Tsallis Regularized Optimal Transport and Ecological Inference — Supplementary Material —

Abstract

This is the supplementary material to the AAAI'17 paper "Tsallis Regularized Optimal Transport and Ecological Inference", by B. Muzellec, R. Nock, G. Patrini and F. Nielsen.

1 Table of contents

Supplementary Material: proofs

2 Proof of Theorem 3

The proof encompasses a more general statement than that of Theorem 3 since we show that M may actually not even be a distance matrix for the Theorem to hold.

($\star$) Let $M \in \mathbb{R}_+^{n\times n}$ be a distance matrix, and $q, q' \in \mathbb{R} - \{1\}, q \neq q'$ (the case when $q = 1$ xor $q' = 1$ can be treated in a similar fashion). We suppose wlog that the support does not reduce to a singleton (otherwise the solution to optimal transport is trivial). Rescaling M and a constant row vector and a constant column vector, the solution of $\textsc{trot}(q, \lambda, M)$ can be written wlog as

$$p_{ij} \quad = \quad \exp_q(-1)\exp_q^{-1}(m_{ij}) \ . \tag{1}$$

Assume there exists a $\lambda' \in \mathbb{R}$ such that the solution of $\textsc{trot}(q', \lambda', M)$ is equal to that of $\textsc{trot}(q, \lambda, M)$. This is equivalent to saying that there exists $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^n$ such that

$$\exp_q(m_{ij}) \quad = \quad \exp_{q'}(\alpha_i + \lambda' m_{ij} + \beta_j) \ , \forall i, j \ . \tag{2}$$

Composing with $\log_{q'}$ and rearranging, this implies that

$$f^{\lambda'}_{q',q}(m_{ij}) \quad = \quad \alpha_i + \beta_j \ , \forall i, j \ , \tag{3}$$

where

$$f^{\lambda'}_{q',q}(x) \quad \doteq \quad \log_{q'} \circ \exp_q - \lambda' \mathrm{Id} \ . \tag{4}$$

Now, remark that, since M is a distance, $m_{ii} = 0, \forall i$ because of the identity of the indiscernibles, and so $\alpha_i + \beta_i = f^{\lambda'}_{q',q}(0) = 0$, implying $\boldsymbol{\alpha} = -\boldsymbol{\beta}$. $f^{\lambda'}_{q',q}$ is differentiable. Let:

$$\begin{aligned} g^{\lambda'}_{q',q}(x) \quad &\doteq \quad \frac{\mathrm{d}}{\mathrm{d}x} f^{\lambda'}_{q',q}(x) \\ &= \quad \exp_q^{q-q'}(x) - \lambda' \ ; \end{aligned} \tag{5}$$

$$\begin{aligned} h^{\lambda'}_{q',q}(x) \quad &\doteq \quad \frac{\mathrm{d}}{\mathrm{d}x} g^{\lambda'}_{q',q}(x) \\ &= \quad (q - q') \cdot \exp_q^{2q-q'-1}(x) \ . \end{aligned} \tag{6}$$

If we assume wlog that $q > q'$, then $g^{\lambda'}_{q',q}$ is increasing and zeroes at most once over $\mathbb{R}$, eventually on some m^* that we define as:

$$m^* \doteq \left\{ \begin{array}{rcl} \log_q\left(\lambda'^{\frac{1}{q-q'}}\right) & \text{if} & (\lambda' > 1) \wedge (0 \in \mathrm{Img} g^{\lambda'}_{q',q}) \\ +\infty & \text{otherwise} & \end{array} \right. \ . \tag{7}$$

