Tsallis Regularized Optimal Transport and Ecological Inference

Boris Muzellec[†], Richard Nock[‡,¶,§], Giorgio Patrini[¶,‡] and Frank Nielsen[†,♭]
[†]Ecole Polytechnique, France; [‡]Data61, Australia; [¶]The Australian National University, Australia
[§]The University of Sydney, Australia; [♭]Sony CS Labs, Inc., Japan
boris.muzellec@polytechnique.edu; richard.nock@data61.csiro.au; giorgio.patrini@anu.edu.au; Frank.Nielsen@acm.org

Abstract

Optimal transport is a powerful framework for computing distances between probability distributions. We unify the two main approaches to optimal transport, namely Monge-Kantorovitch and Sinkhorn-Cuturi, into what we define as Tsallis regularized optimal transport (TROT). TROT interpolates a rich family of distortions from Wasserstein to Kullback-Leibler, encompassing as well Pearson, Neyman and Hellinger divergences, to name a few. We show that metric properties known for Sinkhorn-Cuturi generalize to TROT, and provide efficient algorithms for finding the optimal transportation plan with formal convergence proofs. We also present the first application of optimal transport to the problem of ecological inference, that is, the reconstruction of joint distributions from their marginals, a problem of large interest in the social sciences. TROT provides a convenient framework for ecological inference by allowing to compute the joint distribution — that is, the optimal transportation plan itself — when side information is available, which is *e.g.* typically what census represents in political science. Experiments on data from the 2012 US presidential elections display the potential of TROT in delivering a faithful reconstruction of the joint distribution of ethnic groups and voter preferences.

1 Introduction

Optimal transport (OT) allows to compare probability distributions by exploiting the underlying metric space on their supports (Kantorovitch 1958; Monge 1781). A number of prominent applications allow for a natural definition of this underlying metric space, from image processing (Rubner, Tomasi, and Guibas 2000) to natural language processing (Kusner et al. 2015) and computer graphics (Solomon et al. 2015).

One key problem of OT is its processing complexity — cubic in the support size, ignoring low order terms (on state of the art LP solvers (Cuturi 2013)). Moreover, the optimal transportation plan has often many zeroes, which is not desirable in some applications. An important workaround was found and consists in penalizing the transport cost with a Shannon entropic regularizer (Cuturi 2013). At the price of changing the transport distance, for a distortion with metric related properties, comes an algorithm with geometric convergence rates (Cuturi 2013; Franklin and Lorenz 1989). As a result, we can picture two separate approches to OT: one essentially relies on the initial Monge-Kantorovitch formulation optimizing the transportation cost itself (Villani 2009), but is computationally expensive; the other is based on tweaking the transportation cost by Shannon regularizer (Cuturi 2013). The corresponding optimization algorithm, grounded in a variety of different works (Csiszár 1989; Sinkhorn 1967; Soules 1991), is fast and can be very efficiently parallelized (Cuturi 2013).

Our paper brings *three* contributions. (i) We interpolate these two worlds using a family of entropies celebrated in nonextensive statistical mechanics, Tsallis entropies (Tsallis 1988), and hence we define the Tsallis regularized optimal transport (TROT). We show that the metric properties for Shannon entropy still hold in this more general case, and prove new properties that are key to our application. (ii) We provide efficient optimization algorithms to compute TROT and the optimal transportation plan. (iii) Last but not least, we provide a new application of TROT to a field in which this optimal transportation plan is the key unknown: the problem of ecological inference.

Ecological inference deals with recovering information from aggregate data. It arises in a diversity of applied fields such as econometrics (Cross and Manski 2002; Cho and Manski 2008), sociology and political science (King 1997; King, Tanner, and Rosen 2004) and epidemiology (Wakefield and Shaddick 2006), with a long history (Robinson 1950); interestingly, the empirical software engineering community has also explored the idea (Posnett, Filkov, and Devanbu 2011). Its iconic application is inferring electorate behaviour: given turnout results for several parties and proportions of some population strata, *e.g.* percentages of ethnic groups, for many geographical regions such as counties, the aim is to recover contingency tables for parties $\times$ groups for all those counties. In the language of probability the problem is isomorphic to the following: given two random variables and their respective marginal distributions — conditioned to another variable, the geography —, compute their conditional joint distribution (See Figure 1).

The problem is fundamentally under-determined and any solution can only either provide loose deterministic bounds (Duncan and Davis 1953; Cross and Manski 2002; Cho and Manski 2008) or needs to enforce additional assumptions

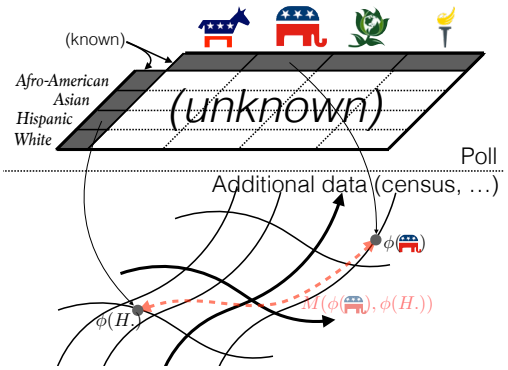

Figure 1: Top: suppose we know (in grey) marginals for the US presidential election (topmost row) and ethnic breakdowns in the US population (leftmost column). Can we recover an estimated joint distribution (white cells) ? If side information is available such as individual level census data (bottom, as depicted on a Hilbert manifold with ϕ-coordinates), then distances can be computed within the supports (dashed red), and optimal transport can provide an estimation of the joint distribution.

and prior knowledge on the data domain (King 1997). More recently, the problem has witnessed a period of renaissance along with the publication of a diversity of methods from the second family, mostly inspired by distributional assumptions as summarised in (King, Tanner, and Rosen 2004). Closer to our approach, (Judge, Miller, and Cho 2004) follows the road of a minimal subset of assumptions and frame the inference as an optimization problem. The method favors one solution according to some information-theoretic solution, *e.g.* the Cressie-Read power divergence, intended as an entropic measure of the joint distribution.

