Commun. Optim. Theory 2025 (2025) 51 https://doi.org/10.23952/cot.2025.51

FULLY PROBABILISTIC DESIGN FOR OPTIMAL TRANSPORT

ANTHONY QUINN[1,2,*], SARAH BOUFELJA Y.[1], MARTIN CORLESS[3], ROBERT SHORTEN[1]

[1]Dyson School of Design Engineering, Imperial College London, UK
[2]Department of Electronic and Electrical Engineering, Trinity College Dublin, Ireland
[3]School of Astronautics and Aeronautics, Purdue University, West Lafayette, USA

Abstract. The relationship between entropy-regularized Kantorovich optimal transport (OT) and fully probabilistic design (FPD) of probability models is derived. In FPD, the (unattainable) zero-cost plan (i.e. probability measure) —called the ideal, π_I—is projected (in a minimum KLD sense) into the set of feasible plans constrained by fixed marginals, μ and ν. We show that π_I has a Gibbs structure. The regularizing measure, ϕ, acts as its base measure, and the cost metric, c, acts as its energy term. Important insights and design opportunities flow from this FPD-OT setting: (i) the fixed objects in regularized OT are classified either as constraints on the actual transport plan (μ, ν) or else as constraints on the ideal plan (ϕ, c and regularization constant, ε); and (ii) the modulation of ϕ by c and ε, in the ideal plan, π_I, allows a c-dependent ϕ to be designed, favouring plans that meet detailed cost-dependent constraints. Extensive examples are presented, illustrating both of these insights. In particular, we show how the FPD-OT setting of discrete regularized OT allows high-cost transport paths to be quenched.

Keywords. Entropic regularization; Fully probabilistic design; Ideal design; Kullback-Leibler divergence; Optimal transport.

2020 MSC. 68T37, 62F15, 94A17.

1. INTRODUCTION

Many machine learning (ML) problems reduce to the question of summarizing and comparing probability measures. For example, problems in domain adaptation, adversarial training and distributed learning fall withing this setting. Quantifying the relationship between two probability measures can be addressed via an appropriate divergence function [28]. However, such functions often do not consider the the semantics of the domain on which the set of measures is defined, be they spatial properties, physical distances, *etc*. Optimal transport (OT), on the other hand, converts this domain structure into a distance between probability measures, and, in doing so, endows the space of probability measures with a topology. If the domain is Euclidean, then the induced distance in the space of probability measures is Wasserstein, and so concepts

*Corresponding author.
E-mail address: a.quinn@imperial.ac.uk (A. Quinn).
Received September 17, 2024; Accepted December 1, 2024.

of interpolation, barycentres and gradient of functions are naturally extended to the space of measures [22]. For these reasons, OT has enjoyed widespread application in areas such as computer vision [24], computer graphics [28] and natural language processing [20], as well as in control [6], filtering [25] and sequential decision-making [31].

Notwithstanding these OT successes, when it comes to the practical requirement of modelling uncertainty in the marginals, or, equivalently, processing nonlinear moment constraints, there exists no systematic or generic methodology. This motivates our interest in fully probabilistic design (FPD). FPD is the axiomatic framework for designing probability measures subject to knowledge constraints, with choices ranked in respect of a (generally unattainable) ideal [17]. The knowledge constraints express any prior information about the unknown probability measure, in the form of a set membership, physical laws, etc. [23]. In this paper, we prove that entropy-regularized OT is a special case of FPD. Indeed, by formulating the OT problem as the constrained design of a joint probability measure, the connection with FPD emerges naturally. We refer to this FPD setting of regularized OT as FPD-OT.

The paper is structured as follows. In Section 2, we review the key concepts of OT, introducing the mathematical objects used later in the document. In Section 3, we review FPD and establish the connection between OT and FPD in Section 4. In Section 5, we illustrate the FPD-OT formalism with an example involving the processing of uncertainty in the marginals. An extensive simulation-based example in provided in Section 6, illustrating the detailed design of a plan subject to a cost-sensitive regularizing base measure. The main insights and additivities that follow from FPD-OT are discussed in Section 7, and the paper concludes with Section 8.

2. Optimal Transport (OT)

Let $(\Omega, \mathfrak{F}, \mathbb{P})$ be a probability space. $X\colon\ \Omega \mapsto \Omega_S$ and $Y\colon\ \Omega \mapsto \Omega_T$ denote two random variables, inducing $(\Omega_S, \mathfrak{F}_S, \mu)$ and $(\Omega_T, \mathfrak{F}_T, \nu)$, the source and target probability spaces, respectively, where Ω_S and Ω_T are two compact metric spaces. In the sequel, μ and ν denote probability measures, described by their Radon-Nikodym densities *w.r.t* the dominating measure, λ, which can be instantiated as either the Lebesgue measure or the counting measure, depending on the context. We overload μ and ν to denote the induced probability density functions (pdfs) in the continuous case, and probability mass functions (pmfs) in the discrete case.

Optimal Transport (OT) was originally introduced by the French mathematician, Gaspard Monge, to study the problem of shovelling—with minimal total cost—a pile of sand into a hole of the same volume [5]. This early formulation was too restrictive since there exists no feasible solution to the Monge problem for many choices of μ and ν. Kantorovitch later proposed a probabilistic relaxation, allowing transported mass to split [30]. In this new setting, the objective is to design a transport plan, i.e. a joint distribution, π, satisfying:

$$\pi^o_{OT}(x,y|\mathscr{K}) \stackrel{\text{def.}}{=} \underset{\pi\in\Pi_{\mathscr{K}}}{\operatorname{argmin}} \left\{ \int_{\Omega_S\times\Omega_T} c(x,y)\pi(x,y)d\lambda(x,y) \right\}. \tag{2.1}$$

$c : \Omega_S \times \Omega_T \mapsto \mathbb{R}^+$ is a measurable cost function, $\pi(x,y)$ denotes an unknown (variational) distribution with support in the product space, $\Omega_S \times \Omega_T$, and $\Pi_{\mathscr{K}}$ denotes the set of joint distributions, $\pi(x,y|\mathscr{K})$, with support in $\Omega_S \times \Omega_T$, on which we impose some knowledge constraints, $\mathscr{K}$. These knowledge constraints relate to any information which should be processed in the optimization problem (2.1) when designing the optimal solution, $\pi^o_{OT}(x,y|\mathscr{K})$. In the context

of classical OT, $\mathscr{K}$ represents the marginal constraints, that is,

$$\Pi_{\mathscr{K}} \overset{\text{def.}}{=} \Pi(\mu,\nu) \overset{\text{def.}}{=} \Big\{\pi \in \mathscr{P}(\Omega_S \times \Omega_T) \mid P_{\Omega_S\#}\pi \equiv \mu,\ P_{\Omega_T\#}\pi \equiv \nu\Big\}. \tag{2.2}$$

$\mathscr{P}(\Omega_S \times \Omega_T)$ denotes the set of distributions with support in $\Omega_S \times \Omega_T$. Furthermore, $P_{\Omega_S\#}$ and $P_{\Omega_T\#}$ are the push-forward operators associated with the (surjective) projections, $P_{\Omega_S}(x,y) = x$ and $P_{\Omega_T}(x,y) = y$, respectively. It follows that

$$\begin{aligned}\int_{\Omega_T} \pi(x,y|\mathscr{K})d\lambda(y) &\equiv \mu(x),\\ \int_{\Omega_S} \pi(x,y|\mathscr{K})d\lambda(x) &\equiv \nu(y).\end{aligned}$$

are the source and target marginal densities imposed as prior knowledge constraints, $\mathscr{K}$.

Remark 2.1. The objective in (2.1) does not require a parametric model of π. This distinguishes OT from parametric design methods for π, mainly via copula design [27].