Notice that $m^* > 0$ and $f^{\lambda'}_{q',q}$ is bijective over $(0, m^*)$. Suppose wlog that $m_{ij} \leq m^*, \forall i, j$. Otherwise, all distances are scaled by the same real so that $m_{ij} \leq m^*, \forall i, j$:

this does not alter the property of M being a distance. A distance being symmetric, we also have $m_{ij} = m_{ji}$ and since $f^{\lambda'}_{q',q}$ is strictly increasing in the range of distances, then we get from eq. (3) that $\alpha_i + \beta_j = \alpha_j + \beta_i, \forall i, j$ and so $\alpha_i - \alpha_j = \beta_i - \beta_j = -(\alpha_i - \alpha_j)$ (since $\boldsymbol{\alpha} = -\boldsymbol{\beta}$). Hence, there exists a real α such that $\boldsymbol{\alpha} = \alpha \cdot \mathbf{1}$. We get, in matrix form

$$f^{\lambda'}_{q',q}(M) = \boldsymbol{\alpha}\mathbf{1}^\top + \mathbf{1}\boldsymbol{\beta}^\top \quad (8)$$

$$= \alpha \cdot \mathbf{1}\mathbf{1}^\top - \alpha \cdot \mathbf{1}\mathbf{1}^\top = 0 \ . \quad (9)$$

Hence, $m_{ij} = m_{ii}, \forall i, j$ and the support reduces to a singleton (because of the identity of the indiscernibles), which is impossible.

($\star$) Remark that the proof also works when M is not a distance anymore, but for example contains all arbitrary non negative matrices. To see this, we remark that the right hand side of eq. (8) is a matrix of rank no larger than 2. Since $f^{\lambda'}_{q',q}$ is continuous, we have

$$\mathrm{Im}(f^{\lambda'}_{q',q}) \doteq \mathcal{I} \subseteq \mathbb{R}$$

where $\mathcal{I}$ is not reduced to a singleton and so the left hand side of eq. (8) spans matrices of arbitrary rank. Hence, eq. (8) cannot always hold.

3 Proof of Theorem 4

Denote

$$f_{ij} : p_{ij} \to p_{ij} m_{ij} - \frac{1}{\lambda(1-q)}(p_{ij}^q - p_{ij}) \ .$$

f_{ij} is twice differentiable on $\mathbb{R}_{+*}$, and

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2} f_{ij}(x) = \frac{q}{\lambda} x^{q-2} > 0$$

for any fixed $q > 0$, and so f_{ij} is strictly convex on $\mathbb{R}_{+*}$. We also remark that $U(\boldsymbol{r}, \boldsymbol{c})$ is a non-empty compact subset of $\mathbb{R}^{n\times n}$. Indeed, $\boldsymbol{r}\boldsymbol{c}^\top \in U(\boldsymbol{r}, \boldsymbol{c}), \forall P \in U(\boldsymbol{r}, \boldsymbol{c}), \|P\|_1 = 1$ (which proves boundedness) and $U(\boldsymbol{r}, \boldsymbol{c})$ is a closed subset of $U(\boldsymbol{r}, \boldsymbol{c})$ (being the intersection of the pre-images of singletons by continuous functions). Hence, since $\langle P, M\rangle - \frac{1}{\lambda} H_q(P) = \sum_{i,j} f_{ij}(p_{ij})$, there exists a unique minimum of this function in $U(\boldsymbol{r}, \boldsymbol{c})$.

To prove the analytic shape of the solution, we remark that $\textsc{trot}(q, \lambda, M)$ consists in minimizing a convex function given a set of affine constraints, and so the KKT conditions are necessary and sufficient. The KKT conditions give

$$p_{ij} = \exp_q(-1) \exp_q^{-1}(\alpha_i + \lambda m_{ij} + \beta_j) \ ,$$

where $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^n$ are Lagrange multipliers.