There is an intriguing link between optimal transport and ecological inference: if we can figure out the computation of the ground metric, then the optimal transportation plan provides a solution to the ecological inference problem. This is appealing because it ties the computation of the joint distribution to a ground individual distance between people. Figure 1 gives an example. As recently advocated in ecological inference (Flaxman, Wang, and Smola 2015), it turns out that we have access to more and more side information that helps to solve ecological inference — in our case, the computation of this ground metric. Polls, census, social networks are as many sources of public or private data that can be of help. It is not our objective to show how to best compute the ground metric, but we show an example on real world data for which a simple approach gives very convincing results.

To our knowledge, there is no former application of optimal transport (regularized or not) to ecological inference. The closest works either assume that the joint distribution follows a random distribution constrained to structural or marginal constraints (Forcina and Marchetti 2011) (and references therein) or modify the constraints to the marginals and / or add constraints to the problem (Donoso, Marín, and Vila 2005). In all cases, there is no ground metric (or anything that looks like a cost) among supports that ties the computation of the joint distribution. More importantly, as noted in (Flaxman, Wang, and Smola 2015), traditional ecological inference would not use side information of the kind that would be useful to estimate our ground metric.

This paper is organized as follows. In Section § 2, we present the main definitions for OT. § 3 presents TROT and its geometric properties. § 4 presents the algorithms to compute TROT and the optimal transportation plan, and their properties. § 5 details experiments. A last Section concludes with open problems. *All proofs*, related comments, and some experiments are deferred to a Supplementary Material (Muzellec et al. 2017).

2 Basic definitions and concepts

In the following, we let $\triangle_n \doteq \{\boldsymbol{x} \in \mathbb{R}^n_+ : \boldsymbol{x}^\top \mathbf{1} = 1\}$ denote the probability simplex (bold faces like $\boldsymbol{x}$ denote vectors). $\langle P, Q \rangle \doteq \boldsymbol{vec}(P)^\top \boldsymbol{vec}(Q)$ denotes Frobenius product ($\boldsymbol{vec}(.)$ is the vectorization of a matrix). For any two $\boldsymbol{r}, \boldsymbol{c} \in \triangle_n$, we define their *transportation polytope* $U(\boldsymbol{r}, \boldsymbol{c}) \doteq \{P \in \mathbb{R}^{n\times n}_+ : P\mathbf{1} = \boldsymbol{r}, P^\top \mathbf{1} = \boldsymbol{c}\}$. For any cost matrix $M \in \mathbb{R}^{n\times n}$, the *transportation distance* between $\boldsymbol{r}$ and $\boldsymbol{c}$ as the solution of the following minimization problem:

$$d_M(\boldsymbol{r}, \boldsymbol{c}) \doteq \min_{P \in U(\boldsymbol{r},\boldsymbol{c})} \langle P, M \rangle \ . \tag{1}$$

Its argument, $P^\star \doteq \arg\min_{P\in U(\boldsymbol{r},\boldsymbol{c})} \langle P, M\rangle$ is the *(optimal) transportation plan* between $\boldsymbol{r}$ and $\boldsymbol{c}$. Assuming $M \neq 0$, $P^\star$ is unique. Furthermore, if M is a *metric matrix*, then d_M is also a metric (Villani 2009, §6.1).

In current applications of optimal transport, the key unknown is usually the distance d_M (Cuturi 2013; Cuturi and Doucet 2014; Genevay et al. 2016; Qian et al. 2016; Solomon et al. 2015) (etc). In the context of ecological inference (Judge, Miller, and Cho 2004), it is rather $P^\star$: $P^\star$ describes a joint distribution between two discrete random variables R and C with respective marginals $\boldsymbol{r}$ and $\boldsymbol{c}$, $p^\star_{ij} = \Pr(\mathsf{R} = r_i \wedge \mathsf{C} = c_j)$, for example the support of R being the votes for year Y US presidential election, and C being the ethnic breakdown in the US population in year Y, see Figure 1. In this case, $p^\star_{ij}$ denotes an "ideal" joint distribution of votes within ethnicities, ideal in the sense that it minimizes a distance based on the belief that votes *correlate* positively with a similarity between an ethnic profile and a party's profile. While we will carry out most of our theory on formal transportation grounds, requiring in particular that M be a distance matrix, it should be understood that requiring just "correlation" alleviates the need for M to formally be a distance for ecological inference.

3 Tsallis Regularized Optimal Transport

For any $\boldsymbol{p} \in \mathbb{R}^n_+, q \in \mathbb{R}$, the *Tsallis entropy* of $\boldsymbol{p}$, $H_q(\boldsymbol{p})$ is:

$$H_q(\boldsymbol{p}) \doteq \frac{1}{1-q} \cdot \sum_i (p_i^q - p_i) \ , \tag{2}$$

and for any $P \in \mathbb{R}^{n\times n}_+$, we let $H_q(P) \doteq H_q(\boldsymbol{vec}(P))$. Notably, we have $\lim_{q\to 1} H_q(\boldsymbol{p}) = -\sum_i p_i \ln p_i \doteq H_1(\boldsymbol{p})$, which is just Shannon's entropy. For any $\lambda > 0$, we define the Tsallis Regularized Optimal Transport (TROT) distance.

