Bayesian Inference for Optimal Transport with Stochastic Cost

Anton Mallasto[*] ANTON.MALLASTO@AALTO.FI
Department of Computer Science, Aalto University, Finland

Markus Heinonen MARKUS.O.HEINONEN@AALTO.FI
Department of Computer Science, Aalto University, Finland

Samuel Kaski SAMUEL.KASKI@AALTO.FI
Department of Computer Science, Aalto University, Finland
Department of Computer Science, University of Manchester, UK

Editors: Vineeth N Balasubramanian and Ivor Tsang

Abstract

In machine learning and computer vision, optimal transport has had significant success in learning generative models and defining metric distances between structured and stochastic data objects, that can be cast as probability measures. The key element of optimal transport is the so called lifting of an *exact* cost (distance) function, defined on the sample space, to a cost (distance) between probability measures over the sample space. However, in many real life applications the cost is *stochastic*: e.g., the unpredictable traffic flow affects the cost of transportation between a factory and an outlet. To take this stochasticity into account, we introduce a Bayesian framework for inferring the optimal transport plan distribution induced by the stochastic cost, allowing for a principled way to include prior information and to model the induced stochasticity on the transport plans. Additionally, we tailor an HMC method to sample from the resulting transport plan posterior distribution.

Keywords: Optimal Transport, Bayesian Inference, Uncertainty Quantification

1. Introduction

Optimal transport (OT) (Villani, 2008; Peyré et al., 2019) is an increasingly popular tool in machine learning and computer vision, where it is used to define similarities between probability distributions: given a *cost function* between samples (e.g. the Euclidean distance), representing the cost of transporting one sample to another, OT extends it to a cost of transporting an entire distribution to another. This *lifting* of the cost function to the space of probability measures is carried out by finding the *OT plan*, which carries out the transport with minimal total cost.

OT assumes a deterministic and exact cost between samples. This is natural for most of OT applications in machine learning, such as defining loss functions for learning probability distributions, e.g., in Wasserstein generative adversial networks (WGANs) (Arjovsky et al., 2017), or defining statistics for stochastic data objects, e.g., between Gaussian processes representing random curves (Mallasto and Feragen, 2017).

[*] Currently at Silo AI.

© 2021 A. Mallasto, M. Heinonen & S. Kaski.

However, the assumption of an exact cost rarely holds in real-life OT applications. This work is motivated by the lack of tools to solve OT problems in such stochastic settings, where the transportation cost $c(x, y)$ between x and y is a random variable. See Fig. 1 for an illustration. We consider two such situations: **1.** the very definition of c is random, e.g., real life logistics, where transport between two points always results in a different cost due to varying traffic conditions. **2.** OT between hierarchical measures, i.e., we have a mass distribution over a collection of random variables X_i, Y_j whose values are uncertain and stochastic, and so $c(X_i, Y_j)$ is also stochastic.

As the transportation cost varies, a natural question arises: how to take this uncertainty into account in the transportation plan, and which of them should be used in practice? Furthermore, it is important to include any prior knowledge in the solution. To answer these questions, we propose to use the Bayesian paradigm in order to infer the distribution of transport plans induced by the stochastic cost, and name the resulting approach as BayesOT. As a special case, we show that the resulting point estimates for the OT plan correspond to well-known regularizations of OT.

We contribute:

1. BayesOT, A Bayesian formulation of the OT problem, which produces full posterior distributions for the OT plans, and allows solving OT problems having stochastic cost functions.

2. The resulting formalism, relying on introducing optimality variables to relate a transport plan to a given cost matrix, generalizes earlier regularisation approaches, as these can be interpreted as maximum a posteriori estimates in our framework.

3. A Hamiltonian Monte Carlo approach for sampling from the transport polytope, i.e., the set of joint distributions with two fixed marginals.

Related Work. We are not aware of earlier works on stochastic costs in OT, but some works are related. For example, *Schrödinger bridges* consider the most likely path of evolution for a gas cloud given an initial and an evolved state, a problem equivalent to entropy-relaxed OT (Di Marino and Gerolin, 2019). The evolution is Brownian, thus the dynamics bring forth a stochastic cost; however, no stochasticity remains as the most likely evolution is considered.

Ecological inference (King et al., 2004) studies individual behavior through aggregate data, by inferring a joint table from two marginal distributions: this is precisely what is done in OT, using the cost function. Frogner and Poggio (2019) consider a prior distribution over the joint tables, computing the maximum likelihood point estimate. Our work, in addition to the prior distribution, adds a likelihood, relating the joint table to the OT cost matrix. Furthermore, instead of just focusing on the MAP estimates, we also study sampling from the posterior. Rosen et al. (2001) consider Markov Chain Monte Carlo (MCMC) sampling from a user-defined prior distribution to estimate the joint table. However, strict marginal constraints are not enforced, which Frogner and Poggio (2019) speculate is due to the difficulty of MCMC inference on the set of joint distributions with perfectly-observed marginals. In contrast, BayesOT takes the marginal constraints strictly into account.

A conceivable alternative to solving the OT problem with stochastic cost would be standard OT with the average cost. An obvious down-side of this approach would be losing all

	Cost			
	exact	stochastic	Prior	Uncertainty
OT	✓	✗	✗	✗
RegularizedOT	✓	✗	✓	✗
BayesOT	✓	✓	✓	✓

Table 1: Comparison between vanilla OT, Regularized OT, and our method BayesOT. The qualities imply whether the approaches are able to incorporate an exact or stochastic cost, prior information, or whether the methods provide uncertainty estimates.

stochasticity, resulting in an average-case analysis. If the measures are hierarchical, i.e., we have mass distributions μ_i, ν_j over spatially varying components given by random variables X_i, Y_j. Then, the cost $c(X_i, Y_j)$ would be stochastic, depending on the realisations of the components. One could then consider extending the sample-wise cost to a component-wise cost using the OT quantity between the two components, i.e., $\tilde{c}(X_i, Y_j) = \mathrm{OT}_c(X_i, Y_j)$ (Chen et al., 2018). However, we would lose all stochasticity again, and the component-wise OT cost would be blind to any natural correlation between the components.