In its discrete form, the Kantorovitch OT problem is a linear program (LP) and so it would be tempting to apply LP optimization to solve it. However, this program can be prohibitively expensive, especially in big data regimes. Indeed, if X and Y are discrete random variables, with $\#(\Omega_S) = n$ and $\#(\Omega_T) = m$, then the complexity of the LP scales at least in $\mathscr{O}(d^3\log(d))$, where $d = max(n,m)$ [12].

Entropy regularization is a widely deployed formulation for computationally efficient OT [22], in which an appropriate entropy functional of π is used to smooth the original problem. Towards defining entropy-regularized OT, recall the Kullback-Leibler divergence (KLD) [19]—also called the relative (or sometimes, confusingly, the cross [26]) entropy—from variational distribution, π, to a fixed one, ζ:

$$KL(\pi||\zeta) \equiv \begin{cases}\displaystyle\int_{\Omega_S\times\Omega_T} \pi(x,y)\log\Big(\frac{\pi(x,y)}{\zeta(x,y)}\Big)d\lambda(x,y) & \text{if } \pi \ll \zeta,\\ +\infty & \text{otherwise.}\end{cases} \tag{2.3}$$

The regularized OT problem reads:

$$\pi^o_{OT,\varepsilon,\phi}(x,y|\mathscr{K}) \equiv \underset{\pi\in\Pi_{\mathscr{K}}}{\operatorname{argmin}} \left\{\int_{\Omega_S\times\Omega_T} c(x,y)\pi(x,y)d\lambda(x,y) + \varepsilon KL(\pi||\phi)\right\}. \tag{2.4}$$

ϕ—which acts as the target or base distribution for regularization—has support in $\Omega_S \times \Omega_T$, and $\varepsilon > 0$ is the regularization constant.

If we specialise ϕ to the uniform distribution, $\mathscr{U}$, with support in $\Omega_S \times \Omega_T$, then (2.4) is the widely adopted Boltzmann-Shannon entropy-regularized OT problem [10], [12]:

$$\pi^o_{OT,\varepsilon,\mathscr{U}}(x,y|\mathscr{K}) \equiv \underset{\pi\in\Pi_{\mathscr{K}}}{\operatorname{argmin}} \left\{\int_{\Omega_S\times\Omega_T} c(x,y)\pi(x,y)d\lambda(x,y) + \varepsilon KL(\pi||\mathscr{U})\right\}. \tag{2.5}$$

It is worth noting that the KLD in (2.4) can be generalised to other entropies and more general Bregman divergences, as proposed in [12].

By introducing an entropy term, the Kantorovitch problem becomes strongly convex, and efficient iterative scaling algorithms can be used to compute the regularized transport plan solving (2.4) (namely, Sinkhorn-Knopp in the discrete case [22], Fortet in the continuous case [13]).

The main reason for introducing an entropy regularizer in OT is the availability of these computationally efficient algorithms. However, it is important to remember that the unique solution of (2.4) has minimum entropy relative to a distribution—called the ideal distribution, π_I—induced by the objective, and parameterized by ϕ, c and ε. This insight—and the benefits that flow from it—are the focus of the FPD-OT reformulation of the regularized OT problem (Section 4).

Particular applications may call for OT plans with specific structures, and/or the preservation of specific properties of the source marginal, μ, under transport to the target, ν. *Structure* may refer to the semantics of protected attributes (e.g. age, gender, ethnicity) in a classifier, spatial correlation in an image, neighbourhood structure in a graph, etc. It can also be expressed as stochastic constraints on the distributions (μ, ν, π) themselves. Existing methodologies to address structured OT rely either on the notion of sub-modular functions (in particular sub-modular cost functions with diminishing returns) [1] or the addition of a regularization term (group Lasso, Laplacian) [8, 15] in order to encourage specific structure-preserving mappings over others. For example, in [14], the authors propose a Laplace regularization scheme for colour transfer in image processing. Laplace regularization encourages transport maps that preserve the graph structure of the nodes in the source domain. In the current paper, we will approach this problem in a different way, by imposing detailed structure on the ideal distribution mentioned in the previous paragraph, and in a way that is sensitive to the cost of such constraints (Section 6).

Next, we introduce fully probabilistic design (FPD), and show that this accommodates regularized OT as a special case. Resetting regularized OT as an FPD problem—which we call FPD-OT—will facilitate the processing of structured knowledge constraints (Section 5) and cost-sensitive ideals (Section 6).

3. Fully Probabilistic Design (FPD)

Fully probabilistic design (FPD) is the axiomatically justified framework for designing probability models under uncertainty [17, 23], and is consistent with the rules of Bayesian decision-making [3]. It generalizes classical Bayesian conditioning and Bayes' rule, allowing the processing of probabilistic knowledge constraints, $\mathscr{K}$, into the conditional distribution, $\pi(x,y|\mathscr{K})$, in cases where the joint distribution, $\pi(x,y,\mathscr{K})$, is unavailable [23].

The axiomatic formulation of FPD as a distributional design problem was first established in [17], where the authors proved that it is an extension of Bayesian decision making. Later, the FPD framework was extended to hierarchical Bayesian models in [23], yielding a stochastic model of the uncertain joint distribution, $\pi(x,y|\mathscr{K})$.

More formally, FPD seeks a distribution, $\pi(x,y|\mathscr{K})$, which satisfies predefined design constraints, formalized as membership of a set, $\Pi_{\mathscr{K}}$, of knowledge-constrained distributions:

$$\pi(x,y|\mathscr{K}) \in \Pi_{\mathscr{K}}. \tag{3.1}$$

With $\Pi_{\mathscr{K}}$ being a singleton only in special cases, the ranking of choices is necessary. In FPD, this relies on the notion of an *ideal design*, which encodes the designer's zero-loss choice of $\pi(x,y|\mathscr{K})$. The ideal design, denoted by $\pi^I(x,y|\mathscr{K})$, does not satisfy the constraints in $\Pi_{\mathscr{K}}$, except in the trivial case; i.e. we assume that $\pi^I(x,y|\mathscr{K}) \notin \Pi_{\mathscr{K}}$. To compute the optimal solution, a utility (loss) function is then used to compare and rank the candidate distributions, based on their degree of closeness to the ideal design. In [3], the KLD is shown to be the

expected utility for ranking these distributional preferences consistent with decision-theoretic foundations:

$$\pi^{o}_{FPD,\pi^I}(x,y|\mathscr{K}) \equiv \underset{\pi\in\Pi_{\mathscr{K}}}{\operatorname{argmin}} \left\{ KL(\pi||\pi^I) \right\} \ , \ \ \pi \ll \pi^I. \tag{3.2}$$

The ideal design in (3.2) is the second argument of the KLD , and, as such, is the (infeasible) zero-KLD datum against which all possible distributions, consistent with the constraint set $\Pi_{\mathscr{K}}$, are ranked [23].

Given that both the Kantorovitch OT and the FPD problems seek the optimal design of a $\mathscr{K}$-consistent joint distribution (3.1), it is natural to investigate possible connections between the two frameworks. We will establish this in the next section.

4. The FPD Framework for Regularized Kantorovich OT (FPD-OT)

The goal of this section is formally to establish the connection between FPD and regularized Kantorovich OT. We start with the general form of FPD-OT, then derive the special case of the entropy-regularized problem.

