mu$-a.e. θ. On the other hand, from the representation $\widehat{\pi} = (\overline{\theta}, \mathrm{Id})_\sharp \pi_0$, for any bounded and measurable function $\phi : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ we have

$$\begin{aligned}
\int_{\mathbb{R}^m} \int_{\mathbb{R}^m} \phi(\theta, \widetilde{\theta})\, d\widehat{\pi}(\theta, \widetilde{\theta}) &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^m} \phi(\overline{\theta}(z), \widetilde{\theta})\, d\pi_0(z, \widetilde{\theta}) \\
&= \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^m} \phi(\overline{\theta}(z), \widetilde{\theta})\, d\pi_0(\widetilde{\theta} \mid z) \right) d\mu(z) \\
&= \int_{\mathbb{R}^d} \int_{\mathbb{R}^m} \left(\int_{\mathbb{R}^m} \phi(\theta, \widetilde{\theta})\, d\pi_0(\widetilde{\theta} \mid z) \right) d\nu(z, \theta) \\
&= \int_{\mathbb{R}^m} \int_{\mathbb{R}^d} \int_{\mathbb{R}^m} \phi(\theta, \widetilde{\theta})\, d\pi_0(\widetilde{\theta} \mid z)\, d\nu(z \mid \theta)\, d(\overline{\theta}_\sharp \mu)(\theta),
\end{aligned}$$

where we recall that ν is the joint distribution of $(Z, \overline{\theta}(Z))$ for $Z \sim \mu$. From this computation and the uniqueness of conditional distributions in the disintegration theorem it follows that

$$\widehat{\pi}(\cdot \mid \theta) = \int_{\mathbb{R}^d} \pi_0(\cdot \mid z)\, d\nu(z \mid \theta),$$

for $\overline{\theta}_\sharp \mu$-a.e. θ. That is, for any Borel measurable $A \subseteq \mathbb{R}^m$ we have

$$\widehat{\pi}(A \mid \theta) = \int_{\mathbb{R}^d} \pi_0(A \mid z)\, d\nu(z \mid \theta).$$

Combining with (4.4), it follows that for $\overline{\theta}_\sharp \mu$-a.e. θ

$$\delta_{T(\theta)}(\cdot) = \widehat{\pi}(\cdot \mid \theta) = \int_{\mathbb{R}^d} \pi_0(\cdot \mid z)\, d\nu(z \mid \theta).$$

For a θ for which the above is true, we may take the singleton $A = \{T(\theta)\}$ and conclude that

$$1 = \int_{\mathbb{R}^d} \pi_0(A \mid z)\, d\nu(z \mid \theta),$$

which implies that $\pi_0(A \mid z) = 1$ for $\nu(\cdot \mid \theta)$-a.e. z. That is,

$$\pi_0(\cdot \mid z) = \delta_{T(\theta)}.$$

for $\nu(\cdot \mid \theta)$-a.e. z. Finally, as discussed right after (4.3), for z in the support of $\nu(\cdot \mid \theta)$ we have $\theta = \overline{\theta}(z)$. It then follows that for $\nu(\cdot \mid \theta)$-a.e. z we have

$$\pi_0(\cdot \mid z) = \delta_{T \circ \overline{\theta}(z)}.$$

At this stage we can apply Fubini's theorem to conclude that

$$\pi_0(\cdot \mid z) = \delta_{T \circ \overline{\theta}(z)}$$

for μ-a.e. $z \in \mathbb{R}^d$, completing in this way the proof. □

We are now ready to prove Theorem 4.

Proof of Theorem 4. Under the assumption that $\overline{\theta}_\sharp \mu$ is absolutely continuous w.r.t. the Lebesgue measure, we can use Brenier's theorem (Theorem 1) to deduce that there exists a unique solution $\widetilde{\pi}$ to (4.1), which has the form

$$\widetilde{\pi} = (\mathrm{Id} \times T)_\sharp (\overline{\theta}_\sharp \mu)$$

for some measurable map T of the form $T = \nabla \varphi$ for a convex function φ; existence of solutions to the dual of (4.1) is guaranteed by Theorem 2.12 in [69]. Further, from Brenier's theorem we also know that $T_\sharp(\overline{\theta}_\sharp \mu) = G^*$ and that T minimizes the objective (2.10). Proposition 14 then implies that

$$\pi := (\mathrm{Id} \times T \circ \overline{\theta})_\sharp \mu$$

is a solution of (2.7).

It remains to show that the obtained solution to (2.7) is unique. To see this, suppose that π_0 is a solution of (2.7), and let $\widehat{\pi} := (\overline{\theta}, \mathrm{Id})_\sharp \pi_0$. It follows that

$$\int |\theta - \vartheta|^2 \, d\widehat{\pi}(\theta, \vartheta) = \int |\overline{\theta}(z) - \vartheta|^2 \, d\pi_0(z, \vartheta) = (2.7) = (4.1),$$

and thus $\widehat{\pi}$ is a solution of (4.1); notice that the latter of the above equalities follows from Proposition 14. From this and the uniqueness of solutions to (4.1), by Assumption 3 as $\overline{\theta}_\sharp \mu$ is absolutely continuous w.r.t. the Lebesgue measure, it follows that $\widehat{\pi} = \widetilde{\pi}$. Using the fact that $\widehat{\pi} = (\overline{\theta} \times \mathrm{Id})_\sharp \pi_0 = (\mathrm{Id} \times T)_\sharp (\overline{\theta}_\sharp \mu)$ in Lemma 1 we can conclude that necessarily $\pi_0 = (\mathrm{Id} \times T \circ \overline{\theta})_\sharp \mu$, proving in this way the uniqueness of solutions to (2.7). □

Remark 15. Suppose that the Bayes estimator $\overline{\theta}$ satisfies Assumption 2 and 3. Then it can be easily seen that the above proof can be used to deduce that, for any $G \in \mathscr{P}(\Omega)$ with finite second moments (not necessarily equal to the prior G^*), the problem

$$\inf_{\pi \in \Gamma(\mu, G)} \int |\theta - \overline{\theta}(z)|^2 \, d\pi(z, \theta)$$

has a unique solution $\widetilde{\pi}$. This unique solution takes the form

$$\widetilde{\pi} = (\mathrm{Id} \times \widetilde{\delta})_\sharp \mu,$$

for $\widetilde{\delta}$ the unique solution to the problem

$$\inf_{\delta : \delta_\sharp \mu = G} \mathbb{E}_{Z\sim\mu}[|\overline{\theta}(Z) - \delta(Z)|^2].$$

4.2 *Proof of Theorem* 5

In order to prove Theorem 5 we begin by relaxing (1.6) as follows:

$$\inf_{\pi,\widetilde{\pi}} \int |\overline{\theta}(z) - \widetilde{\theta}|^2 \, d\pi(z,\widetilde{\theta}) + \frac{1}{2\tau} \int |\widetilde{\theta} - \theta|^2 \, d\widetilde{\pi}(\widetilde{\theta},\theta), \tag{4.5}$$

where the inf is taken over pairs $(\pi, \widetilde{\pi})$ satisfying: $\pi \in \mathscr{P}(\mathbb{R}^d \times \mathbb{R}^m)$, $\widetilde{\pi} \in \mathscr{P}(\mathbb{R}^m \times \mathbb{R}^m)$, $\pi_1 = \mu$, $\widetilde{\pi}_2 = G^*$ and $\pi_2 = \widetilde{\pi}_1$. We will characterize solutions to (4.5) following the proof of a theorem in [1]. We will then relate these solutions with problem (1.6) and with the characterization given in the statement of Theorem 5.

Lemma 2. Let $\tau > 0$. Then problem (4.5) is equivalent to problem

$$\min_{\gamma \in \Gamma(\mu, G^*)} \int B(z,\theta) d\gamma(z,\theta), \tag{4.6}$$

where $B(\cdot,\cdot)$ is the *barycentre cost*:

$$B(z,\theta) := \min_{\widetilde{\theta}\in\mathbb{R}^m} \left\{ |\overline{\theta}(z) - \widetilde{\theta}|^2 + \frac{1}{2\tau}|\widetilde{\theta} - \theta|^2 \right\} = \frac{1}{(1+2\tau)} |\overline{\theta}(z) - \theta|^2.$$

Proof. Let $\gamma^* \in \Gamma(\mu, G^*)$ be a solution to (4.6) (note that a solution to (4.6) indeed exists). For a given (z,θ) in the support of γ^* we consider

$$T(z,\theta) := \operatorname*{argmin}_{\widetilde{\theta}\in\mathbb{R}^m} \left\{ |\overline{\theta}(z) - \widetilde{\theta}|^2 + \frac{1}{2\tau}|\widetilde{\theta} - \theta|^2 \right\} = \frac{2\tau}{1+2\tau}\overline{\theta}(z) + \frac{1}{2\tau+1}\theta. \tag{4.7}$$

Let $\nu \in \mathscr{P}(\mathbb{R}^d \times \Omega \times \mathbb{R}^m)$ be given by

$$\nu := (\mathrm{Id}, T)_\sharp \gamma^*,$$

where (Id, T) is the map $(\mathrm{Id}, T) : (z,\theta) \in \mathbb{R}^d \times \Omega \mapsto (z, \theta, T(z,\theta))$, and let

$$\pi^* := P_{13\sharp}\nu, \quad \widetilde{\pi}^* := P_{32\sharp}\nu,$$

where $P_{13}(z,\theta,\widetilde{\theta}) = (z,\widetilde{\theta})$ and $P_{32}(z,\theta,\widetilde{\theta}) = (\widetilde{\theta},\theta)$. Notice that $(\pi^*,\widetilde{\pi}^*)$ is a feasible pair for (4.5). For this pair we have

$$\begin{aligned}(4.6) \;=\; \int B(z,\theta)\,d\gamma^*(z,\theta) &= \int \left(|\overline{\theta}(z) - T(z,\theta)|^2 + \frac{1}{2\tau}|T(z,\theta)-\theta|^2\right) d\gamma^*(z,\theta)\\ &= \int |\overline{\theta}(z)-\widetilde{\theta}|^2 d\pi^*(z,\widetilde{\theta}) + \frac{1}{2\tau}\int |\widetilde{\theta}-\theta|^2 d\widetilde{\pi}^*(\widetilde{\theta},\theta)\\ &\geq (4.5).\end{aligned} \tag{4.8}$$