Finally, let us show that Lagrange multipliers $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^n$ are unique up to an additive constant. Assume that $\boldsymbol{\alpha}, \boldsymbol{\alpha}', \boldsymbol{\beta}, \boldsymbol{\beta}' \in \mathbb{R}^n$ are such that

$$\begin{aligned}\forall i, j, p_{ij} &= \exp_q(-1) \exp_q^{-1}(\lambda m_{ij} + \alpha_i + \beta_j) \\ &= \exp_q(-1) \exp_q^{-1}(\lambda m_{ij} + \alpha'_i + \beta'_j) \ ,\end{aligned}$$

where P is the unique solution of $\textsc{trot}(q, \lambda, M)$. This implies

$$\alpha_i + \beta_j \quad = \quad \alpha'_i + \beta'_j \ , \forall i, j \ ,$$

i.e.

$$\alpha_i - \alpha'_i \quad = \quad \beta'_j - \beta_j \ , \forall i, j \ .$$

In particular, if there exists i_0 and $C \neq 0$ such that $\alpha_{i_0} - \alpha'_{i_0} = C$, then $\forall j, \beta'_j = \beta_j + C$ and in turn $\forall i, \alpha_i = \alpha'_i + C$, which proves our claim.

4 Proof of Theorems 5 and 6

For reasons that we explain now, we will in fact prove Theorem 6 before we prove Theorem 5.
Had we chosen to follow [4], we would have replaced $\textsc{trot}(q, \lambda, M)$ by:

$$d_{M,\alpha,q}(\boldsymbol{r}, \boldsymbol{c}) \quad \doteq \quad \min_{\substack{P \in U(\boldsymbol{r},\boldsymbol{c}) \\ H_q(P) - H_q(\boldsymbol{r}) - H_q(\boldsymbol{c}) \geq \alpha}} \langle P, M \rangle \ , \tag{10}$$

for some $\alpha > 0$. Both problems are equivalent since λ in $\textsc{trot}(q, \lambda, M)$ plays the role of the Lagrange multiplier for the entropy constraint in eq. (10) [4, Section 3], and so *there exists* an equivalent value of α^* for which both problems coincide:

$$d_{M,\alpha^*,q}(\boldsymbol{r}, \boldsymbol{c}) \quad = \quad d_M^{\lambda,q}(\boldsymbol{r}, \boldsymbol{c}) \ , \tag{11}$$

so eq. (10) indeed matches $\textsc{trot}(q, \lambda, M)$. It is clear from eq. (11) that α does not depend solely on λ, *but also* (eventually) on all other parameters, including $\boldsymbol{r}, \boldsymbol{c}$.

This would not be a problem to state the triangle inequality *for* $d_{M,\alpha,q}$, as in [4] ($\forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \triangle_n$):

$$d_{M,\alpha,q}(\boldsymbol{x}, \boldsymbol{z}) \quad \leq \quad d_{M,\alpha,q}(\boldsymbol{x}, \boldsymbol{y}) + d_{M,\alpha,q}(\boldsymbol{y}, \boldsymbol{z}) \ . \tag{12}$$

However, α is *fixed* and in particular different from the α^* that guarantee eq. (11) — and there might be three different sets of parameters for $d_M^{\lambda,q}$ as it would equivalently appear from eq. (12). Under the simplifying assumption that only λ changes, we might just get from eq. (12):

$$d_M^{\lambda^*,q}(\boldsymbol{x}, \boldsymbol{z}) \quad \leq \quad d_M^{\lambda'^*,q}(\boldsymbol{x}, \boldsymbol{y}) + d_M^{\lambda''^*,q}(\boldsymbol{y}, \boldsymbol{z}) \ , \tag{13}$$

with $\lambda^* \neq \lambda'^* \neq \lambda''^*$. Worse, the transportation plans may change with λ: for example, we may have

$$\arg\min_{P\in U(\boldsymbol{x},\boldsymbol{z})} d_M^{\lambda_1,q}(\boldsymbol{x},\boldsymbol{z}) \quad \neq \quad \arg\min_{P\in U(\boldsymbol{x},\boldsymbol{z})} d_M^{\lambda_2,q}(\boldsymbol{x},\boldsymbol{z}) \enspace ,$$

with $\lambda_1 \neq \lambda_2$ and $\lambda_1, \lambda_2 \in \{\lambda^*, \lambda'^*, \lambda''^*\}$. So, the triangle inequality for $d_M^{\lambda,q}$ that follows from ineq. (12) does not allow to control the parameters of $\mathrm{TROT}(q,\lambda,M)$ nor the optimal transportation plans that follows. It does not show a problem in regularizing the optimal transport distance, but rather that the distance $d_{M,\alpha,q}$ chosen from eq. (11) does not completely fulfill its objective in showing that regularization in $d_M^{\lambda,q}$ still keeps some of the attractive properties that unregularized optimal transport meets.