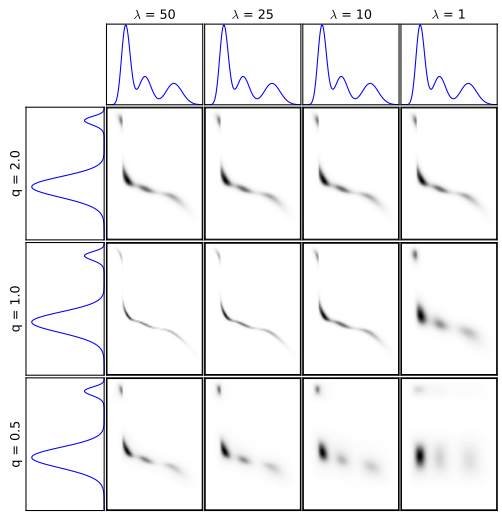

Figure 2: Example of optimal TROT transportation plans (grey levels) for two marginals (blue), with different values of q (in $K_{1/q}$, *Cf* Lemma 2) that corresponds to square Hellinger, Kullback-Leibler and Pearson's χ^2 divergence (top to bottom, conventions follow (Solomon et al. 2015)).

Definition 1 *The* TROT(q, λ, M) *distance (or* TROT *distance for short) between* $\boldsymbol{r}$ *and* $\boldsymbol{c}$ *is:*

$$d_M^{\lambda,q}(\boldsymbol{r},\boldsymbol{c}) \quad \doteq \quad \min_{P \in U(\boldsymbol{r},\boldsymbol{c})} \langle P, M \rangle - \frac{1}{\lambda} \cdot H_q(P) \ . \quad (3)$$

A simple yet important property is that TROT distance unifies both usual modalities of optimal transport. It generalizes optimal transport (OT) when $q \to 0$, since H_q converges to a constant and so the OT-distance is obtained up to a constant additive term (Kantorovitch 1958; Monge 1781). It also generalizes the regularized optimal transport approach of (Cuturi 2013) since $\lim_{q \to 1} d_M^{\lambda,q}(\boldsymbol{r},\boldsymbol{c}) = d_M^{\lambda}(\boldsymbol{r},\boldsymbol{c})$, the Sinkhorn distance between $\boldsymbol{r}$ and $\boldsymbol{c}$ (Cuturi 2013). There are several important structural properties of $d_M^{\lambda,q}$ that motivate the unification of both approaches. To state them, we respectively define the q-logarithm,

$$\log_q(x) \quad \doteq \quad (1-q)^{-1} \cdot (x^{1-q} - 1) \ , \quad (4)$$

the q-exponential, $\exp_q(x) \doteq (1 + (1-q) \cdot x)^{1/(1-q)}$ and Tsallis relative q-entropy between $P, R \in \mathbb{R}_+^{n \times n}$ as:

$$K_q(P,R) \doteq \frac{1}{1-q} \cdot \sum_{i,j} \left(q p_{ij} + (1-q) r_{ij} - p_{ij}^q r_{ij}^{1-q} \right) \quad (5)$$

Taking joint distribution matrices P, R and $q \to 1$ allows to recover the natural logarithm, the exponential and Kullback-Leibler (KL) divergence, respectively (Amari 2016). Other notable examples include (i) Pearson's χ^2 statistic ($q = 2$), (ii) Neyman's statistic ($q = -1$), (iii) square Hellinger distance ($q = 1/2$) and the reverse KL divergence if scaled appropriately by q (Judge, Miller, and Cho 2004), which also allows to span Amari's α divergences for $\alpha = 1 - 2q$ (Amari 2016). For any function $f : \mathbb{R} \to \mathbb{R}$, denoting $f(P)$ for matrix P as the matrix whose general term is $f(p_{ij})$.

Lemma 2 *Let* $\tilde{U} \doteq \exp_q(-1) \exp_q^{-1}(\lambda M)$. *Then:*

$$d_M^{\lambda,q}(\boldsymbol{r},\boldsymbol{c}) \quad = \quad \frac{1}{\lambda} \cdot \min_{P \in U(\boldsymbol{r},\boldsymbol{c})} K_{1/q}(P^q, \tilde{U}^q) + g(M) \quad (6)$$

where $g(M) \doteq (1/\lambda) \cdot \langle \tilde{U}^q, 1 \rangle$ *does not play any role in the minimization of* $K_{1/q}(.\|.)$.

Lemma 2 shows that the TROT distance is a divergence involving *escort* distributions (Amari 2016, § 4), a particularity that disappears in Sinkhorn distances since it becomes an ordinary KL divergence between distributions. Predictably, the generalization is useful to create new solutions to the regularized optimal transport problem that are not captured by Sinkhorn distances (*solution* refers to (optimal) transportation plans, *i.e.* the argument of the min in eq. (3)).

Theorem 3 *Let* $\mathcal{S}_{M,q}(\boldsymbol{r},\boldsymbol{c})$ *denote the set of solutions of eq. (3) when* λ *ranges over* $\mathbb{R}^+$. *Then* $\forall q, q'$ *such that* $q \neq q'$, $\mathcal{S}_{M,q}(\boldsymbol{r},\boldsymbol{c}) \neq \mathcal{S}_{M,q'}(\boldsymbol{r},\boldsymbol{c})$.

Figure 2 provides examples of solutions. Adding the free parameter q is not just interesting for the reason that we bring new solutions to the table: $(1/q) \cdot K_q(\boldsymbol{p},\boldsymbol{r})$ turns out to be Cressie-Read Power Divergence (for $q = \lambda + 1$, (Judge, Miller, and Cho 2004)), and so TROT has an applicability in ecological inference that Sinkhorn distances alone do not have. In addition, we also generalize two key facts already known for Sinkhorn distances (Cuturi 2013). First, the solution to TROT is unique (for $q \neq 0$) and satisfies a simple analytical expression amenable to convenient optimization.