Let us now consider an arbitrary feasible pair $(\pi,\widetilde{\pi})$ for (4.5). From $(\pi,\widetilde{\pi})$ we can construct $\gamma \in \Gamma(\mu,G^*)$ by glueing π and $\widetilde{\pi}$ together as we describe next. Let $\nu_0 = \pi_2 = \widetilde{\pi}_1$. Then, by the disintegration theorem applied to π and $\widetilde{\pi}$, we can decompose π and $\widetilde{\pi}$ in terms of conditionals relative to one of their marginals (in this case ν_0):

$$d\pi(z,\widetilde{\theta}) = d\pi_{z|\widetilde{\theta}}(z)d\nu_0(\widetilde{\theta}), \qquad d\widetilde{\pi}(\widetilde{\theta},\theta) = d\pi_{\theta|\widetilde{\theta}}(\theta)d\nu_0(\widetilde{\theta}).$$

Using these decompositions, we define $\gamma \in \mathscr{P}(\mathbb{R}^d \times \Omega)$ as the probability measure acting on smooth test functions φ according to

$$\int_{\mathbb{R}^d}\int_{\Omega} \varphi(z,\theta)d\gamma(z,\theta) = \int_{\mathbb{R}^m}\left(\int_{\mathbb{R}^d}\int_{\Omega}\varphi(z,\theta)d\pi_{z|\widetilde{\theta}}(z)d\pi_{\theta|\widetilde{\theta}}(\theta)\right)d\nu_0(\widetilde{\theta}).$$

To intuitively explain the joint distribution $(Z,\Theta)\sim\gamma$ above, we consider a joint distribution on three variables $(Z,\Theta,\widetilde{\Theta})$ defined as follows: $\widetilde{\Theta}\sim\nu_0$, Z and Θ are independent given $\widetilde{\Theta}$, with $Z \mid \widetilde{\Theta}\sim\pi_{z|\widetilde{\theta}}$, and $\Theta \mid \widetilde{\Theta}\sim\pi_{\theta|\widetilde{\theta}}$. Thus, γ is the joint distribution of (Z,Θ) according to the above model. It is straightforward to check that $\gamma\in\Gamma(\mu,G^*)$. Moreover, we have the following:

$$\begin{aligned}&\int |\overline{\theta}(z)-\widetilde{\theta}|^2\,d\pi(z,\widetilde{\theta}) + \frac{1}{2\tau}\int|\widetilde{\theta}-\theta|^2\,d\widetilde{\pi}(\widetilde{\theta},\theta)\\ &\quad= \int\int |\overline{\theta}(z)-\widetilde{\theta}|^2\,d\pi_{z|\widetilde{\theta}}(z)\,d\nu_0(\widetilde{\theta}) + \frac{1}{2\tau}\int\int|\widetilde{\theta}-\theta|^2\,d\widetilde{\pi}_{\theta|\widetilde{\theta}}(\theta)\,d\nu_0(\widetilde{\theta})\\ &\quad= \int\left[\int\int\left(|\overline{\theta}(z)-\widetilde{\theta}|^2 + \frac{1}{2\tau}|\widetilde{\theta}-\theta|^2\right)d\pi_{z|\widetilde{\theta}}(z)d\widetilde{\pi}_{\theta|\widetilde{\theta}}(\theta)\right]d\nu_0(\widetilde{\theta})\\ &\quad\geq \int\left(\int\int B(z,\theta)d\pi_{z|\widetilde{\theta}}(z)\,d\widetilde{\pi}_{\theta|\widetilde{\theta}}(\theta)\right)d\nu_0(\widetilde{\theta})\\ &\quad= \int\int B(z,\theta)\,d\gamma(z,\theta) \;\geq\; (4.6).\end{aligned} \tag{4.9}$$

Since the above is true for any arbitrary feasible pair $(\pi,\widetilde{\pi})$, we deduce that (4.5) $\geq$ (4.6). Combining with (4.8) we obtain the equality.

From the equality (4.5) = (4.6), we see from (4.8) and (4.9) that there is an explicit way to map solutions γ^* of (4.6) to solutions $(\pi^*,\widetilde{\pi}^*)$ of (4.5) and vice versa. □

LEMMA 3. For any fixed $\tau > 0$ problem (4.6) is equivalent to (2.7). In particular, its unique solution γ^* has the form

$$\gamma^* = (\mathrm{Id} \times \delta^*)_\sharp \mu$$

for δ^* as defined in (2.9).

Proof. Notice that a direct computation reveals, from Lemma 2, that for all $z \in \mathbb{R}^d$ and $\theta \in \Omega$, $B(z,\theta) = (1+2\tau)^{-1}|\overline{\theta}(z) - \theta|^2$. Therefore, problem (4.6) is equivalent to problem (2.7). □

LEMMA 4. Problem (4.5) has a unique solution, which is given by

$$\pi^* = F_\sharp \mu, \qquad \widetilde{\pi}^* = \widetilde{F}_\sharp \mu, \tag{4.10}$$

where $F : z \in \mathbb{R}^d \mapsto (z, \delta_\tau^*(z))$ and $\widetilde{F} : z \in \mathbb{R}^d \mapsto (\delta_\tau^*(z), \delta^*(z))$. Here δ_τ^* is defined via (2.12).

Proof. First, notice that from the proof of Lemma 2 we know that using $\gamma^* = (\mathrm{Id} \times \delta^*)_\sharp \mu$ we can construct a solution $(\pi, \widetilde{\pi})$ of (4.5) according to

$$\pi = P_{13\sharp}((\mathrm{Id}, T)_\sharp \gamma^*), \quad \widetilde{\pi} = P_{32\sharp}((\mathrm{Id}, T)_\sharp \gamma^*).$$

Recall $T(\cdot,\cdot)$ from (4.7) and note that $T(z, \delta(z)) = \delta_\tau^*(z)$. Using the form of γ^*, it is straightforward to verify that π and $\widetilde{\pi}$ defined above have the form in (4.10). It remains to show that this solution is unique.

To see this, let $(\pi^*, \widetilde{\pi}^*)$ be an arbitrary solution to (4.5). Let Υ be the probability measure over $\mathbb{R}^d \times \mathbb{R}^m \times \Omega$ defined by

$$\int \psi(z, \widetilde{\theta}, \theta)\, d\Upsilon(z, \widetilde{\theta}, \theta) = \int\int\int \psi(z, \widetilde{\theta}, \theta)\, d\pi^*_{z|\widetilde{\theta}}(z)\, d\widetilde{\pi}^*_{\theta|\widetilde{\theta}}(\theta)\, d\nu_0(\widetilde{\theta}),$$

for all smooth test functions ψ; here recall the definitions of $\pi^*_{z|\widetilde{\theta}}(\cdot)$, $\widetilde{\pi}^*_{\theta|\widetilde{\theta}}(\cdot)$ and ν_0 from the proof of Lemma 2. Using (4.9) and the fact that the following inequality is actually an equality (and that both are integrals w.r.t. the measure Υ), we can deduce that for Υ-a.e. $(z, \widetilde{\theta}, \theta)$ we have $\widetilde{\theta} = T(z, \theta)$ (as the integrands must be a.e. equal; cf. (4.7)). From this it follows that the joint distribution Υ is determined by the joint distribution of (z, θ) and the other variable $\widetilde{\theta}$ is a deterministic function of (z, θ), i.e. $\Upsilon = H_\sharp(P_{13\sharp}\Upsilon)$, where $H : (z, \theta) \mapsto (z, T(z, \theta), \theta)$. From (4.9) we can also deduce that $P_{13\sharp}\Upsilon$ is a solution of (4.6), which by Lemma 3 must be equal to $(\mathrm{Id} \times \delta^*)_\sharp \mu$. Therefore,

$$\Upsilon = H_\sharp((\mathrm{Id} \times \delta^*)_\sharp \mu).$$

From this and the fact that by construction we have $\pi^* = P_{12\sharp}\Upsilon$ and $\widetilde{\pi}^* = P_{23\sharp}\Upsilon$ it follows that π^* and $\widetilde{\pi}^*$ are as in (4.10). □

Proof of Theorem 5. Let $\delta : \mathbb{R}^d \to \mathbb{R}^m$ be an arbitrary measurable map. Let $\widetilde{\pi}$ be a 2-OT transport plan between $\delta_\sharp \mu$ and G^*, and let $\pi = (\mathrm{Id}, \delta)_\sharp \mu$. We see that $(\pi, \widetilde{\pi})$ is a feasible pair for (4.5) and that

$$\mathbb{E}_{Z\sim\mu}[|\delta(Z) - \overline{\theta}(Z)|] + \frac{1}{2\tau} W_2^2(\delta_\sharp \mu, G^*) = \int |\widetilde{\theta} - \overline{\theta}(z)|^2\, d\pi(z, \widetilde{\theta}) + \frac{1}{2\tau} \int |\widetilde{\theta} - \theta|^2\, d\widetilde{\pi}(\widetilde{\theta}, \theta).$$

From the above and the form of the unique solution to (4.5) deduced in Lemma 4 it follows that problem (1.6) admits a unique solution, which must have the form (2.12). □

5. Proofs of main results from Section 3

5.1 *Proof of Theorem* 10

Proof. First we establish the existence of solutions to (3.4). Let $\{\gamma^n\}_{n\in\mathbb{N}} \subseteq \mathscr{A}$ be a minimizing sequence for the objective function in (3.4); we recall that $\mathscr{A}$, defined in (3.3), is the feasible set for problem (3.4). In particular, we suppose that

$$\lim_{n\to\infty} \int \mathbf{c}_\tau d\gamma^n = \inf_{\gamma\in\mathscr{A}} \int \mathbf{c}_\tau d\gamma =: M_0 < +\infty.$$