To bypass this problem and establish a statement involving a distance in which all parameters are in the clear and optimal transportation plans still coincide with $d_M^{\lambda,q}$, we chose to rely on measure:

$$\begin{aligned} d_M^{\lambda,q,\beta}(\boldsymbol{r},\boldsymbol{c}) \quad &\doteq \quad \min_{P\in U(\boldsymbol{r},\boldsymbol{c})} \langle P, M\rangle \\ &\quad -\frac{1}{\lambda}\cdot\left(H_q(P) - \beta\cdot(H_q(\boldsymbol{r}) + H_q(\boldsymbol{c}))\right) \enspace , \end{aligned}$$

where β is some *constant*. There is one trivial but crucial fact about $d_M^{\lambda,q,\beta}(\boldsymbol{r},\boldsymbol{c})$: regardless of the choice of β, its optimal transportation plan is the *same* as for $\mathrm{TROT}(q,\lambda,M)$.

Lemma 1 *For any $\boldsymbol{r},\boldsymbol{c}\in\triangle_n$ and constant $\beta\in\mathbb{R}$, let*

$$\begin{aligned} P_1 \quad &\doteq \quad \arg\min_{P\in U(\boldsymbol{r},\boldsymbol{c})} \langle P, M\rangle \\ &\quad -\frac{1}{\lambda}\cdot\left(H_q(P) - \beta\cdot(H_q(\boldsymbol{r}) + H_q(\boldsymbol{c}))\right) \enspace . \qquad (14) \\ P_2 \quad &\doteq \quad \arg\min_{P\in U(\boldsymbol{r},\boldsymbol{c})} \langle P, M\rangle \\ &\quad -\frac{1}{\lambda}\cdot\left(H_q(P)\right) \enspace . \qquad (15) \end{aligned}$$

Then $P_1 = P_2$.

Theorem 2 *The following holds for any fixed $q\geq 1$ (unless otherwise stated):*

- *for **any** $\beta \geq 1$, $d_M^{\lambda,1,\beta}$ satisfies the triangle inequality;*
- *for the choice $\beta = 1/2$, $d_M^{\lambda,q,1/2}$ satisfies the following weak version of the identity of the indiscernibles: if $\boldsymbol{r}=\boldsymbol{c}$, then $d_M^{\lambda,q,1/2}(\boldsymbol{r},\boldsymbol{c})\leq 0$.*
- *for the choice $\beta = 1/2$, $\forall \boldsymbol{r}\in\triangle_n$, choosing the (no) transportation plan $P = \mathrm{Diag}(\boldsymbol{r})$ brings*

$$\langle P, M\rangle - \frac{1}{\lambda}\cdot\left(H_q(P) - \frac{1}{2}\cdot(H_q(\boldsymbol{r}) + H_q(\boldsymbol{r}))\right) \quad = \quad 0 \enspace .$$

Remark: the last property is trivial but worth stating since the (no) transportation plan $P = \mathrm{Diag}(\boldsymbol{r})$ also satisfies $P = \arg\min_{Q \in U(\boldsymbol{r},\boldsymbol{r})} \langle Q, M\rangle$, which zeroes the (no) transportation distance $d_M(\boldsymbol{r},\boldsymbol{r})$. Remark that in this case, $P = \mathrm{Diag}(\boldsymbol{r})$ amounts to making no transportation in the support of the marginal, hence the "(no) transportation" name.

Proof To prove the Theorem, we need another version of the Gluing Lemma with entropic constraints [4, Lemma 1], generalized to handle Tsallis entropy.