Furthermore, one could solve the OT plan associated with each cost matrix sample C^k, and carry out population analysis. This would, however, prevent the use of prior information, and would not result in a likelihood on the OT plans which could be used to estimate the relevancy of a given plan.

Minibatch OT (Fatras et al., 2019, 2021) is somewhat related. It has been applied for example in generative modelling (Mallasto et al., 2019), as it provides a convenient way to approximate the OT problem (Genevay et al., 2016). Although minibatch OT does include stochasticity through sampling the minibatches, only the expected OT problem over these minibatches is considered.

2. BACKGROUND

We now summarize the basics of optimal transport (Villani, 2008; Peyré et al., 2019) and Bayesian inference (Gelman et al., 2013) in order to fix notation.

Optimal Transport (OT) is motivated by a simple problem. Assume we have locations of factories $\{x_i\}_{i=1}^n \subset \mathbb{R}^d$ and of outlets $\{y_j\}_{j=1}^m \subset \mathbb{R}^d$. Each of the factories produces μ_i amount of goods, and the outlets have a demand of ν_j, each positive and normalized to sum to one; $\mu_i, \nu_j \geq 0$ and $\sum_{i=1}^n \mu_i = \sum_{j=1}^m \nu_j = 1$. We represent the distribution of goods over the factories and demands over the outlets by the discrete probability measures

$$\mu(x) = \sum_{i=1}^n \mu_i \delta_{x_i}(x), \qquad \nu(y) = \sum_{j=1}^m \nu_j \delta_{y_j}(y), \tag{1}$$

where $\delta_x(y)$ stands for the Dirac delta function.

Assume that the cost of transporting a unit amount of goods from x_i to y_j is $c(x_i, y_j)$, where $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$ is the *cost function*, inducing the *cost matrix* $C_{ij} = c(x_i, y_j)$. Then,

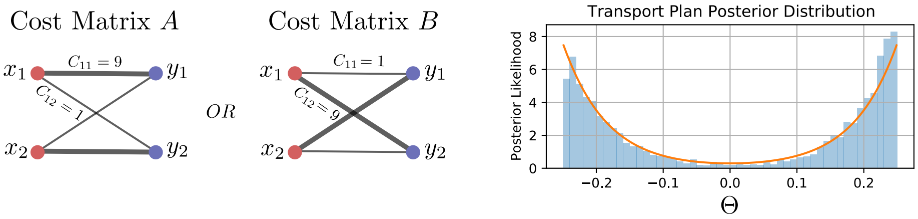

Figure 1: OT with stochastic cost. Assume measures μ, ν having uniform distribution over the atoms x_1, x_2 and y_1, y_2, respectively, and either of the cost matrices A or B is observed with equal probability, so that on average $C_{ij} = 5$ for all i, j. The average cost matrix yields an ill-posed OT problem, as any transport plan would solve the OT problem. On the other hand, the posterior distribution for the transport plan (on the right, blue gives an empirical histogram for posterior samples, orange gives posterior likelihood) encaptures the multimodality, which arises, as there are only two minimizing transport plans for the problem, depending on whether we witness cost matrix A or B. The transport plan is a 2×2 matrix, but can be parameterized with a single real value Θ.

the *OT quantity* between μ and ν is given by

$$\mathrm{OT}(\mu, \nu, C) = \min_{\Gamma \in \Pi(\mu,\nu)} \langle C, \Gamma \rangle \triangleq \min_{\Gamma \in \Pi(\mu,\nu)} \sum_{i,j} \Gamma_{ij} C_{ij}, \tag{2}$$

where the Frobenius inner product $\langle C, \Gamma \rangle$ gives the total transportation cost and the set of joint probability measures with marginals μ and ν, the *transport polytope*, is denoted by

$$\Pi(\mu, \nu) \triangleq \left\{ \Gamma : \sum_{j=1}^{m} \Gamma_{ij} = \mu_i, \quad \sum_{i=1}^{n} \Gamma_{ij} = \nu_j \right\}. \tag{3}$$

Its elements are *transport plans*, as Γ_{ij} is the amount of mass transported from x_i to y_j. The constraints on $\Pi(\mu, \nu)$ enforce the preservation of mass; all the goods produced need to be transported so that the demand of each outlet is satisfied.

This seemingly practical problem produces a geometrical framework for probability measures, by lifting the sample-wise cost function c to a similarity measure $\mathrm{OT}(\mu, \nu, C)$ between the probability measures. Depending on the cost, a metric distance could be produced (i.e., the p-Wasserstein distances), which allows studying probabilities using metric geometry.

Regularized Optimal Transport. The OT problem in (2) is a convex linear program, often producing slow-to-compute, 'sparse' transport plans that might not be unique. This has motivated regularized versions of OT (Cuturi, 2013; Dessein et al., 2018; Di Marino and Gerolin, 2020), which admit unique solutions. We now summarize regularized OT, as it turns out that solving certain *maximum a posteriori* estimates under the BayesOT framework is equivalent to solving regularized OT, as will be discussed in Sec. 3.4.

Given a strictly convex regularizer R, the regularized OT problem is given by (Dessein et al., 2018)

$$\mathrm{OT}_R(\mu, \nu, C) = \min_{\Gamma \in \Pi(\mu,\nu)} \{\langle C, \Gamma \rangle + R(\Gamma)\}, \tag{4}$$

With some technical assumptions on R, such as strict convexity over its domain, there exists a unique minimizer of (4), which in practice can be solved using *iterative Bregman projections*. We denote this minimizer by $\Gamma(R, C, \mu, \nu)$. A popular choice for the regularizer is given by $R = \epsilon H$ (Cuturi, 2013), where the *regularization magnitude* $\epsilon > 0$ is a positive constant, and

$$H(\Gamma) = -\sum_{ij} \Gamma_{ij} \log \Gamma_{ij}, \tag{5}$$

is the *entropy*. This specific regularization strategy has gained much attention, as it is fast to solve with the *Sinkhorn-Knopp iterations* (Knight, 2008), and enjoys better statistical properties compared to vanilla OT (Genevay et al., 2019).