The fact that M_0 is finite follows from the fact that we can take the coupling $\gamma = F_\sharp P_{Z,\Theta}$ (recall $P_{Z,\Theta}$ is the joint distribution of Z and Θ) with $F(z,\theta) := (z,\theta,z,z)$, for which one can see (by Assumption 2) that $\int \mathbf{c}_\tau d\gamma < +\infty$ and $\gamma \in \mathscr{A}$. Without the loss of generality we assume that

$$\sup_{n\in\mathbb{N}} \int \mathbf{c}_\tau d\gamma^n \leq 2M_0.$$

First, we prove that the sequence $\{\gamma^n\}_{n\in\mathbb{N}}$ is precompact in the weak sense. By Prokhorov's theorem it suffices to prove that the sequence $\{\gamma^n\}_{n\in\mathbb{N}}$ is tight. To see this, notice that

$$\int |\theta|^2 d\gamma_2^n(\theta) \leq 2\int |\theta - \overline{\theta}(z_1)|^2 d\gamma^n(z_1,\theta,z_3,z_4) + 2\int |\overline{\theta}(z)|^2 d\mu(z) \leq 4M_0 + 2\int |\overline{\theta}(z)|^2 d\mu(z),$$

which follows from the elementary pointwise inequality $|\theta|^2 \leq 2|\theta-\overline{\theta}(z)|^2 + 2|\overline{\theta}(z)|^2$ and a subsequent integration w.r.t. γ^n on both sides. Likewise,

$$\int |z_3|^2 d\gamma_3^n(z_3) \leq 2\int |z_3 - z_4|^2 d\gamma^n(z_1,\theta,z_3,z_4) + 2\int |z_4|^2 d\mu(z_4) \leq 8\tau M_0 + 2\int |z_4|^2 d\mu(z_4).$$

From the above we can conclude that all second moments of the family of distributions $\{\gamma^n\}_{n\in\mathbb{N}}$ are uniformly bounded, and thus the family $\{\gamma^n\}_{n\in\mathbb{N}}$ is indeed tight. It follows that, up to the extraction of a subsequence that is not relabelled, γ^n converges weakly, as $n\to\infty$, towards a limit that we will denote by γ^*.

Next we show that the limiting γ^* must be feasible for (3.4), i.e. it must belong to the feasible set $\mathscr{A}$. First, observe that $\gamma_1^* = \gamma_4^* = \mu$ follows from the weak convergence of γ^n towards γ^* and the fact that for all $n\in\mathbb{N}$ we have $\gamma_1^n = \gamma_4^n = \mu$. To check that $\int p(\cdot \mid \theta) d\gamma_2^*(\theta) = \gamma_3^*(\cdot)$, and thus conclude that $\gamma^* \in \mathscr{A}$, it is sufficient to show that

$$\int\int \varphi(z) p(z \mid \theta)\, dz\, d\gamma_2^*(\theta) = \int \varphi(z)\, d\gamma_3^*(z) \tag{5.1}$$

for all $\varphi \in C_b(\mathbb{R}^d)$ (here $C_b(\mathbb{R}^d)$ is the set of all bounded continuous functions from $\mathbb{R}^d$ to $\mathbb{R}$). To see this, first notice that

$$\int \int \varphi(z)p(dz|\theta)d\gamma_2^*(\theta) = \int \left(\int \varphi(z)p(dz|\theta) \right) d\gamma_2^*(\theta) = \lim_{n\to\infty} \int \left(\int \varphi(z)p(dz|\theta) \right) d\gamma_2^n(\theta),$$

which follows from the fact that the function $\theta \in \Omega \mapsto \int \varphi(z)p(dz|\theta)$ belongs to $C_b(\Omega)$ by Assumption 8 and the fact that φ is bounded and the weak convergence of γ_2^n towards γ_2^* as $n \to \infty$. On the other hand, the fact that $\gamma^n \in \mathscr{A}$ for all n and the weak convergence of γ_3^n to γ_3^* imply that

$$\lim_{n\to\infty} \int \left(\int \varphi(z)p(dz|\theta) \right) d\gamma_2^n(\theta) = \lim_{n\to\infty} \int \varphi(z)d\gamma_3^n(z) = \int \varphi(z)d\gamma_3^*(z).$$

Combining these identities we deduce (5.1).

To show that γ^* is a solution of (3.4), we start by noticing that, due to Assumption 9, there is a set $B \subseteq \mathbb{R}^d$ with $\mu(B) = 1$ in which the function $\overline{\theta}(\cdot)$ is continuous. Let $\mathscr{B} := B \times \Omega \times \mathbb{R}^d \times \mathbb{R}^d$, and notice that

$$\gamma_n(\mathscr{B}) = \gamma^*(\mathscr{B}) = 1, \quad \forall n \in \mathbb{N},$$

since the first marginals of the γ_n and γ^* are all equal to μ. We deduce that the function $\mathbf{c}_\tau$ is continuous in $\mathscr{B}$. In addition, $\mathbf{c}_\tau$ is lower bounded by a constant (because it is non-negative). We can thus invoke Proposition 5.1.10 in [2] and from the weak convergence of γ_n towards γ^* deduce that

$$M_0 \leq \int \mathbf{c}_\tau \, d\gamma^* \leq \liminf_{n\to\infty} \int \mathbf{c}_\tau \, d\gamma^n = M_0.$$

In the first inequality above we just use the fact that M_0 is the infimum over all couplings.

Next, we discuss the structure of solutions γ^* of (3.4). Consider an arbitrary $\gamma \in \mathscr{A}$ and let π_{12} be optimal for the problem

$$\min_{\pi \in \Gamma(\gamma_1,\gamma_2)} \int |\theta - \overline{\theta}(z)|^2 \, d\pi(z,\theta)$$

and let π_{34} be optimal for the 2-OT problem

$$\min_{\pi \in \Gamma(\gamma_3,\gamma_4)} \int |z_3 - z_4|^2 \, d\pi(z_3,z_4).$$

We define $\widetilde{\gamma} := \pi_{12} \otimes \pi_{34}$, i.e. $\widetilde{\gamma}$ is the product measure between π_{12} and π_{34}. Since $\widetilde{\gamma}_l = \gamma_l$ for all $l = 1, 2, 3, 4$, it follows that $\widetilde{\gamma} \in \mathscr{A}$ as well. Moreover, due to the fact that the cost $\mathbf{c}_\tau$ is the sum of two

terms without shared variables, we have

$$\begin{aligned}\int \mathbf{c}_\tau(z_1,\theta,z_3,z_4)\, d\widetilde{\gamma}(z_1,\theta,z_3,z_4) &= \int |\theta-\overline{\theta}(z_1)|^2 d\pi_{12}(z_1,\theta) + \frac{1}{2\tau}\int |z_3-z_4|^2 d\pi_{34}(z_3,z_4)\\ &\leq \int |\theta-\overline{\theta}(z_1)|^2 d\gamma_{12}(z_1,\theta) + \frac{1}{2\tau}\int |z_3-z_4|^2 d\gamma_{34}(z_3,z_4)\\ &= \int \mathbf{c}_\tau(z_1,\theta,z_3,z_4) d\gamma(z_1,\theta,z_3,z_4).\end{aligned}$$

From the above display we conclude that if $\gamma=\gamma^*$ is a solution to (3.4), then the inequality above must in fact be an equality, and thus we necessarily have

$$\begin{aligned}\int |\theta-\overline{\theta}(z_1)|^2 d\pi_{12}(z_1,\theta) &= \int |\theta-\overline{\theta}(z_1)|^2 d\gamma_{12}(z_1,\theta),\\ \int |z_3-z_4|^2 d\pi_{34}(z_3,z_4) &= \int |z_3-z_4|^2 d\gamma_{34}(z_3,z_4).\end{aligned}$$

This in particular implies that γ_{12}^* is a solution of (3.6) and γ_{34}^* is a 2-OT plan between γ_3^* and γ_4^*. The specific form $(\mathrm{Id}\times\delta)_\sharp\mu$ for γ_{12}^* under Assumption 3 follows from Remark 15. □

Next we present the proof of Theorem 11, which implies the existence of solutions to (1.7) for arbitrary τ.

5.2 *Proof of Theorem* 11

Proof. Let $\delta\in\mathbb{L}^2(\mathbb{R}^d:\mathbb{R}^m;\mu)$, and consider the associated measure γ_δ to δ as defined in (3.7). Then

$$\begin{aligned}(3.4)\leq \int \mathbf{c}_\tau d\gamma_\delta &= \mathbb{E}_{Z\sim\mu}[|\delta(Z)-\overline{\theta}(Z)|^2] + \frac{1}{2\tau}\int |z_3-z_4|^2\, d\pi_{34}^\delta(z_3,z_4)\\ &= \mathbb{E}_{Z\sim\mu}[|\delta(Z)-\overline{\theta}(Z)|^2] + \frac{1}{2\tau}W_2^2(\mu_\delta,\mu).\end{aligned} \tag{5.2}$$

Since δ was arbitrary, the above implies that (3.4) $\leq$ (1.7) $-R_{\mathrm{Bayes}}$, where we recall R_{Bayes} is the Bayes risk (see the beginning of Section 2.2).

Let γ^* be a solution to problem (3.4). By Theorem 10, γ_{12}^* can be written as $\gamma_{12}^*=(\mathrm{Id}\times\delta_{\gamma^*})_\sharp\mu$ for some $\delta_{\gamma^*}\in\mathbb{L}^2(\mathbb{R}^d:\mathbb{R}^m;\mu)$. In particular, $\gamma_2^*=\delta_{\gamma^*\sharp}\mu$ and thus also $\gamma_3=\mu_{\delta_{\gamma^*}}$. From the proof of Theorem 10 we further deduce that

$$\begin{aligned}(3.4)=\int \mathbf{c}_\tau d\gamma^* &= \int |\theta-\overline{\theta}(z_1)|^2 d\gamma_{12}^*(z_1,\theta) + \frac{1}{2\tau}\int |z_3-z_4|^2 d\gamma_{34}^*(z_3,z_4)\\ &= \mathbb{E}_{Z\sim\mu}[|\delta_{\gamma^*}(Z)-\overline{\theta}(Z)|^2] + \frac{1}{2\tau}W_2^2(\mu_{\delta_{\gamma^*}},\mu)\\ &\geq (1.7) - R_{\mathrm{Bayes}}.\end{aligned} \tag{5.3}$$

Combing the above two inequalities we deduce that (3.4) = (1.7) $-R_{\text{Bayes}}$ and that in (5.3) the inequality is actually an equality. In particular, δ_{γ^*} is a solution to (1.7).