Theorem 4 *There exists exactly one matrix* $P \in U(\boldsymbol{r},\boldsymbol{c})$ *solution to* TROT(q, λ, M). *It satisfies:*

$$p_{ij} \quad = \quad \exp_q(-1) \exp_q^{-1}(\alpha_i + \lambda m_{ij} + \beta_j) \ , \forall i, j \ . (7)$$

($\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^n$ *are unique up to an additive constant).*

Second, we can tweak TROT to meet distance axioms. Let

$$d_{M,\alpha,q}(\boldsymbol{r},\boldsymbol{c}) \quad \doteq \quad \min_{\substack{P \in U(\boldsymbol{r},\boldsymbol{c}) \\ H_q(P) - H_q(\boldsymbol{r}) - H_q(\boldsymbol{c}) \geq \alpha}} \langle P, M \rangle \ , \quad (8)$$

where $\alpha \geq 0$. For any $M, \boldsymbol{r}, \boldsymbol{c}, \lambda \geq 0$, $\exists \alpha \geq 0$ such that $d_{M,\alpha,q}(\boldsymbol{r},\boldsymbol{c}) = d_M^{\lambda,q}(\boldsymbol{r},\boldsymbol{c})$. Also, the following holds.

Theorem 5 *For* $q \geq 1, \alpha \geq 0$ *and if* M *is a metric matrix, function* $(\boldsymbol{r},\boldsymbol{c}) \to \mathbb{1}_{\{\boldsymbol{r} \neq \boldsymbol{c}\}} d_{M,\alpha,q}(\boldsymbol{r},\boldsymbol{c})$ *is a distance.*

Theorem 5 is a generalization of (Cuturi 2013, Theorem 1) (for $q = 1$). As we explain more precisely in the supplement (Muzellec et al. 2017), there is a downside to using $d_{M,\alpha,q}$ as proof of the good properties of $d_M^{\lambda,q}$: the triangle inequality, key to Euclidean geometry, transfers to $d_M^{\lambda,q}$ *with varying and uncontrolled parameters* — in the inequality, the three values of λ may all be different! This does not break down the good properties of $d_M^{\lambda,q}$, it just calls for workarounds. We

now give one, which replaces $d_{M,\alpha,q}$ by the quantity ($\beta \in \mathbb{R}$ is a constant):

$$d_M^{\lambda,q,\beta}(\boldsymbol{r},\boldsymbol{c}) \quad \doteq \quad d_M^{\lambda,q}(\boldsymbol{r},\boldsymbol{c}) + \frac{\beta}{\lambda}\cdot(H_q(\boldsymbol{r})+H_q(\boldsymbol{c})) \quad (9)$$

This has another trivial advantage that $d_{M,\alpha,q}$ does not have: the solutions (optimal transportation plans) are always the *same* on both sides. Also, the right-hand side is lowerbounded for any $\boldsymbol{r},\boldsymbol{c}$ and the trick that ensures the identity of the indiscernibles still works on $d_M^{\lambda,q,\beta}$. The good news is that if $q=1$, $d_M^{\lambda,q,\beta}$, *as is*, can satisfy the triangle inequality.

Theorem 6 *$d_M^{\lambda,1,\beta}$ satisfies the triangle inequality, $\forall\beta \geq 1$.*

Hence, the solutions to $d_M^{\lambda,1}$ *are* optimal transport plans for distortions that meet the triangle inequality. This is new compared to (Cuturi 2013). For a general $q \geq 1$, the proof, in the supplement (Muzellec et al. 2017), shows more, namely that $d_M^{\lambda,q,1/2}$ satisfies a weak form of the identity of the indiscernibles. Finally, there always exist a value $\beta \geq 0$ such that $d_M^{\lambda,q,\beta}$ is non negative ($d_M^{\lambda,q,\beta}$ is lowerbounded $\forall\beta \geq 0$).

4 Efficient TROT optimizers

The key idea behind Sinkhorn-Cuturi's solution is that the KKT conditions ensure that the optimal transportation plan $P^\star$ satisfies $P^\star = \mathrm{diag}(\boldsymbol{u})\exp(-\lambda M)\mathrm{diag}(\boldsymbol{v})$. Sinkhorn's balancing normalization can then directly be used for a fast approximation of $P^\star$ (Sinkhorn 1967; 1964). This trick does not fit at first sight for Tsallis regularization because the q-exponential is *not* multiplicative for general q and KKT conditions do not seem to be as favorable. We give however workarounds for the optimization, that work for *any* $q \in \mathbb{R}_+$.

First, we assume wlog that $q \neq 0,1$ since in those cases, any efficient LP solver ($q=0$) or Sinkhorn balancing normalization ($q=1$) can be used. The task is non trivial because for $q \in (0,1)$, the function minimized in $d_M^{\lambda,q}$ is *not Lipschitz*, which impedes the convergence of gradient methods. In this case, our workaround is Algorithm 1 (SO–TROT), which relies on a Second Order approximation of a fundamental quantity used in its convergence proof, auxiliary functions (Della Pietra, Della Pietra, and Lafferty 1997).