Bayesian Inference. Assume we are given a family of models f_θ, with parameter θ, and a dataset $D = \{(x_i, y_i)\}_{i=1}^n \subset X \times Y$, produced by an underlying relationship $y_i = f(x_i) + \varepsilon_i$, where ε_i is a random noise variable, and we want to infer $f : X \to Y$. Given knowledge about θ in the form of a *prior distribution* $\theta \sim \Pr(\theta)$, Bayesian statistics approaches inferring f by conditioning the parameters via the *Bayes' formula*

$$\Pr(\theta|D) = \frac{\Pr(D|\theta)\Pr(\theta)}{\Pr(D)}, \tag{6}$$

where $\Pr(\theta|D)$ is the *posterior distribution*, $\Pr(D)$ is the evidence, which can be viewed as a normalizing constant for the posterior distribution, and $\Pr(D|\theta)$ is the *likelihood*, whose form is a part of the modelling choices.

The posterior distribution can then be used to estimate the uncertainty of predictions $y = f_\theta(x)$ by sampling $\theta \sim \Pr(\theta|D)$ and observing the induced distribution of y. This distribution can also be summarized as a *point estimate*. A common point estimate for $f(x)$ is given by the *maximum a posteriori* (MAP) estimate $f_{\theta^*}(x)$, where $\theta^* = \arg\max_\theta \Pr(\theta|D)$. Another popular point estimate is given by the average prediction $\mathbb{E}_{\theta\sim\Pr(\theta|D)} f_\theta(x)$.

3. BAYESIAN INFERENCE FOR OPTIMAL TRANSPORT

We now detail our approach, BayesOT, to solving OT with stochastic cost via Bayesian inference. First, we motivate the stochastic cost in Sec. 3.1, and then formulate the problem from a Bayesian perspective in Sec. 3.2. We then focus on sampling from the resulting posterior distribution of OT plans in Sec. 3.3, by devising a Hamiltonian Monte Carlo approach. Finally, we discuss resulting *maximum a posteriori* estimates and their connections to regularized OT in Sec. 3.4.

3.1. Optimal Transport with Stochastic Cost

Consider the scenario where instead of an exact cost matrix, we observe samples $C^k \sim \Pr(C)$, $k = 1, ..., N$, from a stochastic cost C, which we view as a random variable. This stochasticity propagates to the regularized OT plan Γ via the OT problem

$$\Gamma \sim \arg\min_{\Gamma\in\Pi(\mu,\nu)} \{\langle C, \Gamma\rangle + R(\Gamma)\}, \quad C \sim \Pr(C). \tag{7}$$

In the rest of this work, our goal is to infer the distribution Γ inherits from C.

Stochastic costs naturally occur when considering OT between hierarchical models $(\mu_i, X_i)_{i=1}^n$ and $(\nu_j, Y_j)_{j=1}^m$, where X_i, Y_j are random variables taking values in $\mathbb{R}^d$, resulting in the stochastic cost matrix $C_{ij} \sim c(X_i, Y_j)$. Here one can view (μ_i, X_i) as a mobile factory with mass μ_i, that has a stochastic location according to the random variable X_i.

On the other hand, the cost c itself can inherently be stochastic, e.g., when transporting goods in real life, as traffic congestions behave stochastically, affecting the cost of transporting mass from point i to point j.

The choice to tackle (7) with Bayesian inference provides a convenient way of expressing uncertainty in parameters, allows the inclusion of prior knowledge on Γ, alleviating problems with sample complexity, and provides a principled way of choosing point-estimates as the *maximum a posteriori* (MAP) estimates.

3.2. Bayesian Formulation of OT

To employ Bayesian machinery, we need to define a prior distribution $\Pr(\Gamma \mid \mu, \nu)$ for Γ with marginals μ, ν, and a likelihood function that relates Γ to a given sample C^k of the cost. As we will mention below, priors on the transport polytope have already been discussed in the literature. Our key contribution is introducing the likelihood, quantifying how likely a given transport plan Γ is optimal for a given cost matrix C^k.

The Likelihood for C^k is defined using *auxiliary optimality variables* O_k inspired by maximum entropy reinforcement learning (Levine, 2018): define a binary variable $O_k \in \{0, 1\}$ indicating whether Γ achieves the minimum in $\mathrm{OT}(\mu, \nu, C^k)$, so that $O_k = 1$ if it achieves the minimum, and $O_k = 0$ otherwise. We consider the distribution

$$\Pr(O_k = 1 \mid \mu, \nu, C^k, \Gamma) = \exp\left(-\langle C^k, \Gamma\rangle\right), \tag{8}$$

which can be interpreted as *the likelihood of* Γ *being optimal for* C_k.

The likelihood is motivated by the fact that $\mathrm{OT}(\mu, \nu, C^k) \geq 0$ always holds, and so if the total cost is zero, $\langle C^k, \Gamma\rangle = 0$, then the likelihood of Γ being optimal for C^k (i.e., $O_k = 1$) is 1. On the other hand, as $\langle C^k, \Gamma\rangle$ decreases, the likelihood increases.

Prior for Γ. Any prior whose support covers the transport polytope could be used, such as the well-behaved ones discussed by Frogner and Poggio (2019): component-wise normal, gamma, beta, chi-square, logistic and Weibull distributions. The authors also considered the Dirichlet distribution, which we find to work well in practice in the experimental section. We also consider the entropy prior, defined as

$$\Pr(\Gamma \mid \mu, \nu) \propto \exp(\epsilon H(\Gamma)), \quad \epsilon > 0, \tag{9}$$

which we use to enforce the positivity of the OT plans.