Conversely, now that we know that (3.4) = (1.7) $-R_{\text{Bayes}}$, (5.2) implies that if δ is optimal for (1.7), then γ_δ, as defined in (3.7), is optimal for (3.4). □

5.3 *Proof of Theorem* 13

Proof. Let $\{\tau_n\}_{n\in\mathbb{N}}$ be a sequence of positive numbers converging to 0. Let δ_n^* be a solution of problem (1.7) for $\tau = \tau_n$. Due to the second part of Theorem 11, the measure $\gamma_{\delta_n^*}$ associated with δ_n^* that was defined in (3.7) is a solution for the problem (3.4) with $\tau = \tau_n$. In what follows, we use γ^n to denote $\gamma_{\delta_n^*}$ in order to make the notation less cumbersome.

Using similar arguments to those in the first part of the proof of Theorem 10, we can show that $\{\gamma^n\}_{n\in\mathbb{N}}$ is precompact in the weak topology of probability measures and that all its possible accumulation points are in $\mathscr{A}$. Let us then take a subsequence of $\{\gamma^n\}_{n\in\mathbb{N}}$ that converges weakly to some $\gamma \in \mathscr{A}$. For notational simplicity let us denote the subsequence also by $\{\gamma^n\}_{n\in\mathbb{N}}$. We will characterize γ_{12}, the projection of γ onto the first two coordinates.

First, observe that, since $\tau_n \to 0$, as $\int \mathbf{c}_{\tau_n} d\gamma^n$ is bounded from above (see the initials steps in the proof of Theorem 5.1),

$$W_2^2(\gamma_3^n, \mu) \le \int |z_3 - z_4|^2 \, d\gamma^n(z_3, z_4) \le 2\tau_n \int \mathbf{c}_{\tau_n} d\gamma^n \to 0.$$

In particular, γ_3, which is the limit of γ_3^n, must be equal to μ. By Assumption 12, we deduce from $\int p(\cdot|\theta)\, d\gamma_2(\theta) = \gamma_3(\cdot) = \mu(\cdot) = \int p(\cdot|\theta)\, dG^*(\theta)$ that $\gamma_2 = G^*$. Therefore, $\gamma_{12} \in \Gamma(\mu, G^*)$. On the other hand, by weak convergence of γ^n to γ (and Assumption 9) we get

$$\begin{aligned} \int |\theta - \overline{\theta}(z)|^2 \, d\gamma_{12}(z,\theta) &\le \liminf_{n\to\infty} \int |\theta - \overline{\theta}(z)|^2 \, d\gamma_{12}^n(z,\theta) \\ &\le \liminf_{n\to\infty} \int \mathbf{c}_{\tau_n} d\gamma^n. \end{aligned} \tag{5.4}$$

Now, an arbitrary $\pi_{12} \in \Gamma(\mu, G^*)$ induces a $\widetilde{\gamma} \in \mathscr{A}$ as follows:

$$\widetilde{\gamma} := \pi_{12} \otimes (\mathrm{Id} \times \mathrm{Id})_\sharp \mu,$$

and as can be easily verified we have

$$\int \mathbf{c}_{\tau_n} d\widetilde{\gamma} = \int |\theta - \overline{\theta}(z)|^2 \, d\pi_{12}(z,\theta).$$

Since γ^n is optimal for (3.4) with $\tau = \tau_n$, it follows that

$$\int \mathbf{c}_{\tau_n} \, d\gamma^n \le \int \mathbf{c}_{\tau_n} d\widetilde{\gamma} = \int |\theta - \overline{\theta}(z)|^2 d\pi_{12}(z,\theta).$$

Taking lim inf on both sides of the above inequality, combining with (5.4), and using the fact that π_{12} was arbitrary, we deduce that γ_{12} is a solution to (2.7). Theorem 4 thus implies that $\gamma_{12} = \pi^* = (\mathrm{Id} \times \delta^*)_\sharp \mu$, where δ^* is our OT-based denoising estimand. We have thus shown that any convergent subsequence of

the original $\{\gamma^n\}_{n\in\mathbb{N}}$ converges to the same limit point $\pi^* := (\mathrm{Id}\times\delta^*)_\sharp\mu$, and as a consequence the original sequence also converges to this same limit point.

At this stage we may use a series of results from functional analysis and measure theory that we present in Appendix C to deduce that $\delta_n^* \to_{\mathbb{L}^2(\mathbb{R}^d;\mathbb{R}^m;\mu)} \delta^*$. Indeed, first notice that Lemma C1 implies that δ_n^* converges to δ^* in μ-measure (see definition in the statement of Lemma C1). In addition, since we also have

$$\sup_{n\in\mathbb{N}} \int |\delta_n^*(z)|^2 d\mu(z) < \infty,$$

as can be easily verified, we can invoke Lemma C2 to conclude that δ_n^* converges weakly in $\mathbb{L}^2(\mathbb{R}^d : \mathbb{R}^m;\mu)$ to δ^* (see Definition C1). In particular, we have

$$\lim_{n\to\infty} \int \delta_n^*(z)\cdot\overline{\theta}(z)\, d\mu(z) = \int \delta^*(z)\cdot\overline{\theta}(z)\, d\mu(z).$$

Since in addition we have

$$\lim_{n\to\infty} \int |\delta_n^*(z)-\overline{\theta}(z)|^2\, d\mu(z) = \int |\delta^*(z)-\overline{\theta}(z)|^2\, d\mu(z),$$

after expanding the square we conclude that

$$\lim_{n\to\infty} \int |\delta_n^*(z)|^2 d\mu(z) = \int |\delta^*(z)|^2 d\mu(z).$$

Lemma C3 now implies that δ_n^* converges in $\mathbb{L}^2(\mathbb{R}^d : \mathbb{R}^m;\mu)$ to δ^*, as we wanted to prove. □

6. Discussion and future work

In this paper we have presented a new perspective on the denoising problem—where one observes Z (from model (1.1)) and the goal is to predict the underlying latent variable $\Theta \sim G^*$—based on OT theory. We define the OT-based denoiser $\delta^*(Z)$ as the function which minimizers the Bayes risk in this problem subject to the distributional stability constraint $\delta^*(Z) \sim G^*$. Moreover, we have developed two approaches to characterize this OT-based denoiser $\delta^*(Z)$, one where we explicitly use G^* (Section 2) and one where we directly involve μ (the marginal distribution of Z) and the likelihood model $\{p(\cdot \mid \theta)\}_{\theta\in\Omega}$ without an explicit use of the prior G^* (Section 3).

One important direction that we believe is worth investigating in future work is the numerical implementation of our proposals in the finite data setting. In Appendix E we outline an approach to implementing the sample version of the Kantorovich relaxation problem (2.7) (by directly plugging in an estimator of G^*) which would lead to an estimator of δ^* (cf. (2.9)). We conjecture that this approach would yield a consistent estimator of δ^* and it would be interesting to study its rate of convergence.

The adaptation of our approach in Section 3 to the finite data setting to find a solution to (1.7) can, in principle, avoid direct estimation of G^*. Here the key challenge is to find a suitable sample version of the Kantorovich relaxation problem (3.4) (which under appropriate conditions yields a solution to (1.7); see Theorem 11). Indeed, in contrast to the gradient descent approach outlined in (6) for solving (1.7) in the finite data setting, the Kantorovich relaxation (3.4) is a linear program whose optimizers are guaranteed to induce global solutions to (1.7). However, the first hurdle in developing an empirical version of

(3.4) is to estimate the cost function $\mathbf{c}_\tau$ in (3.5) which involves the Bayes estimator $\overline{\theta}(\cdot)$. This is where Tweedie's formula (see (B.5) in Appendix B) can be very useful. It expresses the posterior mean $\overline{\theta}(\cdot)$ in an exponential family model (see Appendix A) in terms of the marginal density f_{G^*} of the observations (and its gradient) that can be estimated (non-parametrically) directly from the sample $Z_1, \ldots, Z_n$, say via kernel density estimation. Thus, Tweedie's formula can yield an estimated cost function without directly estimating the unknown prior G^*. The next step would be to solve problem (3.4) with this estimated cost. As problem (3.4) is closely reminiscent of a MOT problem we expect that some adaptations of existing computational OT tools can be useful in solving it. We leave a thorough study of this approach as future work.

Acknowledgements

The authors are thankful to Young-Heon Kim and Brendan Pass for enlightening discussions on topics related to the content of this paper.

Funding

National Science Foundation-Division of Mathematical Sciences (2236447 to N.G.T.); IFDS at UW-Madison and National Science Foundation through TRIPODS (2023239 to N.G.T.); National Science Foundation Division of Mathematical Sciences (2311062 to B.S.).

Data Availability Statement

No new data were generated or analysed in support of this research.

References

1. Agueh, M. & Carlier, G. (2011) Barycenters in the Wasserstein space. *SIAM J. Math. Anal.*, **43**, 904–924.
2. Ambrosio, L., Gigli, N. & Savaré, G. (2008) *Gradient flows: in metric spaces and in the space of probability measures*. Birkäuser Basel: Springer Science & Business Media.
3. Andrews, D. F. & Mallows, C. L. (1974) Scale mixtures of normal distributions. *J. R. Stat. Soc. Ser. B*, **36**, 99–102.
4. Bartholomew, D., Knott, M. & Moustaki, I. (2011) *Latent variable models and factor analysis. Wiley Series in Probability and Statistics*, third edn. Chichester: John Wiley & Sons, Ltd. A unified approach.
5. Blau, Y. and Michaeli, T. (2018) The perception-distortion trade-off. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6228–6237.
6. Böhning, D. (2000) Computer-assisted analysis of mixtures and applications. *Technometrics*, **44**, 442. https://doi.org/10.1080/00401706.2000.10485740.
7. Buttazzo, G., De Pascale, L. & Gori-Giorgi, P. (2012) Optimal-transport formulation of electronic density-functional theory. *Phys. Rev. A*, **85**.
8. Carlier, G. & Ekeland, I. (2008) Matching for teams. *Econ. Theory*, **42**, 397–418.
9. Carroll, R. J. & Hall, P. (1988) Optimal rates of convergence for deconvolving a density. *J. Am. Stat. Assoc.*, **83**, 1184–1186.
10. Chernozhukov, V., Galichon, A., Hallin, M. & Henry, M. (2017) Monge-Kantorovich depth, quantiles, ranks and signs. *Ann. Stat.*, **45**, 223–256.
11. Cotar, C., Friesecke, G. & Klüppelberg, C. (2012) Density functional theory and optimal transportation with coulomb cost. *Commun. Pure Appl. Math.*, **66**, 548–599.
12. Deb, N., Ghosal, P. & Sen, B. (2021) Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections. *Advances in Neural Information Processing Systems*, **34**, 29736–29753.