Theorem 7 (Convergence of SO–TROT) *For any fixed $q \in (0,1)$, matrix P output by* SO–TROT *converges to $P^\star$ with:*

$$P^\star \quad = \quad \arg\min_{P\in\mathbb{R}_+^{n\times n}:P\mathbf{1}=\boldsymbol{r}} K_{1/q}(P^q,\tilde{U}^q) \ .$$

The proof (in the supplement (Muzellec et al. 2017)) is involved but interesting in itself because it represents one of the rare uses of the theory of auxiliary functions outside the realm of Bregman divergences in machine learning (Collins, Schapire, and Singer 2002; Della Pietra, Della Pietra, and Lafferty 1997). Some important remarks should be made. First, since SO–TROT uses only one of the two marginal constraints, it would need to be iterated ("*wrapped*"), swapping the row and column constraints like in Sinkhorn balancing.

Algorithm 1 Second Order Row–TROT (SO–TROT)

Input: marginal $\boldsymbol{r}$, matrix M, params $\lambda \in \mathbb{R}_{+*}, q \in (0,1)$

1: $A \leftarrow \lambda M$
2: $P \leftarrow \exp_q(-1)\exp_q^{-1}(A)$
3: **repeat**
4: $\quad P_1 \leftarrow P \oslash A, P_2 \leftarrow P_1 \oslash A$ //$\oslash$ = Kronecker divide
5: $\quad \boldsymbol{d} \leftarrow \boldsymbol{r} - P\mathbf{1}, \boldsymbol{b} \leftarrow P_1\mathbf{1}, \boldsymbol{a} \leftarrow (2-q)P_2\mathbf{1}$
6: $\quad$ **for** $i = 1,2,...,n$
7: $\quad\quad$ **if** $d_i \geq 0$ **then**
8: $\quad\quad\quad y_i \leftarrow \frac{-b_i+\sqrt{b_i^2+4a_id_i}}{2a_i}$
9: $\quad\quad$ **else**
10: $\quad\quad\quad y_i \leftarrow d_i/b_i$
11: $\quad\quad$ **end if**
12: $\quad\quad$ **if** $|y_i| > \frac{q}{(6-4q)\cdot\max_j p_{ij}^{1-q}}$ **then**

$$y_i \quad \leftarrow \quad \frac{q\cdot\mathrm{sign}(r_i-\sum_j p_{ij})}{(6-4q)\cdot\max_j p_{ij}^{1-q}} \ . \quad (10)$$

13: $\quad A \leftarrow A - \boldsymbol{y}\mathbf{1}^\top$
14: $\quad P \leftarrow \exp_q(-1)\exp_q^{-1}(A)$
15: **until** convergence

Output: P

In practice, this is not efficient. Furthermore, iterating SO–TROT over constraint swapping does not necessarily converge. For these reasons, we swap constraints *in* the algorithm, making one iteration of Steps 4-14 over rows, and then one iteration of Steps 4-14 over columns (this boils down to transposing matrices in SO–TROT), and so on. This converges, but still is not the most efficient. To improve efficiency we perform two modifications, that do not impede convergence experimentally. First, we remove Step 12. In doing so, we not only save $O(n^2)$ computations for *each* outer loop, we essentially make SO–TROT as parallelizable as Sinkhorn balancing (Cuturi 2013). Second, we remarked experimentally that convergence is faster when multiplying y_i by 2 in Step 10, and dividing a by 2 in Step 5.

For simplicity, we still refer to this algorithm (balancing constraints in the algorithm, with the modifications for Steps 5, 10, 12) as SO–TROT in the experiments.

Last, when $q \geq 1$, the function minimized in $d_M^{\lambda,q}$ becomes Lipschitz. In this case, we take the particular geometry of $U(r,c)$ into account by using mirror gradient methods, which are equivalent to gradient methods projected according to some suitable divergence (Beck and Teboulle 2003). In our case, we consider Kullback-Leibler divergence, which can save a factor $O(n/\sqrt{\log n})$ iterations (Beck and Teboulle 2003). Furthermore, the Kullback-Leibler projection can be written in terms of Sinkhorn-Knopp's (SK) algorithm with marginals constraints $\boldsymbol{r},\boldsymbol{c}$ (Sinkhorn and Knopp 1967), as is shown in Algorithm 2, named KL–TROT ($\otimes$ is Kronecker product).

Theorem 8 *If $q > 1$ and the gradient steps $\{t_k\}$ are s.t. $\sum_k t_k \to \infty$ and $\sum_k t_k^2 \ll \infty$, matrix P output by* KL–

Algorithm 2 KL Projected Gradient –TROT (KL–TROT)

Input: Marginals $\boldsymbol{r}, \boldsymbol{c}$, Matrix $\tilde{U}$, Gradient steps $\{t_k\}$

1: $P^{(0)} \leftarrow \tilde{U}$
2: **repeat**
3: $\quad P^{(k+1)} \leftarrow \text{SK}(P^{(k)} \otimes \exp(-t_k \nabla f(P^{(k)})), \boldsymbol{r}, \boldsymbol{c})$
4: **until** convergence

TROT *converges to* $P^\star$ *with:*

$$P^\star \quad = \quad \arg \min_{P \in U(\boldsymbol{r},\boldsymbol{c})} K_{1/q}(P^q, \tilde{U}^q) \ .$$

(proof omitted, follows (Beck and Teboulle 2003; Sinkhorn and Knopp 1967))

5 Experiments

We evaluate empirically the TROT framework with its application to ecological inference. The dataset we use describes about 10 millions individual voters from Florida for the 2012 US presidential elections, as obtained from (Imai and Khanna 2016). The data is much richer than is required for ecological inference: surely we could estimate the joint distribution of every voters' available attributes by counting. This is itself a particularly rare case of data quality in political science, where any analysis is often carried out on aggregate measurements. In fact, since ground truth distributions are effectively available, the Florida dataset has been used to test methodological advances in the field (Flaxman, Wang, and Smola 2015; Imai and Khanna 2016). As a demonstrative example, we focus on inferring the distributions of ethnicity and party for all Florida counties.