Theorem 4.1 (FPD-OT)**.** *Let $(\Omega_S, \mathfrak{F}_S, \mu)$ and $(\Omega_T, \mathfrak{F}_T, \nu)$ be two measure spaces, and $\Pi_{\mathscr{K}}$ be the set of joint distributions, $\pi(x,y|\mathscr{K})$, with support in $\Omega_S \times \Omega_T$, and with prescribed marginals, μ and ν (2.2). Let ϕ be a fixed distribution, which dominates π, i.e., $\pi \ll \phi$, let $c(x,y) \geq 0$ be a measurable cost function, also with support in $\Omega_S \times \Omega_T$, and let $\varepsilon \in \mathbb{R}^+$ be a regularization term. Then the unique solutions of objectives (2.4) and (3.2) are (a.s.[1]) equal,*

$$\pi^{o}_{FPD,\pi^I}(x,y|\mathscr{K}) = \pi^{o}_{OT,\varepsilon,\phi}(x,y|\mathscr{K}), \tag{4.1}$$

if the ideal design, $\pi^I(x,y|\mathscr{K})$, has an extended Gibbs form, defined as

$$\pi^I(x,y|\mathscr{K}) \overset{def.}{=} \frac{1}{K_{\phi,\varepsilon}} \phi(x,y) \exp\left(\frac{-c(x,y)}{\varepsilon}\right). \tag{4.2}$$

Here, $K_{\phi,\varepsilon} \overset{def.}{=} \int_{\Omega_S\times\Omega_T} \phi(x,y) \exp\left(\frac{-c(x,y)}{\varepsilon}\right) d\lambda(x,y)$ is the normalizing constant.

Proof. From (2.4), we have

$$\begin{aligned} \pi^{o}_{OT,\varepsilon,\phi}(x,y|\mathscr{K}) &\overset{\text{def.}}{=} \underset{\pi\in\Pi_{\mathscr{K}}}{\operatorname{argmin}} \int_{\Omega_S\times\Omega_T} \pi(x,y) \log\left\{ \frac{\pi(x,y)\exp(\frac{c(x,y)}{\varepsilon})}{\phi(x,y)} \right\} d\lambda(x,y) \\ &= \underset{\pi\in\Pi_{\mathscr{K}}}{\operatorname{argmin}} KL(\pi||\pi^I) \\ &\overset{(3.2)}{=} \pi^{o}_{FPD,\pi^I}(x,y|\mathscr{K}), \end{aligned} \tag{4.3}$$

in the case where the ideal design is as specified in (4.2). □

We note the following:

- Theorem 4.1 recasts relative-entropy-regularized Kantorovich OT as a specialization of fully probabilistic design (FPD).

[1] The a.s. equality of distributions is assumed throughout.

- The main reason why the additive KLD term (2.4) is adopted in conventional OT is because it strongly convexifies the objective, yielding sufficient conditions for convergence of efficient iterative schemes to the unique minimizer. What has not been reported before is the way in which this KLD regularization shapes the zero-loss (i.e. datum) OT plan—being the ideal design, $\pi^I(x,y|\mathscr{K})$ (4.2)—in a prescribed way.
- Specifically, the optional regularizing term in (2.4) furnishes the base distribution, $\phi(x,y)$, of an ideal design, $\pi^I(x,y|\mathscr{K})$, which is prescriptively of the Gibbs type (4.2). Meanwhile, the pre-prior-imposed cost of transportation, $c(x,y)$, acts as the energy term in the Gibbs structure, modulating the base distribution, $\phi(x,y)$, with ε acting as the (proportional) temperature parameter.
- $\phi(x,y)$ can be chosen judiciously by the designer in order to mitigate the (pre-imposed) cost of transportation, $c(x,y)$, and thereby reduce the expected loss (being the KLD) incurred by the optimal OT plan, $\pi^o_{OT,\varepsilon,\phi}(x,y|\mathscr{K})$ (3.2). We will present an example of this cost-sensitive design in Section 6.
- We refer to this FPD setting for regularized OT—and this distinct role for the regularizing kernel, $\phi(x,y)$, in shaping the ideal design of the transport plan—as FPD-OT.

Remark 4.2. The ideal design, $\pi^I(x,y|\mathscr{K})$ (4.2), in FPD acts as the designer's zero-loss choice for the joint transport plan, $\pi(x,y|\mathscr{K})$, and may be interpreted as a pre-prior in generalized Bayesian inference via FPD. As already noted, this ideal is typically unattainable, in that it fails to satisfy the marginal constraints; i.e. $\pi^I(x,y|\mathscr{K}) \notin \Pi_{\mathscr{K}}$ (2.2). Its role is to induce an expected-loss ranking (equivalent to KLD-ordered preferences [3]) of the elements of $\Pi_{\mathscr{K}}$ [17].

Remark 4.3. KLD is 1-strongly convex, and so the problem stated in (3.2) is ε-strongly convex, thereby yielding a unique solution.

Remark 4.4. When $\varepsilon \to \infty$ in (4.2), $\pi^I \to \phi$; i.e.,

$$\pi^o_{FPD,\pi^I}(x,y|\mathscr{K}) \xrightarrow{\varepsilon\to\infty} \pi^o_{FPD,\phi}(x,y|\mathscr{K}). \tag{4.4}$$

Note that the strong convexity of the KLD objective of FPD (3.2) is lost in the $\varepsilon \to 0$ limit; i.e. the (unregularized) Kantorovich OT problem cannot be expressed as an FPD problem.

Remark 4.5. The Boltzmann-Shannon entropy-regularized OT problem (2.5) is the specialization of FPD-OT in the case where $\pi^I(x,y|\mathscr{K})$ is the Boltzmann distribution, with the transportation cost, $c(x,y)$, as the energy functional, and the regularization term, ε, as the (proportional) temperature, as follows.

Corollary 4.6. *The entropy-regularized OT problem* (2.5) *is a specialization of FPD, where the ideal design, π^I, reduces to the Boltzmann distribution:*

$$\pi^o_{OT,\varepsilon,\mathscr{U}}(x,y|\mathscr{K}) = \pi^o_{FPD,\pi^I}(x,y|\mathscr{K}) \tag{4.5}$$

for the ideal assignment,

$$\pi^I(x,y|\mathscr{K}) \overset{def.}{=} \frac{1}{K_\varepsilon}\exp\Big(\frac{-c(x,y)}{\varepsilon}\Big). \tag{4.6}$$

5. AN EXAMPLE OF FPD-OT: PROCESSING A RELAXATION OF THE MARGINAL CONSTRAINTS

By way of relaxing the standard OT formulation (2.1, 2.2), let us now assume that the marginals of the OT plan, π, are confined to KLD balls, centred around μ and ν, which are *fixed* (i.e. nominal) distributions. The respective KLD ball radii (i.e. KLD upper bounds) are $\eta \geq 0$ and $\zeta \geq 0$, respectively. The knowledge-constrained set, denoted here by $\tilde{\Pi}_{\mathscr{K}}(\eta,\zeta)$, is now the following superset (relaxation) of (2.2):

$$\tilde{\Pi}_{\mathscr{K}}(\eta,\zeta) \stackrel{\text{def.}}{=} \left\{ \pi \in \mathscr{P}(\Omega_S \times \Omega_T) \mid KL(P_{\Omega_S\#}\pi||\mu) \leq \eta \ , KL(P_{\Omega_T\#}\pi||\nu) \leq \zeta \right\}. \tag{5.1}$$

In this way, the parameters, η and ζ, of $\tilde{\Pi}_{\mathscr{K}}(\eta,\zeta)$ encode prior knowledge constraints, along with μ and ν. Note that no stochastic model is posited[2] for π or its push-forwards, $P_{\Omega_S\#}\pi$ and $P_{\Omega_T\#}\pi$. As before, the second prior input to the FPD formalism is the ideal design, $\pi^I(x,y|\mathscr{K})$ (3.2). In this relaxed setting, the FPD-OT primal optimization problem reads as follows:

$$\pi^o_{\eta,\zeta}(x,y|\mathscr{K}) \stackrel{\text{def.}}{=} \underset{\pi\in\tilde{\Pi}_{\mathscr{K}}(\eta,\zeta)}{\operatorname{argmin}} \left\{ KL(\pi||\pi^I) \right\}. \tag{5.2}$$

The associated Lagrangian is

$$\mathscr{L}(\pi,\mathscr{V}) \stackrel{\text{def.}}{=} KL(\pi||\pi^I) + \alpha(KL(P_{\Omega_S\#}\pi||\mu) - \eta) + \beta(KL(P_{\Omega_T\#}\pi||\nu) - \zeta), \tag{5.3}$$

where $\mathscr{V} \stackrel{\text{def.}}{=} (\alpha,\beta) \succcurlyeq 0$ are the Lagrange multipliers.