13. Deb, N. & Sen, B. (2023) Multivariate rank-based distribution-free nonparametric testing using measure transportation. *J. Am. Stat. Assoc.*, **118**, 192–207.
14. Del Barrio & Loubes, J.-M. (2019) Central limit theorems for empirical transportation cost in general dimension. *Ann. Probab.*, **47**, 926–951.
15. Delbracio, M. & Milanfar, P. (2023) Inversion by direct iteration: an alternative to denoising diffusion for image restoration. *Trans. Mach. Learn. Res.* Featured Certification.
16. Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B*, **39**, 1–38 *With discussion.*, 39, 1, 22.
17. Dittmer, S., Schönlieb, C.-B., Maass, P. & Cambridge Image Analysis, Centre for Mathematical Sciences, University of Cambridge, Wilberforce Road, Cambridge CB3 0WA, United Kingdom, Zentrum fr Technomathematik. FB 3 Mathematik und Informatik, Universität Bremen Postfach 330 440, 28344 Bremen, Germany (2024) Ground truth free denoising by optimal transport. *Numerical Algebra, Control and Optimization*, **14**, 34–58.
18. Divol, V., Niles-Weed, J. & Pooladian, A.-A. (2022) Optimal transport map estimation in general function spaces. arXiv preprint arXiv:2212.03722.
19. Dyson, F. (1926) A method for correcting series of parallax observations. *Mon. Not. R. Astron. Soc.*, **86**, 686–706.
20. Efron, B. (2003) Robbins, empirical Bayes and microarrays. *Ann. Statist.*, **31**, 366–378.
21. Efron, B. (2009) Empirical Bayes estimates for large-scale prediction problems. *J. Am. Stat. Assoc.*, **104**, 1015–1028.
22. Efron, B. (2010) *Large-scale inference, volume 1 of Institute of Mathematical Statistics (IMS) Monographs*. Cambridge: Cambridge University Press Empirical Bayes methods for estimation, testing, and prediction.
23. Efron, B. (2011) Tweedie's formula and selection bias. *J. Am. Stat. Assoc.*, **106**, 1602–1614.
24. Efron, B. (2014) Two modeling strategies for empirical Bayes estimation. *Stat. Sci.*, **29**, 285–301.
25. Efron, B. (2016) Empirical Bayes deconvolution estimates. *Biometrika*, **103**, 1–20.
26. Efron, B. (2019) Bayes, oracle Bayes and empirical Bayes. *Stat. Sci.*, **34**, 177–201.
27. Efron, B. (2022) *Exponential families in theory and practice*. Cambridge: Cambridge University Press.
28. Ekeland, I. (2004) An optimal matching problem. *ESAIM. Control. Optim. Calc. Var.*, **11**, 57–71.
29. Fan, J. (1991) On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Stat.*, **19**, 1257–1272.
30. Fan, Z., Guan, L., Shen, Y. & Wu, Y. (2023) Gradient flows for empirical bayes in high-dimensional linear models. arXiv preprint arXiv:2312.12708.
31. Federer, H. (1959) Curvature measures. *Trans. Am. Math. Soc.*, **93**, 418–491.
32. Feller, W. (1971) *An introduction to probability theory and its applications*, vol. **II**, second edn. New York-London-Sydney: John Wiley & Sons, Inc.
33. Fournier, N. & Guillin, A. (2015) On the rate of convergence in Wasserstein distance of the empirical measure. *Probab. Theory Related Fields*, **162**, 707–738.
34. Freirich, D., Michaeli, T. & Meir, R. (2021) A theory of the distortion-perception tradeoff in wasserstein space. *Advances in Neural Information Processing Systems*, vol. **34**. (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang & J. W. Vaughan eds). Curran Associates, Inc., pp. 25661–25672.
35. García Trillos, N. & Slepčev, D. (2016) Continuum limit of total variation on point clouds. *Arch. Rational Mech. Anal.*, **220**, 193–241.
36. Ghodrati, L. & Panaretos, V. M. (2022) Distribution-on-distribution regression via optimal transport maps. *Biometrika*, **109**, 957–974.
37. Ghodrati, L. & Panaretos, V. M. (2023) Transportation of measure regression in higher dimensions. arXiv:2305.17503.
38. Ghosal, P. & Sen, B. (2022) Multivariate ranks and quantiles using optimal transport: consistency, rates and nonparametric testing. *Anna. Stat.*, **50**, 1012–1037.
39. Gu, J. & Koenker, R. (2017) Empirical Bayesball remixed: empirical Bayes methods for longitudinal data. *J. Appl. Econom.*, **32**, 575–599.

40. Gu, J. & Koenker, R. (2023) Invidious comparisons: ranking and selection as compound decisions. *Econometrica*, **91**, 1–41.
41. Hallin, M., del Barrio, Cuesta-Albertos, J. & Matrán, C. (2021) Distribution and quantile functions, ranks and signs in dimension *d*: a measure transportation approach. *Ann. Stat.*, **49**, 1139–1165.
42. Hallin, M., Hlubinka, D. & Hudecová, V. (2023) Efficient fully distribution-free center-outward rank tests for multiple-output regression and MANOVA. *J. Am. Stat. Assoc.*, **118**, 1923–1939.
43. Hütter, J.-C. & Rigollet, P. (2021) Minimax estimation of smooth optimal transport maps. *Ann. Stat.*, **49**, 1166–1194.
44. Ignatiadis, N. & Huber, W. (2021) Covariate powered cross-weighted multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **83**, 720–751.
45. Jiang, W. & Zhang, C.-H. (2009) General maximum likelihood empirical Bayes estimation of normal means. *Ann. Stat.*, **37**, 1647–1684.
46. Keener, R. W. (2010) *Theoretical statistics. Springer texts in statistics*. New York: Springer Topics for a core course.
47. Kiefer, J. & Wolfowitz, J. (1956) Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Stat.*, **27**, 887–906.
48. Koenker, R. & Mizera, I. (2014) Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *J. Am. Stat. Assoc.*, **109**, 674–685.
49. Koenker, R. & Gu, J. (2019) Comment: minimalist *g*-modeling [MR3983318]. *Stat. Sci.*, **34**, 209–213.
50. Laird, N. (1978) Nonparametric maximum likelihood estimation of a mixed distribution. *J. Am. Stat. Assoc.*, **73**, 805–811.
51. Langaas, M., Lindqvist, B. H. & Ferkingstad, E. (2005) Estimating the proportion of true null hypotheses, with application to DNA microarray data. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **67**, 555–572.
52. Lashkari, D. and Golland, P. (2008) Convex clustering with exemplar-based models. In Platt, J., Koller, D., Singer, Y. and S. Roweis, eds, *Advances in neural information processing systems*, vol. 20, pp. 825–832. Curran Associates, Inc.
53. Lindsay, B. G. (1983) The geometry of mixture likelihoods: a general theory. *Ann. Stat.*, **11**, 86–94.
54. Lindsay, B. G. (1995) *Mixture models: theory, geometry and applications*. In *NSF-CBMS regional conference series in probability and statistics*, vol. 5, pp. i–163. JSTOR. http://www.jstor.org/stable/4153184.
55. Louis, T. A. (1984) Estimating a population of parameter values using bayes and empirical bayes methods. *J. Am. Stat. Assoc.*, **79**, 393–398.
56. Manole, T. & Niles-Weed, J. (2024) Sharp convergence rates for empirical optimal transport with smooth costs. *Anna. Appl. Probab.*, **34**, 1108–1135.
57. McLachlan, G. & Peel, D. (2000) *Finite mixture models. Wiley Series in Probability and Statistics: Applied Probability and Statistics*. New York: Wiley-Interscience.
58. Meister, A. (2009) *Deconvolution problems in nonparametric statistics*, vol. **193**. Springer Berlin, Heidelberg: Springer Science & Business Media.
59. Park, T. & Casella, G. (2008) The Bayesian lasso. *J. Am. Stat. Assoc.*, **103**, 681–686.
60. Pass, B. (2015) Multi-marginal optimal transport: theory and applications. *ESAIM. Math. Model. Numer. Anal.*, **49**, 1771–1790.
61. Peyré, G., Cuturi, M., et al. (2019) Computational optimal transport: with applications to data science. *Found. Trends Mach. Learn.*, **11**, 355–607.
62. Rabin, J., Peyré, G., Delon, J., and Bernot, M. (2012) Wasserstein barycenter and its application to texture mixing. In Bruckstein, A. M., ter Haar Romeny, B. M., Bronstein, A. M., and Bronstein, M. M., eds, *Scale Space and Variational Methods in Computer Vision*, pp. 435–446, Berlin, Heidelberg, Springer Berlin Heidelberg.
63. Robbins, H. (1956) An empirical Bayes approach to statistics. In Neyman, J., ed, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I*, pp. 157–163. University of California Press, Berkeley-Los Angeles, Calif.