Lemma 3 *(Refined gluing Lemma) Let $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \triangle_n$. Let $P \in U(\boldsymbol{x},\boldsymbol{y})$ and $Q \in U(\boldsymbol{y},\boldsymbol{z})$. Let $S \in \mathbb{R}^{n\times n}$ defined by general term*

$$s_{ik} \doteq \sum_j \frac{p_{ij}q_{jk}}{y_j} \ . \tag{16}$$

The following holds about S:

1. *$S \in U(\boldsymbol{x},\boldsymbol{z})$;*
2. *if $q \geq 1$, then:*

$$\begin{aligned} & H_q(S) - H_q(\boldsymbol{x}) - H_q(\boldsymbol{z}) \\ \geq \quad & H_q(P) - H_q(\boldsymbol{x}) - H_q(\boldsymbol{y}) \ . \end{aligned} \tag{17}$$

Proof The proof essentially builds upon [4, Lemma 1]. We remark that S can be built by

$$s_{ik} = \sum_j t_{ijk} \ , \tag{18}$$

where $\forall i, j, k \in \{1, 2, ..., n\}$, we have

$$t_{ijk} \doteq \begin{cases} \frac{p_{ij}q_{jk}}{y_j} & \text{if} \quad y_j \neq 0 \\ 0 & \text{otherwise} \end{cases} \ . \tag{19}$$

S is a transportation matrix between $\boldsymbol{x}$ and $\boldsymbol{z}$. Indeed,

$$\begin{aligned} \sum_i \sum_j s_{ijk} &= \sum_j \sum_i \frac{p_{ij}q_{jk}}{y_j} \\ &= \sum_j \frac{q_{jk}}{y_j} \sum_i p_{ij} \\ &= \sum_j \frac{q_{jk}}{y_j} y_j = \sum_j q_{jk} = z_k \ ; \\ \sum_k \sum_j s_{ijk} &= \sum_j \sum_k \frac{p_{ij}q_{jk}}{y_j} \\ &= \sum_j \frac{p_{ij}}{y_j} \sum_k q_{jk} \\ &= \sum_j \frac{p_{ij}}{y_j} y_j = \sum_j p_{ij} = x_i \ . \end{aligned}$$

So, $S \in U(\boldsymbol{x}, \boldsymbol{z})$. To prove ineq. (17), we need the following definition from [6].

Definition 4 *[6] Let* X *and* Y *denote random variables. The Tsallis conditional entropy of* X *given* Y*, and Tsallis joint entropy of* X *and* Y*, are respectively given by:*

$$\begin{aligned} H_q(\mathsf{X}|\mathsf{Y}) &\doteq -\sum_{x,y} p(x,y)^q \log_q p(x|y) \ , \\ H_q(\mathsf{X},\mathsf{Y}) &\doteq -\sum_{x,y} p(x,y)^q \log_q p(x,y) \ . \end{aligned}$$

The Tsallis mutual entropy of X *and* Y *is defined by*

$$\begin{aligned} I_q(\mathsf{X};\mathsf{Y}) &\doteq H_q(\mathsf{X}) - H_q(\mathsf{X}|\mathsf{Y}) \\ &= H_q(\mathsf{X}) + H_q(\mathsf{Y}) - H_q(\mathsf{X},\mathsf{Y}) \ . \end{aligned}$$

We have made use of the simplifying notation that removes variables names when unambiguous, like $p(x) \doteq p(\mathsf{X} = x)$. Let $\mathsf{X}, \mathsf{Y}, \mathsf{Z}$ be random variables jointly distributed as T, that is, for any x, y, z,

$$p(x,y,z) = \frac{p(x,y)p(y,z)}{p(y)} \tag{20}$$

It follows from that and Bayes rule that:

$$\begin{aligned} p(x|y) &= \frac{p(x,y)}{p(y)} \\ &= \frac{p(x,y,z)}{p(y,z)} \ , \forall z \\ &= p(x|y,z) \ , \forall z \ , \end{aligned} \tag{21}$$

and so

$$\begin{aligned} I_q(\mathsf{X};\mathsf{Z}|\mathsf{Y}) &\doteq H_q(\mathsf{X}|\mathsf{Y}) - H_q(\mathsf{X}|\mathsf{Y},\mathsf{Z}) \\ &= 0 \ . \end{aligned} \tag{22}$$