For technical ease, let us assume that the support of π in the product space, $\Omega_S \times \Omega_T$ (2.1), is finite and equipped with the counting measure, so that π is expressible as a pmf; i.e. $\pi \in \mathscr{P}(\Omega_S \times \Omega_T)$ is the probability simplex of finite dimension. Since the KLD is a convex functional of π, the primal problem (5.2) is convex, with unique solution, π^o. Furthermore, the problem satisfies Slater's constraint qualification (see Section 5.3.2 of [4]); i.e. there exists at least one element of (5.1) for which the inequalities there are strict, an example being the product distribution, $\mu \otimes \nu$. These conditions are sufficient for strong duality, i.e.

$$\min_{\pi\in\tilde{\Pi}_{\mathscr{K}}(\eta,\zeta)} \left\{ KL(\pi||\pi^I) \right\} = \max_{\mathscr{V}\succcurlyeq 0} \min_{\pi\in\mathscr{P}(\Omega_S\times\Omega_T)} \mathscr{L}(\pi,\mathscr{V}). \tag{5.4}$$

Denote the dual optimum by

$$\mathscr{V}^*(\eta,\zeta) \stackrel{\text{def.}}{=} (\alpha^*,\beta^*) \stackrel{\text{def.}}{=} \underset{\mathscr{V}\succcurlyeq 0}{\operatorname{argmax}} \min_{\pi\in\mathscr{P}(\Omega_S\times\Omega_T)} \mathscr{L}(\pi,\mathscr{V}), \tag{5.5}$$

and so

$$\begin{aligned} \pi^o_{\eta,\zeta}(x,y|\mathscr{K}) &\stackrel{\text{def.}}{=} \underset{\pi\in\mathscr{P}(\Omega_S\times\Omega_T)}{\operatorname{argmin}} \mathscr{L}(\pi,\mathscr{V}^*) \\ &= \underset{\pi\in\mathscr{P}(\Omega_S\times\Omega_T)}{\operatorname{argmin}} \left\{ KL(\pi||\pi^I) + \alpha^* KL(P_{\Omega_S\#}\pi||\mu) + \beta^* KL(P_{\Omega_T\#}\pi||\nu) \right\}. \end{aligned} \tag{5.6}$$

Note, from (5.3, 5.5), that the regularization constants in objective (5.6) are $\alpha^* \equiv \alpha^*(\eta,\zeta)$ and $\beta^* \equiv \beta^*(\eta,\zeta)$, i.e. deterministic functions of the prior knowledge-constraints (KLD-ball radii), η and ζ (5.1). For this reason, $\eta\alpha^*(\eta,\zeta)$ and $\zeta\beta^*(\eta,\zeta)$ are (finite) constants.

[2] See Section 7.3 for discussion of a future hierarchical FPD attack on this problem.

We conclude that the FPD-OT problem—in the case (5.1) which conditions on (i.e. processes) knowledge in the form of KLD-ball constraints around nominal marginals, μ and ν—is therefore equivalent to the unconstrained minimization of the regularized objective (5.6). The latter objective is the one adopted in the classical unbalanced OT (i.e. UOT) problem [2, 7], but we emphasize that *our FPD-OT problem above (5.1, 5.2) is balanced* (see Remark 5.1, below). A dividend of the FPD-OT formulation of the regularized OT problem (5.6) is the interpretability of its input parameters, η and ζ, as knowledge constraints (i.e. KLD ball radii), something which is not possible for α^* and β^* in (5.6).

Finally, we note that the FPD-OT problem in (5.1, 5.2) specializes in obvious ways at extreme values of the KLD ball radii, η and ζ:

(i) Consider the case in which the KLD-ball radii, η and ζ, are both set to zero. It follows directly from (5.3) and (5.5) that $\alpha^*(0,0) \to \infty$ and $\beta^*(0,0) \to \infty$ for the general case of unconstrained π in (5.6). This forces the minimizer to belong to the set in which the second and third KLD terms in (5.6) are both zero (i.e. the two push-forwards are identically μ and ν, respectively). Hence, in this case, the relaxed FPD-OT problem in (5.6) specializes to one of minimizing $KL(\pi||\pi^I)$, in the set (2.2), which is the original FPD-OT problem (3.2).

(ii) The (trivial) case—in which no knowledge constraints are imposed on π—arises when the KLD-ball radii are unboundedly large. Then, the ideal design is attained:

$$\pi^o_{\eta,\zeta}(x,y|\mathscr{K}) \xrightarrow{\eta\to\infty,\ \zeta\to\infty} \pi^I(x,y|\mathscr{K}) \tag{5.7}$$

Remark 5.1. The processing of uncertainty bounds—in the form of KLD-ball radii, η and ζ, in (5.1)—can be understood as a contribution to robust OT, in a manner similar to the set-up for unbalanced OT [2, 7]. To be clear, however: unit (normalizing) mass is *conserved* in our example, and so the transport is *balanced*. The purpose of the example in Section 5 is to demonstrate how FPD-OT can formulate and process relaxations (5.1) of the conventional knowledge constraints of OT (2.2). A more mature response to the problem of robust OT—involving the elicitation of uncertainty in the optimal plan, π^o (4.1)—will be addressed in future work via hierarchical FPD [23]. See Section 7.3 for further comment on this future direction.

6. Simulation: Influence of the ideal base measure, ϕ

In this Section, we will compute OT plans for various choices of the regularizing distribution, $\phi(x,y)$ (2.4), now repurposed in FPD-OT as the base distribution of our ideal design, $\pi^I(x,y)$ (4.2). The composition of $\phi(x,y)$ and the cost function, $c(x,y)$, in $\pi^I(x,y)$ will facilitate the elicitation of richer structures in the OT plan, providing an alternative to techniques based only on notions of regularization and sub-modular functions. Indeed, the cost function, $c(x,y)$, in (2.4) is generally dictated by the physical or geometrical constraints of the underlying metric space (the ground metric [12]). In contrast, the regularizing distribution, $\phi(x,y)$, can be used to express subjective and subsidiary design preferences, via (4.2), including cost-sensitive choices, $\phi \equiv \phi(c)$. By rewriting the regularized OT objective in (2.4) as the FPD-OT objective (3.2, 4.2)—i.e. as a KLD minimization problem—we explicitly reveal the role of $\phi(x,y)$ as the ideal base distribution, modulated by a cost-dependent Gibbs term.