Posterior for Γ. Given the likelihood and the prior, the posterior for a *single cost matrix* can now be written as

$$\Pr(\Gamma \mid \mu, \nu, O_k = 1, C^k) \propto \Pr(O_k = 1 \mid \mu, \nu, C^k, \Gamma)\Pr(\Gamma \mid \mu, \nu). \tag{10}$$

On the other hand, a *population of cost matrices* C^k, $k = 1, ..., N$, requires defining a joint likelihood for O_k before we can write out the posterior. To this end, we consider two likelihoods:

(L1) The first likelihood encourages Γ to be optimal for **each** C^k, and is given by

$$\begin{aligned}\Pr{}^{(L1)}(O=1|\mu,\nu,C,\Gamma) &= \Pr{}^{(L1)}(O_1=1,...,O_N=1 \mid \mu,\nu,C^1,...,C^N,\Gamma)\\ &= \prod_{k=1}^{N} \Pr(O_k=1 \mid \mu,\nu,C^k,\Gamma).\end{aligned} \tag{11}$$

(L2) The second likelihood encourages Γ to be optimal **for some** C^k, and is given by

$$\begin{aligned}\Pr{}^{(L2)}(O=1|\mu,\nu,C,\Gamma) &= \Pr{}^{(L2)}(O_1=1,...,O_N=1 \mid \mu,\nu,C^1,...,C^N,\Gamma)\\ &= \sum_{k=1}^{N} \Pr(O_k=1 \mid \mu,\nu,C^k,\Gamma).\end{aligned} \tag{12}$$

The resulting posteriors will be denoted by $\Pr^{(L1)}_{\mu,\nu}(\Gamma|C,O=1)$ and $\Pr^{(L2)}_{\mu,\nu}(\Gamma|C,O=1)$, respectively. The negative posterior log-likelihood can then be written as

$$\begin{aligned}Q^{(L1)}(\Gamma) &= -\log \Pr{}^{(L1)}_{\mu,\nu}(\Gamma \mid C,O=1)\\ &= -\log\left(\Pr(\Gamma\mid\mu,\nu)\prod_{k=1}^{N}\Pr(O_k=1\mid\mu,\nu,C^k,\Gamma)\right)\\ &= -\log\Pr(\Gamma\mid\mu,\nu) + \left\langle \sum_{k=1}^{N} C^k,\Gamma\right\rangle + \text{const.},\end{aligned} \tag{13}$$

and for the likelihood (L2) we compute

$$\begin{aligned}Q^{(L2)}(\Gamma) &= -\log \Pr{}^{(L2)}_{\mu,\nu}(\Gamma|C,O=1)\\ &= -\log\left(\Pr(\Gamma\mid\mu,\nu)\sum_{k=1}^{N}\Pr(O_k=1\mid\mu,\nu,C^k,\Gamma)\right)\\ &= -\log\Pr(\Gamma\mid\mu,\nu) - \log\sum_{k=1}^{N}\exp(-\langle C^k,\Gamma\rangle) + \text{const.}\end{aligned} \tag{14}$$

Note that (14) can be viewed as a smooth version of (7). This can be shown as follows. First, let $R(\Gamma) = -\log\Pr(\Gamma)$, and define a family of distributions with parameter τ:

$$\Pr{}_\tau(\Gamma|C^k,O_k=1) \propto \left(\exp\left(-\langle\Gamma,C^k\rangle\right)\Pr(\Gamma)\right)^{\frac{1}{\tau}} = \exp\left(-\frac{1}{\tau}(\langle\Gamma,C^k\rangle - \log\Pr(\Gamma))\right). \tag{15}$$

Then, assuming $R(\Gamma)$ is strongly convex, so that $\log\Pr_\tau(\Gamma|C^k,O_k=1)$ has a unique maximizer, we have the weak convergence $\Pr_\tau(\Gamma|C^k,O_k=1) \to \mathbb{1}_{\Gamma(R,C,\mu,\nu)}(\Gamma)$ as $\tau\to 0$ (Henderson et al., 2003). Then, note that (14) can be written with $\tau=1$ as

$$\Pr{}^{(L2)}_{\mu,\nu}(\Gamma|C,O=1) = \sum_{k=1}^{N}\Pr{}_1(\Gamma|C^k,O_k=1), \tag{16}$$

whereas (7) can be written as

$$\lim_{\tau \to 0} \int \Pr_\tau(\Gamma | C, O = 1) d\Pr(C). \tag{17}$$

We will discuss the relationship of (13) to regularized OT later in Section 3.4.

3.3. Posterior Sampling

We consider a Markov chain Monte Carlo (MCMC), specifically a Hamiltonian Monte Carlo (HMC) method to sample from the OT plan posteriors. This requires a novel way to take the marginal constraints into account, which we do by devising a chart for the transport polytope in (24).

MCMC methods are the main workhorse in Bayesian inference, allowing sampling from a given unnormalized distribution. First, a proposal process $\Pr(\Gamma_{t+1}|\Gamma_t)$ is devised, where t is a sequential sample index. Given a proposed transition $\Gamma_t \to \Gamma_{t+1}$, we filter it through the *Metropolis-Hastings sampler*, ensuring that the resulting Markov chain is reversible with respect to $\Pr_{\mu,\nu}(\Gamma|C, O = 1)$ and satisfies *detailed balance.*

HMC is a popular variant of MCMC, allowing for efficient sampling in high dimensions, which pairs the state Γ with a *momentum* $P \in \mathbb{R}^{n\times m}$ (Neal et al., 2011). One then defines the *kinetic energy* T and *potential energy* U,

$$\begin{aligned} T(\Gamma, P) &= \mathrm{vec}\,(P)\,\mathrm{diag}(M)^{-1}\mathrm{vec}\,(P) = \sum_{ij} P_{ij}^2 M_{ij}^{-1}, \\ U(\Gamma) &= -\log\det\left(\mathrm{diag}(M)\right) - \log\Pr(\Gamma) \\ &= -\frac{1}{2}\sum_{ij}\log M_{ij} - \Pr_{\mu,\nu}(\Gamma|C, O = 1). \end{aligned} \tag{18}$$