64. Santambrogio, F. (2015) *Optimal transport for applied mathematicians, volume 87 of Progress in Nonlinear Differential Equations and their Applications*. Cham: Birkhäuser/Springer Calculus of variations, PDEs, and modeling.
65. Shi, H., Hallin, M., Drton, M. & Han, F. (2022) On universally consistent and fully distribution-free rank tests of vector independence. *Ann. Stat.*, **50**, 1933–1959.
66. Slawski, M. & Sen, B. (2024) Permuted and unlinked monotone regression in R^d: an approach based on mixture modeling and optimal transport. *J. Mach. Learn. Res.*, **25**, 1–57.
67. Soloff, J., Guntuboyina, A. & Sen, B. (2024) Multivariate, heteroscedastic empirical bayes via nonparametric maximum likelihood. *J. R. Stat. Soc. Ser. B Methodol.*, qkae040. https://doi.org/10.1093/jrsssb/qkae040.
68. Stephens, M. (2017) False discovery rates: a new deal. *Biostatistics*, **18**, 275–294.
69. Villani, C. (2003) *Topics in optimal transportation, volume 58 of Graduate Studies in Mathematics*. Providence, RI: American Mathematical Society.
70. Villani, C. (2009) *Optimal transport: old and new, volume 338 of Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Berlin: Springer-Verlag.
71. Wainwright, M. J. & Jordan, M. I. (2008) Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, **1**, 1–305.
72. Wang, W., Wen, F., Yan, Z. & Liu, P. (2023) Optimal transport for unsupervised denoising learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, **45**, 2104–2118.
73. Weed, J. & Bach, F. (2019) Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, **25**, 2620–2648.
74. West, M. (1984) Outlier models and prior distributions in Bayesian linear regression. *J. R. Stat. Soc. Ser. B*, **46**, 431–439.
75. Woodroofe, M. & Sun, J. (1993) A penalized maximum likelihood estimate of $f(0+)$ when f is nonincreasing. *Stat. Sin.*, **3**, 501–515.
76. Yaqing Chen, Z. L. & Müller, H.-G. (2023) Wasserstein regression. *J. Am. Stat. Assoc.*, **118**, 869–882.
77. Zhang, C.-H. (1990) Fourier methods for estimating mixing densities and distributions. *Ann. Stat.*, **18**, 806–831.
78. Zhang, Y., Cui, Y., Sen, B. & Toh, K.-C. (2024) On efficient and scalable computation of the nonparametric maximum likelihood estimator in mixture models. *J. Mach. Learn. Res.*, **25**, 1–46.
79. Zhong, X., Su, C. & Fan, Z. (2022) Empirical Bayes PCA in high dimensions. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **84**, 853–878.

A. Exponential families

Consider a random vector $Z \in \mathbb{R}^d$ having a density w.r.t. a dominating measure λ, parametrized by $\theta := (\theta_1, \ldots, \theta_m) \in \mathbb{R}^m$ and expressible as

$$p(z \mid \theta) := \exp\left[\sum_{j=1}^{m} \theta_j T_j(z) - A(\theta)\right] h(z), \qquad \text{for } z \in \mathbb{R}^p. \tag{A.1}$$

Here $h : \mathbb{R}^d \to \mathbb{R}$ is a non-negative function, $T = (T_1, \ldots, T_m)$ is a measurable function from $\mathbb{R}^d$ to $\mathbb{R}^m$ and the parameter space is the set

$$\Omega := \{\theta \in \mathbb{R}^m : A(\theta) < \infty\}, \tag{A.2}$$

where the function $A : \Omega \to \mathbb{R}$ (sometimes referred to as the *cumulant function* or the *log-partition function*) is defined as

$$A(\theta) := \log \int \exp\left[\sum_{i=1}^{m} \theta_i T_i(z)\right] h(z)\, d\lambda(z). \tag{A.3}$$

Through the discussion in this appendix we will assume that Ω is a non-empty open subset of $\mathbb{R}^m$ for simplicity.

In this case, Z is said to belong to a *regular m-parameter exponential family*, and θ is the natural or canonical parametrization. There are many examples of parametric families belonging to an exponential family, e.g. Gaussian, binomial, multinomial, Poisson, gamma and beta distributions, as well as many others. Here are some examples.

EXAMPLE A1 (Exponential distribution). Consider the exponential distribution parametrized by $\beta \in (0, \infty)$:

$$p_\beta(z) = \beta e^{-\beta z} \mathbf{1}_{(0,\infty)}(z). \tag{A.4}$$

The above family is indeed a one-parameter exponential family with natural parameter $\theta := -\beta$ and $\Omega = (-\infty, 0)$. Here $T(x) = x$, $h(x) = \mathbf{1}_{(0,\infty)}(x)$ and $A(\theta) = \log \int_0^\infty e^{\theta x} dx = \log(-\theta^{-1})$.

EXAMPLE A2 (Multivariate normal). Consider the family of multivariate normal distributions on $\mathbb{R}^d$ with a fixed known non-singular covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ and unknown mean vector $\beta = (\beta_1, \ldots, \beta_d) \in \mathbb{R}^d$, i.e. $Z \sim N_d(\beta, \Sigma)$ has density given by

$$p_\beta(z) = \frac{e^{-\frac{1}{2}(z-\beta)^\top \Sigma^{-1}(z-\beta)}}{\sqrt{(2\pi)^d |\Sigma|}}, \qquad \text{for } z \in \mathbb{R}^d. \tag{A.5}$$

It is easy to check that (A.5) can be expressed in the form (A.1) where we take

$$\theta := \Sigma^{-1}\beta, \qquad T(z) := z, \qquad A(\theta) := \frac{1}{2}\theta^\top \Sigma \theta \quad \text{and} \quad h(z) \equiv p_0(z) = \frac{e^{-\frac{1}{2}z^\top \Sigma^{-1} z}}{\sqrt{(2\pi)^d |\Sigma|}}.$$

Suppose that $Z \sim p(\cdot \mid \theta)$ as in (A.1). Here are some important properties of exponential families.

1. The support of Z (i.e. z such that $p(z \mid \theta) > 0$) does not depend on θ.
2. It is clear that the statistic $T(Z)$ is a *sufficient statistic* for this family. It can be shown that[4]
$$\mathbb{E}_\theta[T_j(Z)] = \frac{\partial A(\theta)}{\partial \theta_j}, \qquad \text{for } j = 1, \ldots, m. \tag{A.6}$$
3. The natural parameter space Ω is a *convex set* and the cumulant function $A(\cdot)$ is a *convex function*.
4. The moment generating function of $T \equiv (T_1(Z), \ldots, T_m(Z))$ is, for $u \in \mathbb{R}^m$ such that $u + \theta \in \Omega$,
$$\begin{aligned} M_T(u) := \mathbb{E}[e^{u^\top T}] &= \int e^{u^\top T} e^{\theta^\top T - A(\theta)} h(z)\, d\lambda(z) \\ &= e^{A(u+\theta)-A(\theta)} \int p(z \mid u+\theta) d\lambda(z) = e^{A(u+\theta)-A(\theta)}. \end{aligned}$$

[4] A proof of this can be obtained as follows. Recall (C.3). Thus, $e^{A(\theta)} = \int e^{\theta^\top T(z)} h(z)\, d\lambda(z)$. Differentiating this expression w.r.t. θ_j, which can be done under the integral if $\theta \in \Omega^o$ (here Ω^o is the interior of Ω), gives $e^{A(\theta)} \frac{\partial A(\theta)}{\partial \theta_j} = \int T_j(z) e^{\theta^\top T(z)} h(z)\, d\lambda(z) \quad \Rightarrow \frac{\partial A(\theta)}{\partial \theta_j} = \int T_j(z) p(z \mid \theta)\, d\lambda(z) = \mathbb{E}_\theta[T_j(Z)]$.

5. The *cumulant generating function* is

$$K_T(u) := \log M_T(u) = A(u+\theta) - A(\theta). \tag{A.7}$$

6. Noting that if $M_T(\cdot)$ is finite in some neighbourhood of the origin, then M_T has continuous derivatives of all orders at the origin, and for $r_j \geq 0$, for $j = 1, \ldots, m$,

$$\mathbb{E}[T_1^{r_1}(Z) \times \cdots \times T_s^{r_s}(Z)] = \frac{\partial^{r_1}}{\partial u_1^{r_1}} \cdots \frac{\partial^{r_s}}{\partial u_s^{r_s}} M_T(u)\Big|_{u=0}.$$

Thus, when $r_j = 1$ and $r_k = 0$ for all $k \neq j$, we obtain (A.6).

See Keener, [46, Chapter 10] and [27] for a more detailed study of exponential families.

B. Tweedie's formula

Now suppose that Θ is assumed to have a prior distribution G^* (on $\Omega \subset \mathbb{R}^m$). Thus our model becomes

$$\Theta \sim G^* \qquad \text{and} \qquad Z \mid \Theta = \theta \sim p(\cdot \mid \theta), \tag{B.1}$$

where we assume that $p(\cdot \mid \theta)$ comes from the exponential family (A.1). Then the marginal density of Z (w.r.t. λ) is

$$f_{G^*}(z) := \int p(z \mid \theta)\, dG^*(\theta), \qquad \text{for } z \in \mathbb{R}^d.$$

Let $\mathcal{Z} \subset \mathbb{R}^d$ be the support of the marginal distribution of Z. Now Bayes rule provides the posterior density of Θ given Z. Suppose that Θ has density $g(\cdot)$, w.r.t. a dominating measure ξ, with support contained in the set $\Omega \subset \mathbb{R}^m$. Then, the posterior density of Θ given $Z = z$ (w.r.t. ξ) is given by, for $\theta \in \Omega$ and $z \in \mathcal{Z}$,

$$p_{\Theta|Z}(\theta \mid z) = \frac{p(z \mid \theta) g(\theta)}{f_{G^*}(z)} = \frac{e^{\theta^\top T(z) - A(\theta)} h(z) g(\theta)}{f_{G^*}(z)} = e^{\theta^\top T(z) - \kappa(z)} e^{-A(\theta)} g(\theta), \tag{B.2}$$

where

$$\kappa(z) := \log\left(\frac{f_{G^*}(z)}{h(z)}\right), \qquad \text{for} \quad z \in \mathcal{Z}. \tag{B.3}$$

This implies that $\Theta \mid Z = z$ is also an *exponential family* with canonical parameter $T(z)$, sufficient statistic Θ and log-partition function $\kappa(z)$. Thus, the cumulant generating function is (cf. (A.7))

$$\log \mathbb{E}[e^{\Theta^\top t} \mid Z = z] = \kappa(t+z) - \kappa(z) \tag{B.4}$$

for $z \in \mathcal{Z}$ such that $t + z \in \mathcal{Z}$.

Tweedie's formula, given below, calculates the posterior expectation of Θ given $Z = z$ in the setting (B.1).

LEMMA B1 (Tweedie's formula). For $z \in \mathcal{Z}$, we have

$$\mathbb{E}[\Theta \mid Z = z] = \nabla \kappa(z) = \frac{\nabla f_{G^*}(z)}{f_{G^*}(z)} - \frac{\nabla h(z)}{h(z)}. \tag{B.5}$$

Proof. The result is a direct consequence of the fact that the distribution of $\Theta \mid Z = z$ is an m-parameter exponential family with log-partition function $\kappa(\cdot)$ defined via (B.3): By property 2. above (see (A.6))

the expectation of the sufficient statistic Θ can then be expressed as the gradient of the log-partition function. □

For $d = m = 1$, the above formula for the Gaussian case was given in Robbins, [63]. Efron, [23] calls this Tweedie's formula since Robbins attributes it to M.C.K. Tweedie; however it appears earlier in Dyson, [19] who credits it to the English astronomer Arthur Eddington.