It comes from [6, Theorem 4.3],

$$I_q(\mathsf{X};\mathsf{Y},\mathsf{Z}) = I_q(\mathsf{X};\mathsf{Z}) + I_q(\mathsf{X};\mathsf{Y}|\mathsf{Z}) \tag{23}$$

$$\phantom{I_q(\mathsf{X};\mathsf{Y},\mathsf{Z})} = I_q(\mathsf{X};\mathsf{Y}) + I_q(\mathsf{X};\mathsf{Z}|\mathsf{Y}) \ , \tag{24}$$

but since $I_q(\mathsf{X};\mathsf{Z}|\mathsf{Y}) = 0$, we obtain

$$I_q(\mathsf{X};\mathsf{Y}) = I_q(\mathsf{X};\mathsf{Z}) + I_q(\mathsf{X};\mathsf{Y}|\mathsf{Z}) \ . \tag{25}$$

It also follows from [6, Theorem 3.4] that $I_q(\mathsf{X};\mathsf{Y}|\mathsf{Z}) \geq 0$ whenever $q \geq 1$, and so

$$I_q(\mathsf{X};\mathsf{Y}) \geq I_q(\mathsf{X};\mathsf{Z}) \ , \forall q \geq 1 \ . \tag{26}$$

Now, it comes from Definition 4 and the definition of X, Y and Z from eq. (20),

$$\begin{aligned} -I_q(\mathsf{X};\mathsf{Y}) &= H_q(\mathsf{X},\mathsf{Y}) - H_q(\mathsf{X}) - H_q(\mathsf{Y}) \\ &= H_q(P) - H_q(\boldsymbol{x}) - H_q(\boldsymbol{y}) \ , \qquad (27) \\ -I_q(\mathsf{X};\mathsf{Z}) &= H_q(\mathsf{X},\mathsf{Z}) - H_q(\mathsf{X}) - H_q(\mathsf{Z}) \\ &= H_q(S) - H_q(\boldsymbol{x}) - H_q(\boldsymbol{z}) \ . \qquad (28) \end{aligned}$$

Since $P \in U_\lambda(\boldsymbol{x}, \boldsymbol{y})$, by assumption, we obtain from ineq. (26) that whenever $q \geq 1$,

$$H_q(S) - H_q(\boldsymbol{x}) - H_q(\boldsymbol{z}) \geq H_q(P) - H_q(\boldsymbol{x}) - H_q(\boldsymbol{y}) \ ,$$

as claimed. ■

We can now prove Theorem 2. Shannon's entropy is denoted H_1 for short.

Define for short

$$\Delta \doteq H_1(P) + H_1(Q) - H_1(S) - 2\beta \cdot H_1(\boldsymbol{y}) \ , \qquad (29)$$

where P, Q, S are defined in Lemma 3. It follows from the definition of S and [4, Proof of Theorem 1] that