Here $\mathrm{diag}(M) \in \mathbb{R}^{nm\times nm}$ is a *diagonal mass matrix* induced by the matrix $M \in \mathbb{R}_+^{n\times m}$. A default choice would be $M_{ij} = 1$. The kinetic and potential energies form the *Hamiltonian*

$$\mathcal{H}(\Gamma, P) = T(\Gamma, P) + U(\Gamma), \tag{19}$$

which induces the *Hamiltonian system* whose trajectories preserve the Hamiltonian. The HMC procedure then samples a momentum P_t, and evolves the pair (Γ_t, P_t) according to the Hamiltonian with a symplectic integrator, e.g., the *leapfrog algorithm* Betancourt (2017). The resulting pair (Γ_{t+1}, P_{t+1}) is then accepted with probability

$$\alpha(\Gamma_t, \Gamma_{t+1}) = \min\left\{1, \exp\left(\mathcal{H}(\Gamma_t, P_t) - \mathcal{H}(\Gamma_{t+1}, P_{t+1})\right)\right\}. \tag{20}$$

Constraints on Γ, can be taken into account by utilizing the geometry of the transport polytope, which we present below. The positivity constraints $\Gamma \geq 0$ can be enforced coordinate-wise through the prior, assigning zero mass for negative values. For example, one could consider the uniform distribution over the probability simplex.

Transport Polytope. To accommodate the somewhat complicated constraints on the transport polytope, we cast the polytope as a set concentrated on an affine plane bound by positivity constraints. This allows parameterizing the polytope using a linear chart (given in (24) below), which we put in use when sampling viable transport plans.

Rigorously, we formulate the constraints on $\Pi(\mu,\nu)$ in a linear fashion as

$$\begin{bmatrix}\Gamma & 0\\ 0 & \Gamma^T\end{bmatrix}\begin{bmatrix}\mathbb{1}_m\\ \mathbb{1}_n\end{bmatrix}=\begin{bmatrix}\mu\\ \nu\end{bmatrix}, \quad \Gamma_{ij}\geq 0,\ \forall i,j. \tag{21}$$

where $\mathbb{1}_n$ is the n-vector with all coordinates 1. Hence, $\Pi(\mu,\nu)$ is a convex polytope, and furthermore, it lies on the affine plane (see Fig. 2)

$$\Gamma_0+\mathbb{V}_0=\left\{\Gamma_0+M:\sum_j M_{ij}=0,\ \sum_i M_{ij}=0,\ \forall i,j\right\}, \tag{22}$$

for some $\Gamma_0\in\Pi(\mu,\nu)$. Thus, given any $\Gamma\in\Pi(\mu,\nu)$, we can find $M\in\mathbb{V}_0$, so that $\Gamma=\Gamma_0+M$. The vector space $\mathbb{V}_0$ is isomorphic to $\mathbb{R}^{(n-1)\times(m-1)}$ via

$$\varphi:\mathbb{R}^{(n-1)\times(m-1)}\to\mathbb{V}_0,\quad \Theta\mapsto\begin{bmatrix}\Theta & -\Theta^R\\ -(\Theta^C)^T & \sum_{ij}\Theta_{ij}\end{bmatrix}, \tag{23}$$

where $\Theta^R_i=\sum_j\Theta_{ij}$ is the row sum vector of Θ and $\Theta^C_j=\sum_i\Theta_{ij}$ is the respective column sum vector. Thus, φ provides a linear chart for $\Pi(\mu,\nu)$ through

$$\Gamma(\Theta)=\Gamma_0+\varphi(\Theta)\geq 0, \tag{24}$$

where the inequality is enforced for all coordinates.

In practice, we choose Γ_0 to be the independent joint distribution of μ and ν.

3.4. Maximum A Posteriori Estimation as Regularized OT

We now consider the MAP estimate for the posterior distribution $\Pr^{(L1)}_{\mu,\nu}(\Gamma|C,O=1)$ under the likelihood (L1). The (L2) case is more demanding due to the non-convexity of the smooth minimum appearing in $Q^{(L2)}$, whereas $Q^{(L1)}$ is convex if $-\log\Pr(\Gamma\mid\mu,\nu)$ is convex. Now considering $Q^{(L1)}$ in (13), we see that computing the MAP estimate,

$$\Gamma^*_{\forall}=\underset{\Gamma\in\Pi(\mu,\nu)}{\arg\min}\,Q^{(L1)}(\Gamma)=\underset{\Gamma\in\Pi(\mu,\nu)}{\arg\min}\left\{-\log\Pr(\Gamma\mid\mu,\nu)+\left\langle\sum_{k=1}^N C^k,\Gamma\right\rangle\right\}, \tag{25}$$

is equivalent to solving the regularized OT problem (4) with the regularizer

$$R(\Gamma)=-\log\Pr(\Gamma\mid\mu,\nu), \tag{26}$$

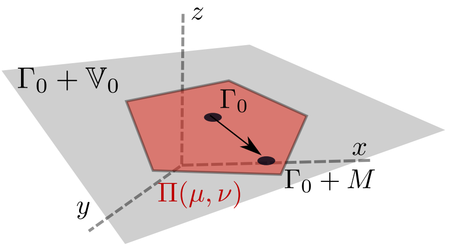

Figure 2: Illustration of the transport polytope as described in (22).

the marginals μ, ν, and the cost matrix $\sum_k C^k$.

For the sake of illustration, we discuss the MAP estimate in three example cases.

Constant Prior. With a constant prior $\Pr(\Gamma) = \text{const.}$, solving (25) corresponds to the vanilla OT problem (2).