LEMMA B2. Consider model (B.1) where we assume that $p(\cdot \mid \theta)$, for $\theta \in \Omega$, is a member of an exponential family of distributions as in (A.1) with $T(z) = z$ and $m = d$. Suppose further that $h(\cdot)$ in (A.1) integrates to 1 (w.r.t. λ). Then $\kappa(\cdot)$, as defined in (B.3), is a convex function. As a consequence, $\mathbb{E}[\Theta \mid Z = z]$ is the gradient of a convex function.

Proof. Observe that under the assumptions of the lemma, from (B.3) we see that the distribution of $\Theta \mid Z = z$ is an m-parameter exponential family with log-partition function $\kappa(\cdot)$ defined by (B.3). As the log-partition function $\kappa(\cdot)$ is known to be convex, the result follows. □

REMARK B1 (Tweedie's formula for multivariate normal distribution). Suppose now that Z has multivariate normal distribution with known covariance matrix as in Example A2. Then, for $z \in \mathbb{R}^d$,

$$\mathbb{E}[\Theta \mid Z = z] = \nabla\kappa(z) = \Sigma^{-1}z + \frac{\nabla f_{G^*}(z)}{f_{G^*}(z)},$$

where the last equality follows from (B.5) and the fact that $\nabla h(z) = -h(z)(\Sigma^{-1}z)$. Thus, the Bayes estimator of mean μ in (A.5) is

$$\mathbb{E}[\mu \mid Z = z] = z + \Sigma\frac{\nabla f_{G^*}(z)}{f_{G^*}(z)}. \tag{B.6}$$

C. Auxiliary results from measure theory and functional analysis

LEMMA C1. Let μ be a Borel probability measure over $\mathbb{R}^d$. Suppose that $\{\mathrm{T}_n\}_{n\in\mathbb{N}}$ is a sequence of (vector valued) Borel measurable maps $\mathrm{T}_n : \mathbb{R}^d \to \mathbb{R}^m$ and suppose that T is another Borel measurable map from $\mathbb{R}^d$ into $\mathbb{R}^m$.

The sequence of measures $\pi_n =: (\mathrm{Id} \times \mathrm{T}_n)_\sharp\mu$ converges weakly to $\pi := (\mathrm{Id} \times \mathrm{T})_\sharp\mu$ if and only if T_n converges in μ-measure to T, i.e. for every $\eta > 0$ we have

$$\lim_{n\to\infty} \mu\left(\left\{z \in \mathbb{R}^d \,:\, |\mathrm{T}_n(z) - \mathrm{T}(z)| \geq \eta\right\}\right) = 0.$$

Proof. Recall that weak convergence of probability measures is equivalent to convergence in Levy–Prokhorov metric, which we recall is defined as

$$d_{LP}(\pi_n, \pi) := \inf\left\{\varepsilon > 0 \,:\, \pi_n(A) \leq \pi(A^\varepsilon) + \varepsilon \text{ and} \pi(A) \leq \pi_n(A^\varepsilon) + \varepsilon, \quad \forall A \in \mathfrak{B}(\mathbb{R}^d \times \mathbb{R}^m)\right\}.$$

In the above, for an arbitrary A the set A^ε is defined as the set of points (z, θ) such that there exists $(\widetilde{z}, \widetilde{\theta}) \in A$ with $|z - \widetilde{z}| + |\theta - \widetilde{\theta}| < \varepsilon$.

Let us first assume that π_n converges weakly to π and let $\varepsilon_n := 2d_{LP}(\pi_n, \pi)$. Fix $\eta > 0$ and $r > 0$. From the fact that μ is a Borel probability measure over $\mathbb{R}^d$, it follows that T can be approximated in the μ-a.e. convergence sense by a sequence of Lipschitz continuous functions (with possibly growing Lipschitz constants). Indeed, by density (in the μ-a.e. sense) of simple functions in the set of all measurable functions and the fact that we are considering the Borel σ-algebra (which is generated by

open sets) one can reduce the problem to approximating (scalar) indicator functions of open sets. In turn, using rescaled distance functions (which are Lipschitz), one can easily approximate indicator functions of open sets with Lipschitz continuous functions as desired. It thus follows that there exists a Lipschitz function $\psi_r : \mathbb{R}^d \to \mathbb{R}^m$ such that

$$\mu(G_r) \leq r$$

for the set G_r defined as

$$G_r := \left\{ z \in \mathbb{R}^d \, : \, |\psi_r(z) - \mathrm{T}(z)| \geq \frac{\eta}{3} \right\}.$$

The above says that we can approximate the Borel measurable function T up to accuracy $\eta/3$ by the Lipschitz function ψ_r on a set with 'large' μ-probability. Intersecting the set $\left\{ z \in \mathbb{R}^d \, : \, |\mathrm{T}_n(z) - \mathrm{T}(z)| \geq \eta \right\}$ with G_r and, separately, with G_r^c, we get the inequality

$$\mu\left(\left\{ z \in \mathbb{R}^d \, : \, |\mathrm{T}_n(z) - \mathrm{T}(z)| \geq \eta \right\} \right) \leq r + \mu\left(\left\{ z \in \mathbb{R}^d \, : \, |\mathrm{T}_n(z) - \psi_r(z)| \geq \frac{2}{3}\eta \right\} \right).$$

Let us now consider the set

$$A_r := \left\{ (z, \theta) \in \mathbb{R}^d \times \mathbb{R}^m \, : \, |\psi_r(z) - \theta| \geq \frac{2}{3}\eta \right\}.$$

Due to the specific form of the measure π_n, we can write

$$\mu\left(\left\{ z \in \mathbb{R}^d \, : \, |\psi_r(z) - \mathrm{T}_n(z)| \geq \frac{2}{3}\eta \right\} \right) = \pi_n(A_r) \leq \pi(A_r^{\varepsilon_n}) + \varepsilon_n.$$

On the other hand,

$$\begin{aligned} \pi(A_r^{\varepsilon_n}) &= \mu\left(\left\{ z \in \mathbb{R}^d \, : \, \exists\, (\widetilde{z}, \widetilde{\theta}) \text{ s.t.} |z - \widetilde{z}| + |\widetilde{\theta} - \mathrm{T}(z)| < \varepsilon_n \text{ and} |\widetilde{\theta} - \psi_r(\widetilde{z})| \geq \frac{2}{3}\eta \right\} \right) \\ &\leq \mu\left(\left\{ z \in \mathbb{R}^d \, : \, \exists \widetilde{z} \text{ s.t.} |z - \widetilde{z}| < \varepsilon_n \text{ and} |\mathrm{T}(z) - \psi_r(\widetilde{z})| \geq \frac{2}{3}\eta - \varepsilon_n \right\} \right). \end{aligned}$$

In turn, we see that

$$\mu\left(G_r \cap \left\{ z \in \mathbb{R}^d \, : \, \exists \widetilde{z} \text{ s.t.} |z - \widetilde{z}| < \varepsilon_n \text{ and} |\mathrm{T}(z) - \psi_r(\widetilde{z})| \geq \frac{2}{3}\eta - \varepsilon_n \right\} \right) \leq r,$$

and $\mu(G_r^c \cap \{ z \in \mathbb{R}^d \, : \, \exists \widetilde{z} \text{ s.t.} |z - \widetilde{z}| < \varepsilon_n \text{ and} |\mathrm{T}(z) - \psi_r(\widetilde{z})| \geq \frac{2}{3}\eta - \varepsilon_n \})$ is smaller than

$$\mu\left(\left\{ z \in \mathbb{R}^d \, : \, \exists \widetilde{z} \text{ s.t.} |z - \widetilde{z}| < \varepsilon_n \text{ and} |\psi_r(z) - \psi_r(\widetilde{z})| \geq \frac{1}{3}\eta - \varepsilon_n \right\} \right).$$

Since the function ψ_r is Lipschitz and $\varepsilon_n \to 0$ as $n \to \infty$, it follows that

$$\lim_{n \to \infty} \mu\left(\left\{ z \in \mathbb{R}^d \, : \, \exists \widetilde{z} \text{ s.t.} |z - \widetilde{z}| < \varepsilon_n \text{ and} |\psi_r(z) - \psi_r(\widetilde{z})| \geq \frac{1}{3}\eta - \varepsilon_n \right\} \right) = 0.$$

From all the above inequalities it follows that

$$\limsup_{n\to\infty} \mu\left(\left\{z \in \mathbb{R}^d \ : \ |\mathrm{T}_n(z) - \mathrm{T}(z)| \geq \eta\right\}\right) \leq 2r.$$

Since $r > 0$ was arbitrary, we conclude that

$$\lim_{n\to\infty} \mu\left(\left\{z \in \mathbb{R}^d \ : \ |\mathrm{T}_n(z) - \mathrm{T}(z)| \geq \eta\right\}\right) = 0,$$

as we wanted to prove.

Conversely, if T_n converges in μ-measure, then we can assume without loss of generality that the convergence is actually μ-a.e. (as we can work along subsequences). It follows now that for every $\phi \in C_b(\mathbb{R}^d \times \mathbb{R}^m)$,

$$\begin{aligned}
\lim_{n\to\infty} \int \phi(z,\theta)\, d\pi_n(z,\theta) &= \lim_{n\to\infty} \int \phi(z, \mathrm{T}_n(z))\, d\mu(z) \\
&= \int \lim_{n\to\infty} \phi(z, \mathrm{T}_n(z))\, d\mu(z) \\
&= \int \phi(z, \lim_{n\to\infty} \mathrm{T}_n(z))\, d\mu(z) \\
&= \int \phi(z, \mathrm{T}(z))\, d\mu(z) \\
&= \int \phi(z,\theta)\, d\pi(z,\theta),
\end{aligned}$$

where the second equality follows from the dominated convergence theorem, and the third equality follows from the continuity of ϕ. This shows that π_n converges weakly to π. □

Remark C1. Lemma C1 is analogous to the characterization of the TL^p convergence in Proposition 3.12. in [35]. In Lemma C1, however, we have restricted our attention to the case where the base measure for the entire approximating sequence is the same (i.e. μ).