$$\begin{aligned} & d_M^{\lambda,q,\beta}(\boldsymbol{x}, \boldsymbol{z}) \\ \doteq\ & \min_{R \in U(\boldsymbol{x},\boldsymbol{z})} \langle R, M\rangle - \frac{1}{\lambda} \cdot (H_1(R) - \beta \cdot (H_1(\boldsymbol{x}) + H_1(\boldsymbol{z}))) \\ \leq\ & \langle S, M\rangle - \frac{1}{\lambda} \cdot (H_1(S) - \beta \cdot (H_1(\boldsymbol{x}) + H_1(\boldsymbol{z}))) \\ \leq\ & \langle P, M\rangle + \langle Q, M\rangle - \frac{1}{\lambda} \cdot (H_1(S) - \beta \cdot (H_1(\boldsymbol{x}) + H_1(\boldsymbol{z}))) \\ & = \langle P, M\rangle - \frac{1}{\lambda} \cdot (H_1(P) - \beta \cdot (H_1(\boldsymbol{x}) + H_1(\boldsymbol{y}))) \\ & + \langle Q, M\rangle - \frac{1}{\lambda} \cdot (H_1(Q) - \beta \cdot (H_1(\boldsymbol{y}) + H_1(\boldsymbol{z}))) \\ & + \frac{1}{\lambda} \cdot (H_1(P) + H_1(Q) - H_1(S) - 2\beta \cdot H_1(\boldsymbol{y})) \\ \doteq\ & d_M^{\lambda,q,\beta}(\boldsymbol{x}, \boldsymbol{y}) + d_M^{\lambda,q,\beta}(\boldsymbol{y}, \boldsymbol{z}) + \frac{1}{\lambda} \cdot \Delta \ . \qquad (30) \end{aligned}$$

We now show that $\Delta \leq 0$. For this, observe that ineq. (17) yields:

$$\begin{aligned} & \Delta \\ \leq\ & (H_1(S) + H_1(\boldsymbol{y}) - H_1(\boldsymbol{z})) \\ & + H_1(Q) - H_1(S) - 2\beta \cdot H_1(\boldsymbol{y}) \\ & = H_1(Q) - H_1(\boldsymbol{y}) - H_1(\boldsymbol{z}) + 2(1-\beta) H_1(\boldsymbol{y}) \ , \qquad (31) \end{aligned}$$

and, by definition of $Q, \boldsymbol{y}, \boldsymbol{z}$,

$$\begin{aligned} & H_1(Q) - H_1(\boldsymbol{y}) - H_1(\boldsymbol{z}) \\ \doteq\ & H_1(\mathsf{Y},\mathsf{Z}) - H_1(\mathsf{Y}) - H_1(\mathsf{Z}) \ . \qquad (32) \end{aligned}$$

Shannon's entropy of a joint distribution is maximal with independence: $H_1(\mathsf{Y},\mathsf{Z}) \leq H_1(\mathsf{Y}\times\mathsf{Z}) = H_1(\mathsf{Y}) + H_1(\mathsf{Z})$, so we get from eq. (31) after simplifying

$$\Delta \quad \leq \quad 2(1-\beta)H_1(\boldsymbol{y}) \ . \tag{33}$$

Hence if $\beta \geq 1$, then $\Delta \leq 0$. We get that for any $\beta \geq 1$,

$$d_M^{\lambda,1,\beta}(\boldsymbol{x},\boldsymbol{z}) \quad \leq \quad d_M^{\lambda,1,\beta}(\boldsymbol{x},\boldsymbol{y}) + d_M^{\lambda,1,\beta}(\boldsymbol{y},\boldsymbol{z}) \ , \tag{34}$$

and $d_M^{\lambda,1,\beta}$ satisfies the triangle inequality. For $\beta = 1/2$, it is trivial to check that for any $\boldsymbol{x} \in \triangle_n$, the (no) transportation plan $P = \mathrm{Diag}(\boldsymbol{x})$ is in $U(\boldsymbol{x},\boldsymbol{x})$ and satisfies

$$\begin{aligned}&\langle P, M\rangle - \frac{1}{\lambda}\cdot\left(H_q(P) - \frac{1}{2}\cdot(H_q(\boldsymbol{x}) + H_q(\boldsymbol{x}))\right)\\ &= \quad 0 - \frac{1}{\lambda}\cdot(H_q(\boldsymbol{x}) - H_q(\boldsymbol{x})) = 0 \ .\end{aligned} \tag{35}$$