As a motivating example, we consider the problem of energy transportation from a source domain (producers) to a target domain (consumers) with the objective of minimizing the total

(A) (B) (C)

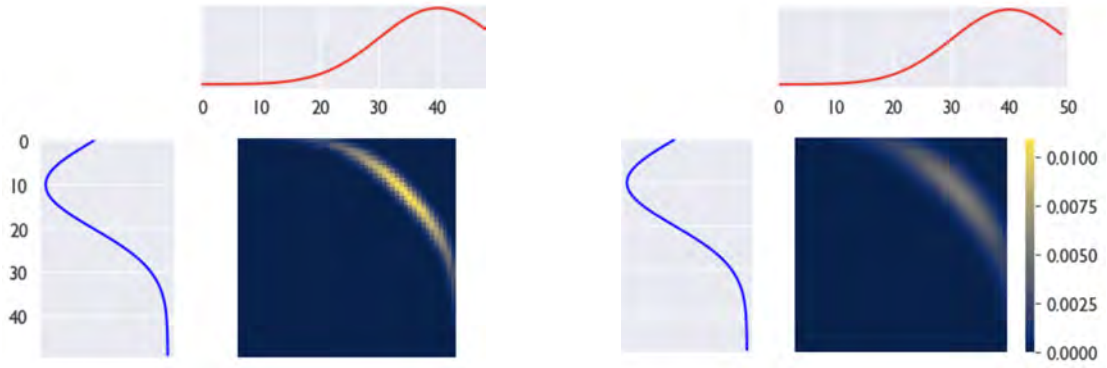
FIGURE 1. (A) Quadratic cost matrix, $c(x,y)$; (B) regions of low- and high-cost pathways, (x,y), as defined in (6.5); and (C) binary map of quenched (0) and unquenched (1) pathways when $\phi(x,y)$ is the cost-sensitive Bernoulli random matrix realization (6.2, 6.3), with $\alpha = 1$.

transportation cost, and, consequently, the carbon footprint of the electric grid [21]. Responding dynamically to the demand, and adjusting the transportation plan, is a key element in the optimization of the overall carbon budget. In addition to the cost, we may need to implement additional constraints, imposed by technical and physical considerations. For example, we may not be able to use all transportation pathways simultaneously for capacity reasons, or we may want to block (i.e. quench) a proportion of the available pathways. Designing OT plans with particular capacity constraints was studied in [18], where the set of feasible solutions, $\Pi_{\mathscr{K}}$ (2.2), was dominated by a fixed transportation plan, modelling the capacity constraint. The authors showed that the resulting OT plan was sparse. In the FPD-OT approach in Section 6.2, we instead impose sparsity constraints in the optimization procedure via zeroes in the base distribution, $\phi(x,y)$, of the ideal (4.2). We achieve this in either a cost-insensitive way (Section 6.2.1) or as a function of the cost metric (Section 6.2.2).

In Section 6.1, we choose ϕ with a sampled Gaussian profile, which includes the special case of uniform ϕ. Then, in Section 6.2, we realize ϕ from a Bernoulli random matrix process. This choice of ϕ allows the implementation of capacity constraints in the OT plan (above), where only a proportion $\theta < 1$ of the pathways are active. We aim for a constant proportion, $\theta = 0.85$. Note that—although we simulate the base distribution, ϕ, from an appropriate prior in Section 6.2 (i.e. as a realization of a random process satisfying the structural constraints we wish to impose on π via π^I (4.2))—this prior is *not* part of the knowledge structure (2.2) processed by FPD-OT.

In both experiments (Sections 6.1 and 6.2), we adopt the following settings:

- $\Omega_S \equiv \{0,\dots,m-1\}$, $m \equiv 50$; $\Omega_T \equiv \{0,\dots,n-1\}$, $n \equiv 50$;
- $c(x,y) \equiv ||x-y||^2$ (Fig. 1 (A));
- $\varepsilon \equiv 10^{-2}$;
- Maximum iterations of the Sinkhorn-Knopp (SK) algorithm [9] $\equiv 1000$;
- Stopping threshold for SK $\equiv 10^{-9}$.



(A) $\Sigma_r = \Sigma_1$ (B) $\Sigma_r = 100 \times \Sigma_1$

FIGURE 2. OT plans in the case of ideal base pmf, $\phi \propto \mathcal{N}(\upsilon, \Sigma_r)$, where $\Sigma_r \equiv r\Sigma_1$, for two cases of $r > 0$: (A) $r \equiv 1$ and (B) $r \equiv 100$.

6.1. **Ideal base pmf, $\phi \propto \mathcal{N}(\upsilon, \Sigma_r)$ (sampled Gaussian).** In this first experiment, we choose the marginals as follows ($\propto$ denotes 'proportional to'):

- $\mu(x) \propto \mathcal{N}(10, 10)$, i.e. the pmf defined by evaluating a scalar Gaussian pdf at $x \in \Omega_S$;
- $\nu(y) \propto \mathcal{N}(40, 10)$, i.e. the pmf defined by evaluating a scalar Gaussian pdf at $x \in \Omega_T$.

Furthermore, $\phi(x, y)$—the base pmf in the Gibbs ideal design (4.2)—is chosen to be the pmf induced by a bivariate Gaussian pdf, when confined to the support $\Omega_S \times \Omega_T$. The mean of the underlying bivariate Gaussian is chosen as $\upsilon \equiv [20, 20]^T$ (where the superscript denotes transposition). Its covariance matrix is parameterized as $\Sigma_r \equiv r\Sigma_1 \in \mathbb{R}^{2\times 2}$, with

$$\Sigma_1 \equiv \begin{bmatrix} 2 & 2 \\ 2 & 6 \end{bmatrix},$$

and we will vary $r \in \mathbb{R}^+$, in order to study its effect on the smoothness of the OT plan. Inserting this base pmf, along with the quadratic cost metric, $c(x, y)$, into (4.2), we obtain the ideal pmf,

$$\pi^I(x, y|\mathscr{K}) \propto \exp\Big\{-(\mathbf{x}-\mathbf{m})^T \Big(\frac{1}{2r}\Sigma_1^{-1} + \frac{1}{\varepsilon}\begin{bmatrix} +1 & -1 \\ -1 & +1 \end{bmatrix}\Big)(\mathbf{x}-\mathbf{m})\Big\}, \tag{6.1}$$

where $\mathbf{x} \equiv [x, y]^T$, and the ideal mean, $\mathbf{m}$, is a (deterministic) function of the remaining parameters. This reveals the fact that the regularization constant, ε (2.4), and the variance controller, r, of the base distribution, ϕ, both have the same role as temperature (annealing) parameters of the Gibbs-type ideal plan (4.2). The entropy (smoothness) of the induced OT plan therefore increases as either (or both) are increased. This is corroborated empirically by the results shown in Fig. 2, where we display the OT plan obtained for two values of $r \in \{1, 100\}$. The former yields an OT plan relatively concentrated on the graph of the corresponding Monge map [30], whereas the latter yields an OT plan with high entropy. Indeed, as $r \to \infty$ in (4.2), then $\pi^I(x, y) \to K_\varepsilon^{-1} \exp\Big(\frac{-c(x,y)}{\varepsilon}\Big)$ (4.6), and we recover the solution of the entropy-regularized OT problem (2.5), corresponding to the case where ϕ is the uniform pmf, i.e. $\phi \equiv \mathscr{U}$. The resulting OT plan then has maximum entropy among all members of the knowledge-constrained set of transport plans, $\Pi_{\mathscr{K}}$ (2.2) [10].

6.2. **Ideal base pmf, ϕ, with elements realized as i.i.d. Bernoulli r.v.s.** In this experiment, we again adopt marginal pmfs with a Gaussian profile, but this time they are chosen equal:

- $\mu(x) \propto \mathcal{N}(20,10)$, i.e. the pmf defined by sampling a scalar Gaussian pdf at $x \in \Omega_S$;
- $\nu(y) \propto \mathcal{N}(20,10)$, i.e. the pmf defined by sampling a scalar Gaussian pdf at $x \in \Omega_T$.

Our goal now is to study the FPD-OT problem when ϕ is a realization of a Bernoulli random matrix. As mentioned earlier, this design choice is of practical relevance, since it allows only some OT pathways, (x,y) to be active, with the rest being quenched (i.e. zeroes for certain pathways). A practical example is the optimization of an electricity grid designed to match producers to consumers of electricity, and in a situation where only a subset of the producer-consumer pathways can be used. The expected proportion of active pathways, denoted by θ, is dictated by the technical characteristics of the grid. In the simulations below, we fix $\theta = 0.85$.