We recall the definition of convergence in the weak topology of the Hilbert space $\mathbb{L}^2(\mathbb{R}^d : \mathbb{R}^m; \mu)$.

Definition C1. We say that the sequence $\{\mathrm{T}_n\}_{n\in\mathbb{N}}$ converges weakly in $\mathbb{L}^2(\mathbb{R}^d : \mathbb{R}^m; \mu)$ to T if

$$\lim_{n\to\infty} \int \mathrm{T}_n(z) \cdot g(z)\, d\mu(z) = \int \mathrm{T}(z) \cdot g(z)\, d\mu(z) \quad \forall g \in \mathbb{L}^2(\mathbb{R}^d : \mathbb{R}^m; \mu).$$

The next two lemmas are well-known results in measure theory and functional analysis.

Lemma C2. Suppose that $\mathrm{T}_n \to \mathrm{T}$ in μ-measure (as defined in Lemma C1) and that

$$\sup_{n\in\mathbb{N}} \int |\mathrm{T}_n(z)|^2\, d\mu(z) < \infty.$$

Then T_n converges weakly in $\mathbb{L}^2(\mathbb{R}^d : \mathbb{R}^m; \mu)$ to T, as $n \to \infty$.

Proof. From the second moment condition we deduce that the sequence $\{\mathrm{T}_n\}_{n\in\mathbb{N}}$ is uniformly integrable. This, together with the dominated convergence theorem, allows us to conclude that

$$\lim_{n\to\infty}\int \mathrm{T}_n(z)\cdot g(z)\,d\mu(z) = \int \mathrm{T}(z)\cdot g(z)\,d\mu(z)$$

for every $g\in\mathbb{L}^2(\mathbb{R}^d:\mathbb{R}^m;\mu)$, which is what we wanted to show. □

Lemma C3. If T_n converges weakly in $\mathbb{L}^2(\mathbb{R}^d:\mathbb{R}^m;\mu)$ to T, and in addition we have

$$\lim_{n\to\infty}\int |\mathrm{T}_n(z)|^2 d\mu(z) = \int |\mathrm{T}(z)|^2\,d\mu(z),$$

then

$$\lim_{n\to\infty}\|\mathrm{T}_n-\mathrm{T}\|_{\mathbb{L}^2(\mathbb{R}^d:\mathbb{R}^m;\mu)}=0.$$

Proof. Expanding the square, we get

$$\int |\mathrm{T}_n(z)-\mathrm{T}(z)|^2\,d\mu(z) = \int |\mathrm{T}_n(z)|^2\,d\mu(z)+\int |\mathrm{T}(z)|^2\,d\mu(z)-2\int \mathrm{T}_n(z)\cdot\mathrm{T}(z)\,d\mu(z).$$

The result follows now from the above display, the assumed consistency of second moments and the fact that T_n converges weakly to T (see Definition C1). □

D. More general loss functions $\ell(\cdot,\cdot)$

If the squared error loss in problem (1.5) is substituted with an arbitrary loss function $\ell(\cdot,\cdot)$, the resulting problem

$$\inf_{\delta:\mathbb{R}^d\to\mathbb{R}^m}\mathbb{E}_{(Z,\Theta)\sim P_{Z,\Theta}}\left[\ell(\delta(Z),\Theta)\right] \qquad \text{subject to} \quad \delta(Z)\sim G^* \tag{D.1}$$

can still be written as a standard OT problem in Monge form:

$$\min_{\delta\,:\,\delta_\sharp\mu=G^*}\int_{\mathbb{R}^d} c_\ell(\delta(z),z)\,d\mu(z) \tag{D.2}$$

for the cost function

$$c_\ell(\theta,z) := \mathbb{E}[\ell(\theta,\Theta)\mid Z=z], \qquad \text{for } \theta\in\mathbb{R}^m,\ z\in\mathbb{R}^d.$$

The existence of solutions for (1.5) then reduces to proving existence of an OT map for (D.2).

Investigating the existence of OT maps for specific transport problems is an important topic in the theory of OT. A general strategy that can be followed for proving existence of optimal maps (also-called Monge maps) is based on the analysis of the optimality conditions of solutions to the Kantorovich relaxation [69,chapters 1–3] of the original Monge problem; an important property of Kantorovich relaxations is that they can be shown to have solutions under very mild lower semicontinuity assumptions on the transportation cost function (see e.g. Villani, [70,chapter 5]). Notice that the Kantorovich relaxation of (D.2) takes the form

$$\min_{\pi\in\Gamma(\mu,G^*)}\int c_\ell(\theta,z)\,d\pi(z,\theta).$$

Under appropriate assumptions on the cost function and marginals of a general OT problem, one can show that a solution to the Kantorovich relaxation must be supported on a graph of a function, and from this one can infer the existence of a solution to the original Monge problem. In principle, one could attempt to carry out this program for the OT problem (D.2), but we notice that the dependence of the cost $c_\ell(\cdot,\cdot)$ on the loss function ℓ and on the model $P_{Z,\Theta}$ may, in general, be rather intractable. For this reason, in this paper we have focused on one tractable and important case, namely, the setting of the squared error loss, for which we can prove a variety of theoretical results and discuss a variety of algorithmic consequences.

E. An empirical Bayes approach to estimating the OT-based denoiser δ^*

Suppose that we observe $Z_1,\dots,Z_n$ from model (1.1) where the unobserved latent variables are $\Theta_1,\dots,\Theta_n$ drawn i.i.d. from G^*. We assume here that G^* is unknown and belongs to a (sub)-family $\mathscr{P}$ of $\mathscr{P}(\Omega)$, the space of all probability measures on $\Omega \subset \mathbb{R}^m$. In the following text we discuss an approach to estimate the OT-based denoiser δ^* based on the observed data $Z_1,\dots,Z_n$. We plan to pursue a more thorough analysis of this framework in future work.

Our approach can be broken down into three steps: (a) first we estimate the unknown prior G^*, say by $\widehat{G}$, using the method of maximum likelihood, and (b) then use $\widehat{G}$ as a plug-in estimator to solve an empirical version of the Kantorovich relaxation problem in (2.7). This yields an optimal coupling (based on the data) which can (c) then be used to define an estimator of the OT-based denoiser δ^*.

Let us describe each step in a bit more detail now.

(a) We apply the method of maximum likelihood (ML) to estimate G^*. Marginally, the observations Z_i's are i.i.d. with density f_{G^*} (as defined in (1.2)). An ML estimator is any $\widehat{G} \in \mathscr{P}$ which maximizes the marginal likelihood of the observations $(Z_i)_{i=1}^n$, i.e.

$$\widehat{G} \in \operatorname*{argmax}_{G\in\mathscr{P}} \frac{1}{n}\sum_{i=1}^{n} \log f_G(Z_i). \tag{E.1}$$

Note that when $\mathscr{P} = \mathscr{P}(\Omega)$, the space of all probability measures on $\Omega \subset \mathbb{R}^m$, this estimator is called the non-parametric MLE (NPMLE) of G^* and has been studied in detail in the statistics literature; see [45, 47, 53, 54, 67] and the references therein. In particular, in this case (E.1) is an infinite-dimensional convex optimization problem for which several algorithms have been proposed; see e.g. [6, 50, 52, 67, 78]. Moreover, this approach can be applied even when $\mathscr{P} \subsetneq \mathscr{P}(\Omega)$ and/or $\mathscr{P}$ is finite-dimensional. In the empirical Bayes literature this approach falls under the general framework of G-modelling as we directly estimate the unknown prior G^* ([26]).

(b) In our second step, given an estimate $\widehat{G}$ of the prior G^*, empirical Bayes imitates the optimal Bayesian analysis [26]. If G^* were known, the Bayes estimator of Θ_i (under the squared error loss) would be the posterior mean $\mathbb{E}_{G^*}[\Theta_i \mid Z_i]$ as defined in (1.4) (here by $\mathbb{E}_{G^*}[\dots]$ we emphasize the dependence on G^*). The NPMLE (E.1) yields a fully data-driven, empirical Bayes estimate of this posterior mean via

$$\widehat{\overline{\theta}}(Z_i) := \mathbb{E}_{\widehat{G}}\big[\widehat{\Theta}_i \mid Z_i\big], \quad \text{where} \quad \widehat{\Theta}_i \sim \widehat{G} \quad \text{and} \quad Z_i \mid \widehat{\Theta}_i = \theta \sim p(\cdot \mid \theta). \tag{E.2}$$

Once we obtain an estimator ($\widehat{\overline{\theta}}$ as above) of $\overline{\theta}(\cdot)$, we can solve an empirical version of (2.7) defined via

$$\widehat{\pi} \in \operatorname*{argmin}_{\pi\in\Gamma(\mu_n,\widehat{G})} \int |\widehat{\overline{\theta}}(z) - \vartheta|^2 \, d\pi(\mu_n,\vartheta), \tag{E.3}$$

where $\mu_n := \frac{1}{n}\sum_{i=1}^n \delta_{Z_i}$ is the empirical distribution of the Z_i's. If $\widehat{G}$ is the NPMLE over $\mathscr{P}(\Omega)$, the above computation is quite straightforward as it is known that $\widehat{G}$ is finitely supported (see [54]) and thus (E.3) reduces to a discrete-discrete OT problem which can be solved using the various computational OT tools available in the literature (see e.g. [61]).

(c) The optimal coupling $\widehat{\pi}$ obtained in (E.3) can now be used to construct an estimator of the OT-based denoiser $\delta^*(\cdot) \equiv \nabla\varphi^*(\widehat{\overline{\theta}}(\cdot))$ (see (2.9) and (2.10)) via the *barycentric projection* of $\widehat{\pi}$:

$$\delta_{\widehat{\pi}}(z) := \int_{\Omega} \vartheta \, d\widehat{\pi}(\vartheta|z), \qquad \text{for } z \in \mathscr{Z}. \tag{E.4}$$

We conjecture that $\delta_{\widehat{\pi}}$ will be a consistent estimator of δ^*; see [12] and [66] where the barycentric projection estimator has been investigated and shown to be consistent for estimating OT maps.