Entropy Prior. Assume we have a prior proportional to the exponential of the ϵ-scaled entropy of Γ defined in (9), we get the regularizer

$$R(\Gamma) = -\log \Pr(\Gamma \mid \mu, \nu) = -\epsilon H(\Gamma). \tag{27}$$

Thus, solving (25) corresponds to solving the entropy-relaxed OT problem (Cuturi, 2013).

Gaussian Prior. Consider a Gaussian prior $\text{vec}(\Gamma) \sim \Pr(\bar{\Gamma}, \Sigma)$ for the vectorized transport plan, with mean $\bar{\Gamma}$ and covariance matrix Σ. Then, one gets

$$R(\Gamma) = \frac{1}{2}(\text{vec}(\Gamma) - \bar{\Gamma})\Sigma^{-1}(\text{vec}(\Gamma) - \bar{\Gamma}), \tag{28}$$

and so if $\bar{\Gamma} = 0$, the Gaussian prior results in quadratically regularized OT (Lorenz et al., 2019; Dessein et al., 2018), where the quadratic term is the norm with respect to the *Mahalanobis metric* given by $\frac{1}{2}\Sigma^{-1}$.

	No Cost				**With Cost**			
Prior	Error	Correlation	1 STD	2 STD	Error	Correlation	1 STD	2 STD
Dirichlet	$\mathbf{1.92 \times 10^{-3}}$	**0.702**	**62.8%**	**82.9%**	1.91×10^{-3}	0.686	61.6%	82.4%
Tsallis	2.62×10^{-3}	0.304	47.9%	62.8%	2.54×10^{-3}	0.350	47.1%	60.8%
Entropic	2.44×10^{-3}	0.319	45.4%	60.4%	2.58×10^{-3}	0.240	47.4%	63.0%
Gaussian	2.41×10^{-3}	0.340	44.3%	59.1%	2.42×10^{-3}	0.322	46.5%	60.7%
Uniform	2.52×10^{-3}	0.284	44.3%	58.3%	2.46×10^{-3}	0.290	48.2%	61.9%

Table 2: BayesOT yields meaningful uncertainty estimates for the Florida vote registration dataset. The *median* error is computed for the mean posterior prediction, the correlation is between standard deviations of the posterior (for an entry in the joint distribution) and the absolute error, and the two last columns give the percentage of data points lying inside the confidence bounds given by 1 and 2 standard deviations, respectively. The first four columns omit the OT likelihood term, whereas the four last columns include it.

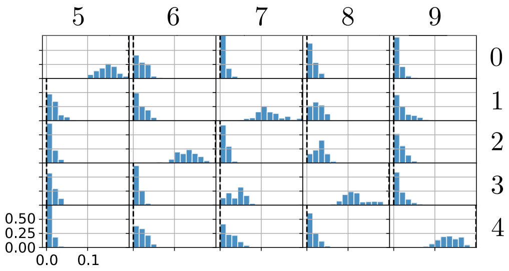

Figure 3: Demonstration of BayesOT between instances of digits 5-9 (columns) and 0-4 (rows). Each histogram shows the posterior of Γ_{ij}.

4. EXPERIMENTS

We now demonstrate BayesOT on one toy data set (MNIST) and give empirical results on two sets: Florida vote registration dataset shows how BayesOT provides useful uncertainty estimates while building on top of traditional OT approaches. The New York City taxi dataset presents real traffic data, which we use to transport persons around Manhattan, comparing the BayesOT posterior to the average case analysis.

We implement BayesOT with the Pyro probabilistic programming framework (Bingham et al., 2019), and use the NUTS sampler (Hoffman and Gelman, 2014) for HMC to automatically tune the hyperparameters. We typically get 60 samples per second on a Macbook Pro 2015 in each of the experiments.

MNIST. As a toy-example with real data, we consider transport between two measures over 32×32 images of hand-written digits in the MNIST dataset (LeCun et al., 1998). The digits $0-9$ are arbitrarily split into two groups of $0-4$ and $5-9$, forming two measures μ and ν with uniform weights. We sample images of each digit from the dataset to compute $N = 100$ samples from the stochastic cost matrix using the squared Euclidean metric. We sample 10^4 points from the posterior with 10^3 burn-in samples with a step size of 10^{-4}, and use the entropy prior with $\epsilon = 1$.

The resulting posterior over the transport plans, with the likelihood (L1), is illustrated in Fig. 3. The results positively match intuition, as we most often see the mappings $0 \mapsto 5$, $1 \mapsto 7$, $2 \mapsto 6$, $3 \mapsto 8$ and $4 \mapsto 9$. However, some of the assignments are not as clear-cut as others. For instance, $0 \mapsto 5$ is very dominant, whereas $3 \mapsto 8$ is not that dominant, as in some cases $3 \mapsto 7$ might be more favorable, depending on the drawing style of the digit. This effect is ignored by the MAP solution

Florida Vote Registration. We apply BayesOT to infer a joint table given two marginals, a common task in ecological inference. On top of point estimates, BayesOT provides uncertainty estimates, which are shown to be meaningful by the experiment.

The Florida dataset (Imai and Khanna, 2016) describes $\approx 10^6$ individual voters in Florida for the 2012 US presidential elections. From the data, we aggregate two marginals per county (of which there are 68), namely a marginal of the party vote ('Democrat', 'Republican', 'Other') and another for ethnicity ('White', 'Black', 'Hispanic', 'Asian', 'Native', 'Other'). Then, we infer a posterior over joint tables between these features (Flaxman et al., 2015), which we compare to ground truth joint tables for each county.

Muzellec et al. (2017) apply OT to this problem by using side information to compute a cost matrix as

$$C_{ij} = \sqrt{2 - 2\exp\left(-\gamma \|v_i^p - v_j^e\|_2\right)}, \tag{29}$$

where $\gamma = 10$, v_i^p is the average profile for party i, consisting of: age normalized to lie within $[0, 1]$, gender represented as a binary number, and an indicator variable expressing whether they voted in 2008 or not. The v_j^e is an analogous profile, but for ethnicity j. Muzellec et al. (2017) employed *Tsallis-regularized OT* to infer the joint table, which in our framework can be viewed as a MAP estimate with Tsallis-entropy prior. We show here how BayesOT, even when the cost is exact, allows us to provide uncertainty estimates for regularized OT, including Tsallis-regularized OT.