In order to select the active OT pathways, we study two different settings:

(1) In Section 6.2.1 below, we choose ϕ as a realization of a Bernoulli random matrix with i.i.d. (i.e. independent, identically distributed) Bernoulli entries (i.e. transport pathways), each with parameter θ; i.e. each pathway in the OT plan independently has probability θ of being active *a priori*[3]. We denote this by $\phi(x,y) \overset{\text{iid}}{\sim} Bern(\theta)$ below.

(2) In Section 6.2.2 below, we impose more structure on the problem. Once again, we assume independent Bernoulli pathways in ϕ, but now with probabilities that are a decreasing function, $\theta_c(x,y)$, of the cost, $c(x,y)$. We denote this by $\phi(x,y) \overset{\text{id}}{\sim} Bern(\theta_c)$ below. We still require that the proportion of active pathways in the OT plan be equal to θ, as explained below. By choosing $\theta_c(x,y)$ to be cost-sensitive, we are effectively encoding spatial correlation in the Bernoulli random matrix.

6.2.1. $\phi(x,y) \overset{\text{iid}}{\sim} Bern(\theta)$. Given that all pathways, (x,y), are quenched will equal probability, $1-\theta$, we are imposing a notion of fairness, in the sense that no pathway is favoured (or penalized) more than any other. The i.i.d. Bernoulli r.v.s realizing the base distribution, ϕ, are therefore

$$\begin{cases} \phi(x,y) = 1 & \text{with probability } \theta, \\ \phi(x,y) = 0 & \text{with probability 1- } \theta. \end{cases} \tag{6.2}$$

The resulting OT plan—with quenched paths uniformly distributed across the domain, at an expected rate of 0.15—is illustrated in Fig. 3. Note how both low-cost and high-cost pathways have the same probability (0.85) of being active.

6.2.2. $\phi(x,y) \overset{\text{id}}{\sim} Bern(\theta_c)$. In this case, we assume that the pathways of the OT plan are quenched independently but not identically (id), by realizing the ideal base pmf entries, $\phi(x,y)$, as Bernoulli r.v.s with parameters $\theta_c(x,y)$, chosen to be a decreasing function of the cost, $c(x,y)$. This yields an OT plan where high-cost pathways are penalized by being assigned lower probabilities of being active *a priori*. It is worth noting that this non-uniform choice of θ_c enables the modelling of spatial correlations in the OT plan, where clusters of entries with similar transport cost (and, therefore, spatial neighbours) have similar activation probabilities. There exist other techniques

[3]Since the iterative SK algorithm is initialized by the ideal, $\pi^I(x,y|\mathscr{K})$ (4.2), the (exact) zeroes of $\pi^o_{FPD,\pi^I}(x,y|\mathscr{K})$ (4.1) are equal to those of $\phi(x,y)$ for a finite number of SK iterations [9].

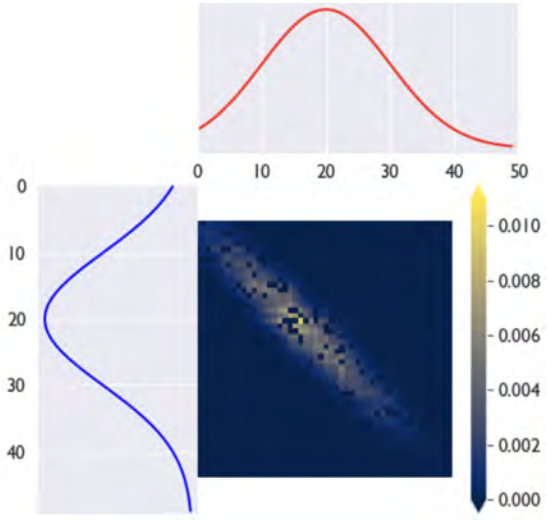


FIGURE 3. OT plan for ideal base pmf, $\phi(x,y) \overset{\text{iid}}{\sim} Bern(\theta)$, $\theta = 0.85$.

for modelling spatial interactions in Bernoulli random matrices, among them the Ising model and the multivariate Bernoulli distribution [11, 29]. In our simulation, we adopt the following strictly decreasing function of the cost, $c(x,y)$, parameterized by $\alpha \geq 0$ and $\beta > 0$:

$$\theta_c(x,y) \equiv \beta \exp(-\alpha c(x,y)). \tag{6.3}$$

$\beta \in (0,1]$ is designed to satisfy the constraint that the proportion of active pathways in $\Omega_S \times \Omega_T$ be $\theta = 0.85$:

$$\beta \equiv \frac{m^2 \theta}{\sum\limits_{(x,y)\in\{0,\ldots,m-1\}^2} \exp(-\alpha c(x,y))} \leq 1. \tag{6.4}$$

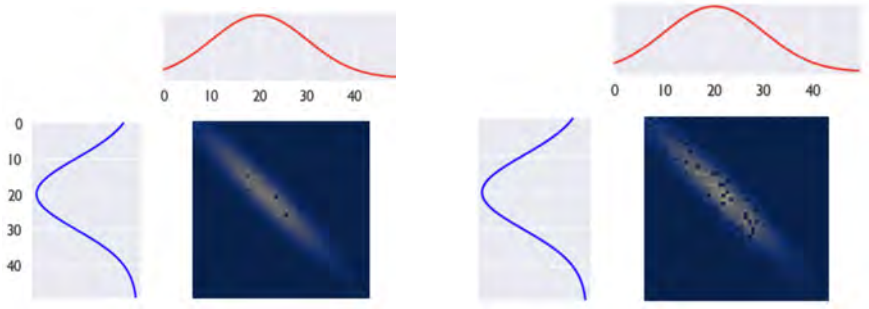
We visualize the OT plans for four cases of $\alpha \in \{1.0, 0.5, 0.1, 0.05\}$, in Fig. 4. Lower values of α induce more high-cost active pathways in the OT plan. When $\alpha \to 0$, we recover the i.i.d. Bernoulli matrix realization of $\phi(x,y)$ in Section 6.2.1, with $\theta_c(x,y) \to \beta \equiv \theta$ (6.3). Therefore, α provides control over the proportion of quenched high-cost pathways.

Conversely, $\alpha = 1$ yields OT plans with quenched pathways mostly located remotely from the main diagonal of the plan, being the regions of high cost (Fig. 4(A)). In Fig. 4(B)–(D), we can clearly see that decreasing α causes the quenched pathways to concentrate less in the high-cost region of the plan and more of them to concentrate in the low-cost region (near the main diagonal). Since the proportion of quenched pathways is being held constant at $1-\theta = 0.15$, it follows that lowering α sweeps the quenched pathways away from the high-cost regions towards the low-cost region. For convenience, we define the high-cost threshold to be

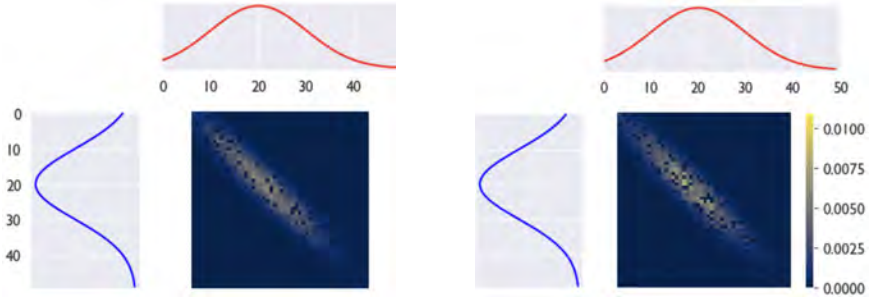
$$c(x,y) \geq \ \bar{c} + 1.9 c_0, \tag{6.5}$$

where $\bar{c}$ and c_0 are the average and standard deviation, respectively, of the pathway costs, $c(x,y)$. With this definition, the number of high-cost entries is equal to $\kappa = 182$. Fig. 1(B) shows the binary map of high- and low-cost entries in the transport map, induced by the quadratic cost function, $c(x,y)$.