Dataset description and preprocessing. The data contains the following attributes *for each voter*: location (district, county), gender, age, party (Democrat, Republican, Other), ethnicity (White, African-american, Hispanic, Asian, Native, Other), 2008 vote (yes, no). About 800K voters with missing attributes are excluded from the study. Thanks to the richness of the data, marginal probabilities of ethnic groups and parties can be obtained by counting: for each county we obtain marginals $\boldsymbol{r}, \boldsymbol{c}$ for the optimal transport problems.

Evaluation assumptions. Two assumptions are made in terms of information available for inference. First, the ground truth joint distributions for one district are known; we chose district number 3 which groups 9 out of 68 counties of about $285K$ voters in total. This information will be used to tune hyper-parameters. Second, a cost matrix M^{RBF} is computed based on mean voter's attributes at state level. For the sake of simplicity, we retain only age (normalized in $[0, 1]$), gender and the 2008 vote; notice that in practice geographical attributes may encode relevant information for computing distances between voter behaviours (Flaxman, Wang, and Smola 2015). We do not use this. For distance matrix M^{RBF}, we aggregate those features over all Florida for each party to obtain the vectors $\boldsymbol{\mu}^{\mathsf{p}}$ of the party's expected profile and for each ethnic group to obtain the vectors $\boldsymbol{\mu}^{\mathsf{e}}$ of the ethnicity's expected profile. The dissimilarity measure

party \ ethnicity	white	afro.	hispanic	asian	native	other
Democrat	0.29	0.38	0.55	0.55	0.37	0.57
Republican	0.18	0.63	0.76	0.84	0.54	0.72
Other	0.74	0.62	0.27	0.24	0.41	0.23

Table 1: Visualization of the cost matrix as M: small values indicate high similarity. Highest similarity: (white, Republican); lowest similarity: (asian, Republican) followed by (hispanic, Republican).

Algorithm	M	q	λ	KL-divergence $\pm$ SD	Abs. error $\pm$ SD
Florida-Average	-	-	-	0.251 ± 0.187	0.025 ± 0.011
Simplex	M^{RBF}	-	-	0.280 ± 0.108	0.023 ± 0.008
Simplex	M^{sur}	-	-	0.136 ± 0.098	0.013 ± 0.009
Sinkhorn	M^{RBF}	$1.0^\dagger$	10^0	0.054 ± 0.036	0.009 ± 0.005
Sinkhorn	M^{sur}	$1.0^\dagger$	10^1	0.035 ± 0.027	0.007 ± 0.004
TROT	M^{RBF}	1.0	10^0	0.054 ± 0.036	0.009 ± 0.005
TROT	M^{sur}	2.8	10^1	$\mathbf{0.007 \pm 0.009}$	$\mathbf{0.003 \pm 0.002}$
TROT	M^{no}	0.8	10^0	0.076 ± 0.048	0.011 ± 0.005

Table 2: Average KL-divergence and absolute error with standard deviation (SD) of algorithms inferring joint distributions of all Florida counties. Parameters noted with † are not cross-validated but defined by the algorithm.

relies on a Gaussian kernel between average county profiles:

$$m_{ij}^{\text{RBF}} \quad \doteq \quad \sqrt{2 - 2\exp(-\gamma \cdot \|\boldsymbol{\mu}_i^{\mathsf{p}} - \boldsymbol{\mu}_j^{\mathsf{e}}\|_2)} \ , \quad (11)$$

with $\gamma = 10$. The given function is actually the Hilbert metric in the RBF space. Table 1 shows the resulting cost matrix. Notice how it does encode some common-sense knowledge: White and Republican is the best match, while Hispanic and Asians are the worst match with Republican profiles. It is rather surprising that only 3 features such as age, gender and whether people voted at the last election can reflect so well those relative political traits; these results are indeed much in line with survey-based statistics (Gallup 2013). We also try another cost matrix M, M^{sur}, derived from the ID proportions of parties composition given in (Gallup 2013); m_{ij}^{sur} is computed as $1 - p_{ij}$, where p_{ij} is the proportion of people registered to party j belonging to ethnic group i. Finally, we consider a "no prior" matrix M^{no}, in which $m_{ij}^{\text{no}} = 1, \forall i, j$.

Cross-validation of q. We study the solution of TROT for a grid of $\lambda \in [0.01, 1000], q \in [0.5, 4]$, inferring the joint distributions of all counties of district number 3. We measure average KL-divergence between inferred and ground truth joint distributions. Notice that each county defines a *different* optimal transport problem; inferring the joint distributions for multiple counties at a time is therefore trivial to parallelize. This is somewhat counter-intuitive since we may believe that geographically wider spread data should improve inference at a local level, that is, more data better inference. Indeed, the implicit coupling of the problem is represented by cost matrix, which expresses some prior knowledge of the problem by means of all data from Florida.

Baselines and comparisons with other methods. To evaluate quantitatively the solution of TROT is useful to de-

fine a set of baseline methods: i) Florida-average, which the same state-level joint distribution (assumed prior knowledge) for each of the 67 county; ii) Simplex, that is the solution of optimal transport with no regularization as given by the Simplex algorithm; iii) Sinkhorn(-Cuturi)'s algorithm, which is TROT with $q = 1$; iv) TROT. ii-iv are tested with $M \in \{M^{\text{RBF}}, M^{\text{sur}}\}$, and we provide in addition the results for TROT with $M = M^{\text{no}}$. Hyper-parameters are cross-validated independently for each algorithm.