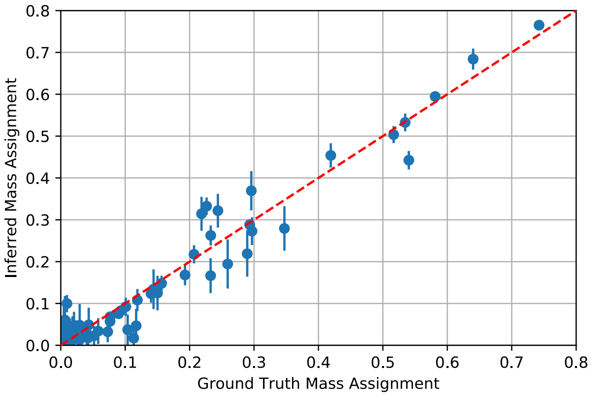

Figure 4: Ground-truth assignments against the posterior mean assignments $\bar{\Gamma}_{ij}$ for the 10 first counties in the Florida vote registration dataset. The posterior utilizes Dirichlet prior with the cost matrix computed over the individual counties. A perfect inference would produce a scatter plot lying on the red diagonal line.

The approach by Frogner and Poggio (2019) discussed in Sec. 1 is also related. They choose a prior distribution, whose most likely joint table is chosen. Our HMC approach, which takes the marginal constraints into account, can be applied to their work, by sampling from the prior distribution, yielding uncertainty estimates for the point estimate.

For each county, we vary the prior distribution between the Diriclet prior, the Tsallis-entropy prior and the entropy prior, and choose whether to use the likelihood associated with the OT cost or not (second term in (14) and (13)). In each case, the HMC chain is initialized with 10^2 burn in samples with an initial step size of 10^{-4}, after which 10^3 posterior samples are acquired. This number of samples is quite low, especially for higher dimensions, but the results show that meaningful uncertainty estimates are still obtained.

The results are summarized in Table 2, presenting the median error, and to assess the uncertainty estimates, the correlation between the uncertainty estimates and absolute error, and how many test values lie within the 1 STD and 2 STD confidence intervals of the point estimate. Furthermore, the results obtained using the Dirichlet prior and the cost matrix on the 10 first counties are illustrated in Fig. 4.

The results indicate clearly that the Dirichlet prior performs the best, as it achieves the lowest median error and highest correlation between the posterior standard deviations and absolute errors. This might be as the prior is supported on the probability simplex, and thus concentrates more mass there compared to the other priors. On the other hand, it is surprising that the cost matrix does not seem to provide meaningful information, as the results over each prior remain quite unaffected when we leave the OT likelihood term out.

NYC Taxi Dataset. We consider data collected from Yellow cabs driving in Manhattan in January 2019, totalling 7.7 million trips. For μ, we consider the 5 most common pick-up zones, and for ν the 6-15 most common pick-up zones, presented in Fig. 5. The weights for μ (and ν) are computed according to the number of trips departing (and arriv-

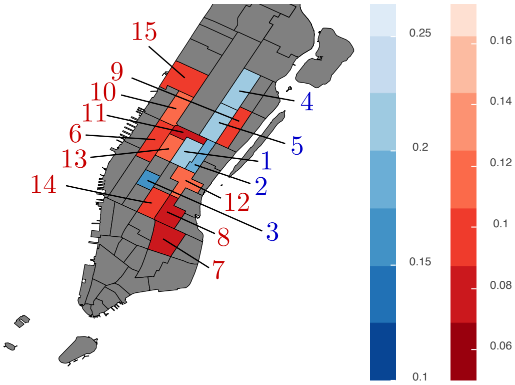

Figure 5: μ (blue) and ν (red) distributions used for taxi zones on Manhattan.

ing) from the location. The cost matrix C_{ij} is computed by sampling trips between locations i and j, and dividing the fare by the amount of passengers on board. Thus, our task is to transport persons from pick-up locations to drop-off locations in an optimal way.

For this experiment, we pick the uniform prior and obtain 1000 samples from the stochastic cost matrix. We initialize the HMC chain with 2×10^3 warm-up iterations, after which we sample 10^4 points from the posterior, induced by the likelihood (L1), which is illustrated in Fig. 6, alongside with the average cost OT solution.

In many cases where the average case analysis assigns considerable mass (e.g., $1 \mapsto 13, 2 \mapsto 12, 5 \mapsto 11$), we see a larger variation in the histogram towards larger mass assignments. This agrees with intuition, as there should be many individual cost matrices encouraging a large assignment, if the average OT plan has a large assignment. However, the histogram also supports low assignments, implying that it is not always optimal to match these taxi zones together. We do also observe contradicting cases, such as $3 \mapsto 14$, which might be caused by a situation, where the assignment on average is optimal, but otherwise is not. On the other end, we also observe cases where no mass is assigned on the average $(1 \mapsto 9, 3 \mapsto 10)$, but the histogram still tends to assign some mass. This could be caused by a similar case as above, where this is suboptimal on average, but in many cases one should still assign some mass.

5. DISCUSSION

We introduced BayesOT, expanding the scope of OT to systems with stochastic cost, a common scenario in the real world. The experiments endorse BayesOT as a successful approach to model the stochasticity propagating to the OT plans from the cost, and proves to be useful in providing uncertainty estimates for use cases of OT with an exact cost.