In Fig. 5, we plot the proportion of high-cost pathways that are quenched as a function of $\alpha \in [0,1]$. Here, we randomize $\phi(x,y)$ over 100 Monte Carlo trials, and graph the average



(A) $\alpha = 1$ (B) $\alpha = 0.5$



(C) $\alpha = 0.1$ (D) $\alpha = 0.05$

FIGURE 4. OT plans for ideal base pmf, $\phi(x,y) \overset{\text{id}}{\sim} Bern(\theta_c)$, with cost-sensitive θ_c (6.3), and for $\alpha \in \{1.0, 0.5, 0.1, 0.05\}$.

proportion of high-cost pathways that are quenched, with the corresponding standard deviation. This confirms our design aim: that increasing α increases the proportion of quenched pathways among those of high cost. In contrast, small values of α yield OT plans in which the quenched pathways are distributed uniformly, so that the proportion of quenched high-cost pathways drops to $1 - \theta = 0.15$.

In these examples, we have shown how FPD-OT facilitates the design of an ideal (i.e. zero-loss, but unattainable) distribution (4.2) with a cost-dependent base distribution, ϕ, chosen as a realization of a non-stationary, cost-sensitive Bernoulli field (6.2, 6.3). This has enabled sophisticated, multi-objective design constraints to be satisfied, in this case the concentration of quenched (i.e. zero-transport) (x,y) paths into high-cost regions of the plan, while maintaining a constant average rate, $\theta = 0.85$, of active paths. As we increase α in (6.3), we can push more of the quenched paths into these high-cost regions (Fig. 5). Conversely, as we dial α down to zero in (6.3), then $\theta_c \to \beta \equiv 0.85$ (in this simulation), and the design reverts to the cost-independent case (Fig. 3(A)). It is the direct interaction between the regularizing base distribution, $\phi(x,y)$, and the cost metric, $c(x,y)$, in the ideal distribution (4.2) that facilitates this kind of design. The classical regularized Kantorovich setting of the OT problem (2.4) renders such multi-objective, cost-dependent designs harder to achieve, perhaps explaining why $\phi(x,y)$ in (2.4) has not been actively exploited in structured OT. Instead, the choice, $\phi \equiv \mathscr{U}$, of entropic OT is the usual default.

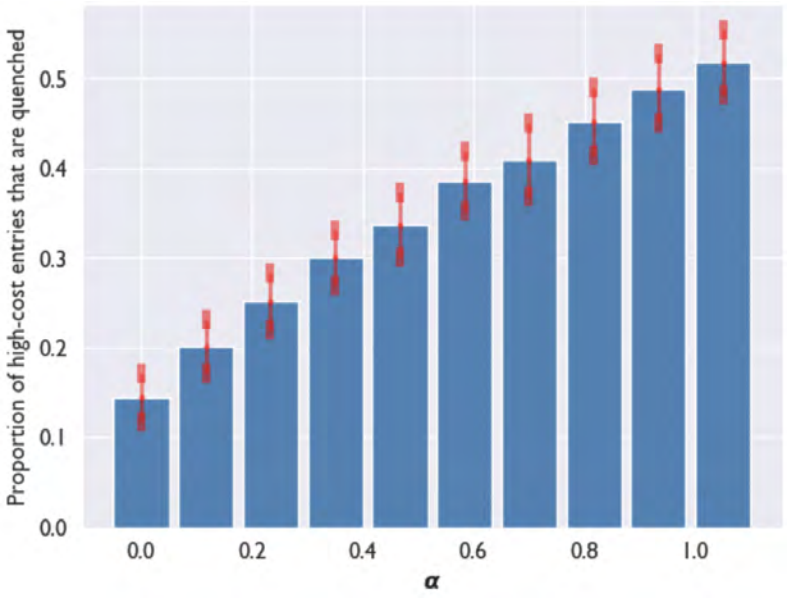


FIGURE 5. Proportion of high-cost entries that are quenched, as a function of α. The blue bars represent the average proportion obtained from 100 Monte Carlo simulations, with the corresponding standard deviations in red.

7. DISCUSSION

The purpose of this paper has been to recast the regularized Kantorovich optimal transport (OT) problem (2.4) as one of fully probabilistic design (FPD) (3.2). A number of important conceptual and practical benefits flow from this FPD-OT framework, as summarized next:

7.1. **The benefits of an FPD setting of regularized Kantorovich OT.** FPD (3.2) is the minimum-KLD projection of the ideal distribution, π^I, into the set, $\pi_{\mathscr{K}}$ (2.2), constrained by the fixed marginals, μ and ν. In this sense, it specifies an optimal update of π^I when processing these knowledge constraints:

$$\pi^I(x,y|\mathscr{K}) \xrightarrow{\mu,\nu} \pi^o_{OT,\varepsilon,\phi}(x,y|\mathscr{K}) \tag{7.1}$$

In this way, the Gibbs-type OT ideal transport plan (4.2) acts as the pre-prior, yielding the joint distribution, $\pi^o_{OT,\varepsilon,\phi}(x,y|\mathscr{K})$ (4.1), as the optimally and sequentially $\mathscr{K}$-conditioned joint model. The alignment of regularized OT to the rich context and literature of FPD [3, 16, 23] is not well known currently. It provides a more mature justification—beyond the usual regularization notions of smoothness and computational convenience—for designing minimum-KLD plans in OT, in place of (unregularized) plans which attain a Wasserstein distance between μ and ν.

7.2. **Cost-dependent ideal design.** The FPD-OT example in Section 6 has made clear the potential for this resetting of the regularized Kantorovich OT problem to reveal new, structured OT plans. In particular, the facility to trade off the ideal base distribution, $\phi(x,y)$, against the pre-specified cost of transportation, $c(x,y)$, in specifying the ideal transport plan, $\pi^I(x,y|\mathscr{K})$ (4.2), points to interesting new criteria for finding lower Bayes-risk (i.e. lower KLD) OT designs (4.1) (see the fourth bullet point after Theorem 4.1).

7.3. **Future opportunities in hierarchical FPD-OT.** OT—and its FPD-OT setting in this paper—involves the choice of an uncertain, $\mathscr{K}$-constrained augmented model, $\pi^o_{OT,\varepsilon,\phi}(x,y|\mathscr{K})$, via deterministic optimization (4.3). However, in a fully Bayesian setting, the Kantorovich plan, $\pi(x,y)$ is a random process [23], and must be equipped with a hyper-prior, $\pi(x,y) \sim \Pi$, whose design should process suitable relaxations of the marginal constraints of conventional OT (2.2). In this way, optimization is replaced by randomization. Several opportunities to explore new directions for OT then emerge. These include (i) the processing of—fully modelled—noisy and uncertain marginals (conferring robustness on the design); (ii) the formal quantification of uncertainty in $\pi(x,y)$; (iii) the opportunity to process nonlinear functionals of $\pi(x,y)$; and (iv) the deployment of the mature armoury of stochastic simulation tools. These propensities for hierarchical FPD-OT will be reported in future publications on this topic.