Table 2 reports a quantitative comparison. From the most general to the most specific, there are three remarks to make. First, optimal transport can be (but is not always) better than the default distribution (Florida average). Second, *regularizing* optimal transport consistently improves upon these baselines. Third, TROT successfully matches Sinkhorn's approach when $q = 1$ is be the best solution in TROT's range of q ($M = M^{\text{RBF}}$), and manages to tune q to significantly beat Sinkhorn's when better alternatives exist: with $M = M^{\text{sur}}$, TROT divides the expected KL divergence by more than *seven* (7) compared to Sinkhorn. This is a strong advocacy to allow for the tuning of q. Notice that in this case, λ is larger compared to $M = M^{\text{RBF}}$, which makes sense since $M = M^{\text{sur}}$ is more accurate for the optimal transport problem (see the Simplex results) and so the weight of the regularizer predictably decreases in the regularized optimal transport distance. We conjecture that $M = M^{\text{sur}}$ beats $M = M^{\text{RBF}}$ in part because it is somehow finer grained: M^{RBF} is computed from sufficient statistics for the marginals alone, while M^{sur} exploits information computed from the cartesian product of the supports. Figure 3 compares all 1 836 inferred probabilities (3×6 per county) with respect to the ground truth for Sinkhorn vs TROT using $M = M^{\text{sur}}$. Remark that the figures in Table 2 translate to per-county ecological inference results that are significantly more in favor of TROT, which basically has no "hard-to-guess" counties compared to Sinkhorn for which the absolute difference between inference and ground truth can exceed 10%.

To finish up, additional experiments, displayed in the supplement (Muzellec et al. 2017) also show that TROT with $M = M^{\text{sur}}$ manages to have a distribution of per county errors extremely peaked around zero error, compared to the simplest baselines (Florida average and TROT with $M = M^{\text{no}}$). These are good news, but there are some local discrepancies. For example, there exists *one* county on which TROT with $M = M^{\text{sur}}$ is beaten by TROT with $M = M^{\text{no}}$.

6 Discussion and conclusion

In this paper, we have bridged Shannon regularized optimal transport and unregularized optimal transport, via Tsallis entropic regularization. There are three main motivations to the generalization, the two first have already been discussed: TROT allows to keep the properties of Sinkhorn distances, and fields like ecological inference bring natural applications for the general TROT family. The application to ecological inference is also interesting because the main unknown is the optimal transportation plan and not necessarily the transportation distance obtained. The third and last motivation is important for applications at large

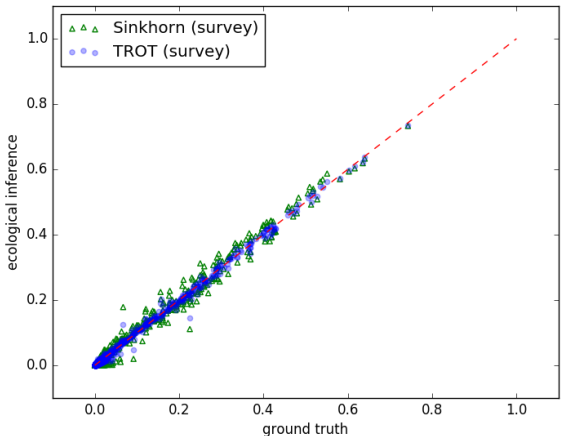

Figure 3: Correlation between TROT vs Sinkhorn inferred probabilities and ground truth for all Florida counties (the closer to $y = x$, the better).

and ecological inference in particular. TROT spans a subset of f-divergences, and f-divergences satisfy the information monotonicity property that coarse graining does not increase the divergence (Amari 2016, § 3.2). Furthermore, f-divergences are invariant under diffeomorphic transformations (Qiao and Minematsu 2010, Theorem 1). This is a powerful statement: if the ground metric is affected by such a transformation h (for example, we change the underlying manifold coordinate system, *e.g.* for privacy reasons), then, from the optimal TROT transportation plan $P^\star$, the transportation plan corresponding to the initial coordinate system can be recovered from the *sole* knowledge of h^{-1}.

The algorithms we provide allow for the efficient optimization of the regularized optimal transport for all values of $q \geq 0$, and include notable cases for which conventional gradient-based approaches would probably not be the best approaches due to the fact that the function to optimize is not Lipschitz for the q chosen. In fact, the main notable downside of the generalization is that we could not prove the same (geometric) convergence rates as the ones that are known for Sinkhorn's approach (Franklin and Lorenz 1989). This is an important avenue for future work, as recent related work on generalizations of regularized optimal transport (Dessein, Papadakis, and Rouas 2016) also remain far from such rates — our second order approximation of the auxiliary function for the convergence proof should provide an uplift for better convergence rates (Muzellec et al. 2017) than just linear or quadratic (Dessein, Papadakis, and Rouas 2016) .

Our results display that there can be significant discrepancies in the regularized optimal transport results depending on how cost matrix M is crafted, yet the information we used for our best experiments is readily available from public statistics (matrices $M^{\text{RBF}}, M^{\text{sur}}$). Even the instantiation without prior knowledge ($M = \mathbf{1}\mathbf{1}^\top$) does not strictly fail in returning useful solutions (compared *e.g.* to Florida average and unregularized optimal transport). This may be a strong advocacy to use TROT even on domains for which little prior knowledge is available.

References

Amari, S.-I. 2016. *information geometry and its applications*. Springer-Verlag, Berlin.

Beck, A., and Teboulle, M. 2003. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters* 31:167 – 175.