A notable challenge for the use of BayesOT is formed by the posterior sampling method used (see supplementary material), which is orthogonal to the scope of this work, where the focus was on deriving a general framework for stochastic OT. As we consider marginal distributions with an increasing number of atoms, the dimensionality of the problem increases, subsequently increasing the mixing time for the MCMC method used. However, MCMC methodology is still under extensive consideration, and new advances are likely to

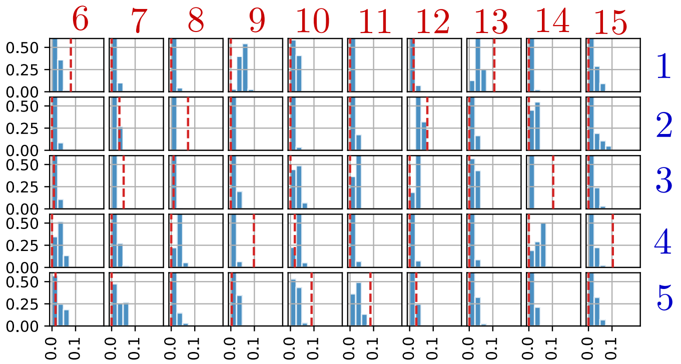

Figure 6: BayesOT posterior with uniform prior for transport plans between zones 1-5 and 6-15. Each histogram shows the posterior of Γ_{ij}, and the red lines give the standard OT solution for the average case.

scale BayesOT to even larger problems. As a current example, an alternative to HMC could be the stochastic gradient Riemann Hamiltonian Monte Carlo (Ma et al., 2015), or more approximative inference methods such as variational inference (Blei et al., 2017).

Future directions for BayesOT include modelling the joint distribution (C, Γ) of the cost and the OT plan explicitly, which allows computing a posterior distribution for the total OT cost. One could also consider regression problems, where at a given time with no observations, a distribution over potential OT plans could be inferred based on previous data. Although advances are needed to scale our approach to large problems, based on the experiments, we view BayesOT as a useful first step towards making OT-based analysis possible in uncertain environments.

Acknowledgments

This work was supported by the Academy of Finland (Flagship programme: Finnish Center for Artificial Intelligence FCAI, Grants 294238, 319264, 292334, 334600, 328400). We acknowledge the computational resources provided by Aalto Science-IT project.

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, 06–11 Aug 2017.

Michael Betancourt. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*, 2017.

Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul A. Szerlip, Paul Horsfall, and Noah D. Goodman.

Pyro: Deep universal probabilistic programming. *J. Mach. Learn. Res.*, 20:28:1–28:6, 2019.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Yongxin Chen, Tryphon T Georgiou, and Allen Tannenbaum. Optimal transport for Gaussian mixture models. *IEEE Access*, 7:6269–6278, 2018.

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

Arnaud Dessein, Nicolas Papadakis, and Jean-Luc Rouas. Regularized optimal transport and the rot mover's distance. *The Journal of Machine Learning Research*, 19(1):590–642, 2018.

Simone Di Marino and Augusto Gerolin. An optimal transport approach for the Schrödinger bridge problem and convergence of Sinkhorn algorithm. *arXiv preprint arXiv:1911.06850*, 2019.

Simone Di Marino and Augusto Gerolin. Optimal transport losses and Sinkhorn algorithm with general convex regularization. *arXiv preprint arXiv:2007.00976*, 2020.

Kilian Fatras, Younes Zine, Rémi Flamary, Rémi Gribonval, and Nicolas Courty. Learning with minibatch Wasserstein: asymptotic and gradient properties. *arXiv preprint arXiv:1910.04091*, 2019.

Kilian Fatras, Younes Zine, Szymon Majewski, Rémi Flamary, Rémi Gribonval, and Nicolas Courty. Minibatch optimal transport distances; analysis and applications. *arXiv preprint arXiv:2101.01792*, 2021.

Seth R Flaxman, Yu-Xiang Wang, and Alexander J Smola. Who supported Obama in 2012? Ecological inference through distribution regression. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 289–298, 2015.

Charlie Frogner and Tomaso Poggio. Fast and flexible inference of joint distributions from their marginals. In *International Conference on Machine Learning*, pages 2002–2011, 2019.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian Data Analysis*. CRC press, 2013.

Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, pages 3440–3448. Curran Associates, Inc., 2016.

Aude Genevay, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of Sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1574–1583. PMLR, 2019.

Darrall Henderson, Sheldon H Jacobson, and Alan W Johnson. The theory and practice of simulated annealing. In *Handbook of Metaheuristics*, pages 287–319. Springer, 2003.

Matthew D Hoffman and Andrew Gelman. The No-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

Kosuke Imai and Kabir Khanna. Improving ecological inference by predicting individual ethnicity from voter registration records. *Political Analysis*, pages 263–272, 2016.

Gary King, Martin A Tanner, and Ori Rosen. *Ecological inference: New methodological strategies*. Cambridge University Press, 2004.

Philip A Knight. The Sinkhorn–Knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018.

Dirk A Lorenz, Paul Manns, and Christian Meyer. Quadratically regularized optimal transport. *Applied Mathematics & Optimization*, pages 1–31, 2019.

Yi-An Ma, Tianqi Chen, and Emily Fox. A complete recipe for stochastic gradient MCMC. In *Advances in Neural Information Processing Systems*, pages 2917–2925, 2015.

Anton Mallasto and Aasa Feragen. Learning from uncertain curves: The 2-Wasserstein metric for Gaussian processes. In *Advances in Neural Information Processing Systems*, pages 5660–5670, 2017.

Anton Mallasto, Jes Frellsen, Wouter Boomsma, and Aasa Feragen. (q, p)-Wasserstein GANs: Comparing ground metrics for Wasserstein Gans. *arXiv preprint arXiv:1902.03642*, 2019.

Boris Muzellec, Richard Nock, Giorgio Patrini, and Frank Nielsen. Tsallis regularized optimal transport and ecological inference. In *AAAI*, 2017.

Radford M Neal et al. MCMC using Hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Ori Rosen, Wenxin Jiang, Gary King, and Martin A Tanner. Bayesian and frequentist inference for ecological inference: The R×C case. *Statistica Neerlandica*, 55(2):134–156, 2001.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.