8. Conclusion

In this paper, we have recast the entropy-regularized Kantorovitch OT problem as one of fully probabilistic design (FPD). Probability models—confined to a knowledge-constrained set of alternatives—are ranked against a zero-loss ideal case, using Kullback-Leibler divergence (KLD) as the induced expected loss (i.e. risk) [3]. In effect, the optimization is reformulated as a problem of generalized Bayesian conditioning [17, 23], providing important inferential insights into the resulting designs. For instance, in Section 5, the regularization constants in objective (5.6) are recast as deterministic functions of KLD-ball radii. In Section 6, we showed how the regularizing distribution, $\phi(x,y)$, of conventional OT is recast in FPD-OT as the base distribution of a Gibbs-type ideal OT plan, $\pi^I(x,y|\mathscr{K})$, and so can be chosen to quench pathways between source-target pairs, (x,y), with high transport costs. More sophisticated spatial correlation structures might be satisfied by choosing the base distribution, $\phi(x,y)$ (4.2), as the realization of a Markov random field with cost-dependent clique potentials. In future work, we plan to investigate hierarchical relaxations of FPD [23] for OT (i.e. HFPD-OT). Important problems of robust OT and nonlinear moment processing will be accommodated by this framework.

Authors' Note

The fourth author (RS) first met Ezra during a visit to Imperial College London in 2003. Since then, their collaborations spanned many research topics: spectral theory of matrix products; robustness of switched linear systems; strict positive realness of linear systems; the stability of switched descriptor systems; and fractional dynamical systems. Their past work together spawned a community of international collaborators, including Oliver Mason, Shravan Sajjja, Martin Corless, Ted Davison, Yirmeyahu Kaminski, Kai Wulff, Paul Curran, Chris King and Selim Solmaz, some of who have contributed to this volume. These collaborations have also straddled industry and academia, starting with Imperial College London, then the Hamilton Institute, Technion, IBM Research, Holon Institute of Technology, University College Dublin, and finally back to Imperial College London, where it all began in 2003. Their journey together, over 20 years, has been extremely rich and fulfilling for RS, who mourns the recent passing of Ezra. He was a great scientist, a wonderful teacher, and a fantastic scholar, but, most of all, he was a wonderful human being: kind, gentle and compassionate in equal measure. RS has learned much from Ezra and it is a privilege for him to have been Ezra's co-author, colleague, and friend.

REFERENCES

[1] D. Alvarez-Melis, T. Jaakkola, S. Jegelka, Structured optimal transport, In: A. Storkey, F. Perez-Cruz (ed.), Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics, volume 84 of Proceedings of Machine Learning Research, pages 1771–1780. PMLR, 2018. URL https://proceedings.mlr.press/v84/alvarez-melis18a.html.

[2] J. D. Benamou, Y. Brenier, Mixed L 2-Wasserstein optimal mapping between prescribed density functions, Journal of Optimization Theory and Applications 111 (2001) 255-271.

[3] J. M. Bernardo, Expected information as expected utility, The Annals of Statistics, 7 (1979) 686-690.

[4] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, Cambridge, 2004. URL https://web.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf.

[5] F.R.S. Cayley, On Monge's "Mémoire sur la Théorie des Déblais et des Remblais.", Proceedings of the London Mathematical Society, s1-14 (1882) 139–143.

[6] Y. Chen, T. Georgiou, M. Pavon, Optimal transport in systems and control, Annual Review of Control, Robotics, and Autonomous Systems, 4 (2021) 89-113.

[7] L. Chizat, G. Peyré, B. Schmitzer, F.-X. Vialard, Scaling algorithms for unbalanced transport problems, arXiv: Optimization and Control, 2016. URL https://arxiv.org/pdf/1607.05816.pdf.

[8] N. Courty, R. Flamary, D. Tuia, A. Rakotomamonjy, Optimal transport for domain adaptation, IEEE Transactions on Pattern Analysis and Machine Intelligence, 39 (2017) 1853–1865.

[9] M. Cuturi, Sinkhorn distances: Lightspeed computation of optimal transport, In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS13, pages 2292–2300, 2013.

[10] M. Cuturi, G. Peyré, Semidual regularized optimal transport, SIAM Review, 60 (2018) 941-965.

[11] B. Dai, S. Ding, G. Wahba, Multivariate Bernoulli distribution, Bernoulli, 19 (2013) 1465-1483.

[12] A. Dessein, N. Papadakis, J.-L. Rouas, Regularized optimal transport and the rot mover's distance, J. Mach. Learn. Res. 19 (2018), 1-53.

[13] M. Essid, M. Pavon, Traversing the Schrödinger bridge strait: Robert Fortet's marvelous proof redux, Journal of Optimization Theory and Applications, 181 (2019) 23-60.

[14] S. Ferradans, N. Papadakis, G. Peyré, J.-F. Aujol, Regularized discrete optimal transport, 2013. URL https://doi.org/10.48550/arXiv.1307.5551.

[15] R. Flamary, N. Courty, A. Rakotomamonjy, D. Tuia, Optimal transport with laplacian regularization. 2014. URL https://dumas.ccsd.cnrs.fr/OCA/hal-01103076v1.

[16] M. Kárný, Axiomatisation of fully probabilistic design revisited, Systems and Control Letters, 141 (2020) 104719.

[17] M. Kárný, T. Kroupa, Axiomatisation of fully probabilistic design, Information Sciences, 186 (2012) 105-113.

[18] J. Korman and R. J. McCann, Optimal transportation with capacity constraints, 2012. URL https://doi.org/10.48550/arXiv.1201.6404.

[19] S. Kullback and R. A. Leibler, On information and sufficiency, The Annals of Mathematical Statistics, 22 (1951) 79-86.

[20] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, From word embeddings to document distances, In: F. Bach, D. Blei (ed.), Proceedings of the 32nd International Conference on Machine Learning, volume 37 of Proceedings of Machine Learning Research, pages 957–966, 2015. PMLR. URL https://proceedings.mlr.press/v37/kusnerb15.html.

[21] T. Nasiri, M. Moeini-Aghtaie, M. Foroughi, and M. Azimi, Energy optimization of multi-carrier energy systems to achieve a low carbon community, Journal of Cleaner Production, 390 (2023) 136154.

[22] G. Peyré and M. Cuturi, Computational optimal transport, 2018. URL https://arxiv.org/abs/1803.00567.

[23] A. Quinn, M. Kárný, and T. V. Guy, Fully probabilistic design of hierarchical Bayesian models, Information Sciences, 369 (2016) 532–547.

[24] Y. Rubner, C. Tomasi, and L. J. Guibas, The earth mover's distance as a metric for image retrieval, International Journal of Computer Vision, 40 (2024) 99-121.

[25] S. Shafieezadeh-Abadeh, V. Nguyen, D. Kuhn, and P. Esfahani, Wasserstein distributionally robust Kalman filtering, Advances in Neural Information Processing Systems, 31 (2018) 8483–8492.
[26] J. Shore and R. Johnson, Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy, IEEE Transactions on Information Theory, 26 (1980) 26–37.
[27] A. Sklar, Random variables, joint distribution functions, and copulas, Kybernetika, 44 (1973) 449-460.
[28] J. Solomon, F. Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas, Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains, ACM Transactions on Graphics, 34 (2015), 66.
[29] J. L. Teugels, Some representations of the multivariate Bernoulli and binomial distributions, Journal of Multivariate Analysis, 32 (1990) 256-268.
[30] C. Villani, Optimal transport: Old and new, 2008. URL https://cedricvillani.org/sites/dev/files/old_images/2012/08/preprint-1.pdf.
[31] R. Zhang, Z. Wen, C. Chen, C. Fang, T. Yu, and L. Carin, Scalable Thompson sampling via optimal transport, volume 89 of Proceedings of Machine Learning Research, pages 87–96. PMLR, 2019. URL https://proceedings.mlr.press/v89/zhang19a.html.