Cho, W.-K.-T., and Manski, C.-F. 2008. Cross level/ecological inference. *Oxford Handbook of Political Methodology* 547–569.

Collins, M.; Schapire, R.; and Singer, Y. 2002. Logistic regression, adaboost and Bregman distances. *MLJ* 253–285.

Cross, P.-J., and Manski, C.-F. 2002. Regressions, short and long. *Econometrica* 70(1):357–368.

Csiszár, I. 1989. A geometric interpretation of Darroch and Ratcliff's generalized iterative scaling. *Ann. of Stat.* 17:1409–1413.

Cuturi, M., and Doucet, A. 2014. Fast computation of wasserstein barycenters. In *31st ICML*, 685–693.

Cuturi, M. 2013. Sinkhorn distances: lightspeed computation of optimal transport. In *NIPS*26*, 2292–2300.

Della Pietra, S.; Della Pietra, V.-J.; and Lafferty, J.-D. 1997. Inducing features of random fields. *IEEE Trans. PAMI* 19(4):380–393.

Dessein, A.; Papadakis, N.; and Rouas, J.-L. 2016. Regularized optimal transport and the rot mover's distance. *CoRR* abs/1610.06447.

Donoso, S.; Marín, N.; and Vila, M.-A. 2005. Systems of possibilistic regressions: A case study in ecological inference. *Mathware and Soft Computing* 12:169–184.

Duncan, O.-D., and Davis, B. 1953. An alternative to ecological correlation. *American sociological review* 665–666.

Flaxman, S.-R.; Wang, Y.-X.; and Smola, A.-J. 2015. Who supported obama in 2012?: Ecological inference through distribution regression. In *21st KDD*, 289–298.

Forcina, A., and Marchetti, G.-M. 2011. The Brown and Payne model of voter transition revisited. In Ingrassia, S.; Rocci, R.; and Vichi, M., eds., *New Perspectives in Statistical Modeling and Data Analysis*. Springer. 481–488.

Franklin, J., and Lorenz, J. 1989. On the scaling of multidimensional matrices. *Linear Algebra and Applications* 114:717–735.

Gallup. 2013. http://www.gallup.com/poll/160373/democrats-racially-diverse-republicans-mostly-white.aspx.

Genevay, A.; Cuturi, M.; Peyré, G.; and Bach, F. 2016. Stochastic optimization for large-scale optimal transport. In *NIPS*29*.

Imai, K., and Khanna, K. 2016. Improving ecological inference by predicting individual ethnicity from voter registration records. *Political Analysis* 24:263–272.

Judge, G.-G.; Miller, D.-J.; and Cho, W.-K.-T. 2004. An information theoretic approach to ecological estimation and inference. In King, G.; Rosen, O.; and Tanner, M., eds., *Ecological inference: New methodological strategies*. Cambridge University Press. 162–187.

Kantorovitch, L. 1958. On the translocation of masses. *Management Science* 1–4.

King, G.; Tanner, M.-A.; and Rosen, O. 2004. *Ecological inference: New methodological strategies*. Cambridge University Press.

King, G. 1997. *A solution to the ecological inference problem: reconstructing individual behavior from aggregate data*. Princeton University Press.

Kusner, M.-J.; Sun, Y.; Kolkin, N.-I.; and Weinberger, K.-Q. 2015. From word embeddings to document distances. In *32nd ICML*, 957–966.

Monge, G. 1781. Mémoire sur la théorie des déblais et des remblais. *Académie Royale des Sciences de Paris* 666–704.

Muzellec, B.; Nock, R.; Patrini, G.; and Nielsen, F. 2017. Tsallis regularized optimal transport and ecological inference. In *31st AAAI*. (Supplementary Material).

Posnett, D.; Filkov, V.; and Devanbu, P. 2011. Ecological inference in empirical software engineering. In *Proceedings of the 2011 26th IEEE/ACM International Conference on Automated Software Engineering*, 362–371.

Qian, W.; Hong, B.; Cai, D.; He, X.; and Li, X. 2016. Non-negative matrix factorization with Sinkhorn distance. In *25 th IJCAI*, 1960–1966.

Qiao, Y., and Minematsu, N. 2010. A study on invariance of f-divergence and its application to speech recognition. *IEEE Trans. SP* 58:3884–3890.

Robinson, W.-S. 1950. Ecological correlations and the behavior of individuals. *American Sociological Review* 15(3):351–357.

Rubner, Y.; Tomasi, C.; and Guibas, L.-J. 2000. The earth movers distance as a metric for image retrieval. *Int. J. Comp. Vis.* 40:99–121.

Sinkhorn, R., and Knopp, P. 1967. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific J. Math.* 21:343–348.

Sinkhorn, R. 1964. A relationship between arbitrary positive matrices and doubly stochastic matrices. *Annals of Mathematical Statistics* 35:876–879.

Sinkhorn, R. 1967. Diagonal equivalence to matrices with prescribed row and column sums. *American Mathematical Monthly* 74:402–405.

Solomon, J.; de Goes, F.; Peyré, G.; Cuturi, M.; Butscher, A.; Nguyen, A.; Du, T.; and Guibas, L. 2015. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics* 34:66:1–66:11.

Soules, G.-W. 1991. The rate of convergence of Sinkhorn balancing. *Linear Algebra and Applications* 3–40.

Tsallis, C. 1988. Possible generalization of Boltzmann-Gibbs statistics. *J. of Statistical Physics* 52:479–487.

Villani, C. 2009. *Optimal transport: old and new*. Springer.

Wakefield, J., and Shaddick, G. 2006. Health-exposure modeling and the ecological fallacy. *Biostatistics* 7(3):438–455.