This ends the proof of Theorem 2. ■

Notice that Theorem 6 is in fact a direct consequence of Theorem 2. To finish up, we now prove Theorem 5. To simplify notations, let

$$U_\alpha(\boldsymbol{r},\boldsymbol{c}) \quad \doteq \quad \left\{ \begin{array}{l} P \in U(\boldsymbol{r},\boldsymbol{c}) : \\ H_q(P) - H_q(\boldsymbol{r}) - H_q(\boldsymbol{c}) \geq \alpha(\lambda) \end{array} \right\} \ . \tag{36}$$

Suppose P, Q in Lemma 3 are such that $P, Q \in U_\lambda(\boldsymbol{x},\boldsymbol{y})$. In this case,

$$H_q(P) - H_q(\boldsymbol{x}) - H_q(\boldsymbol{y}) \quad \geq \quad \alpha \tag{37}$$

and so point 2. in Lemma 3 brings

$$H_q(S) - H_q(\boldsymbol{x}) - H_q(\boldsymbol{z}) \quad \geq \quad \alpha \ , \tag{38}$$

so $S \in U_\lambda(\boldsymbol{x},\boldsymbol{z})$. The proof of [4, Theorem 1] can then be used to show that $\forall \boldsymbol{x},\boldsymbol{y},\boldsymbol{z} \in \triangle_n$,

$$d_{M,\alpha,q}(\boldsymbol{x},\boldsymbol{z}) \quad \leq \quad d_{M,\alpha,q}(\boldsymbol{x},\boldsymbol{y}) + d_{M,\alpha,q}(\boldsymbol{y},\boldsymbol{z}) \ . \tag{39}$$

It is easy to check that $d_{M,\alpha,q}$ is non negative and that $\mathbb{1}_{\{\boldsymbol{r}=\boldsymbol{c}\}} d_{M,\alpha,q}(\boldsymbol{r},\boldsymbol{c})$ meets, in addition, the identity of the indiscernibles. This achieves the proof of Theorem 5.

5 Proof of Theorem 7

Basic facts and definitions — In this proof, we make two simplifying assumptions: (i) we consider matrices either as matrices or as vectorized matrices without ambiguity, and (ii) we let $\phi(P) \doteq -H_q(P)$, noting that the domain of ϕ is $\triangle_{n^2}$ (nonnegative matrices with row- and column-sums in the simplex) when $P \in U(\boldsymbol{r},\boldsymbol{c})$. Since ϕ is convex, we can define a *Bregman divergence* with generator D_ϕ [2] as:

$$D_\phi(P\|R) \quad \doteq \quad \phi(P) - \phi(R) - \langle\nabla\phi(R), P - R\rangle \ .$$

We define

$$a_{ij} \doteq \alpha_i + \lambda m_{ij} + \beta_j \ , \tag{40}$$

so that

$$p_{ij} = \exp_q(-1)\exp_q^{-1}(a_{ij}) \tag{41}$$

in eq. (7) (main file). Finally, let us denote for short

$$D_q(P\|R) \doteq K_{1/q}(P^q, R^q) \ , \tag{42}$$

so that we can, reformulate eq. (6) (main file) as:

$$d_M^{\lambda,q}(\boldsymbol{r},\boldsymbol{c}) = \frac{1}{\lambda}\cdot \min_{P\in U(\boldsymbol{r},\boldsymbol{c})} D_q(P\|\tilde{U}) + g(M) \ , \tag{43}$$

and our objective "reduces" to the minimization of $D_q(P\|\tilde{U})$ over $U(\boldsymbol{r},\boldsymbol{c})$. In SO–TROT (Algorithm 1), we just care for a single constraint out of the two possible in $U(\boldsymbol{r},\boldsymbol{c})$, so we will focus without loss of generality on the row constraint and therefore to the solution of:

$$P^\star \doteq \arg\min_{P\in\mathbb{R}_+^{n\times n}:P\boldsymbol{1}=\boldsymbol{r}} D_q(P\|\tilde{U}) \ . \tag{44}$$

The same result would apply to the column constraint.
Convergence proof — We reuse the theory of *auxiliary functions* developed for the iterative constrained minimization of Bregman divergences [2, 5]. We reuse notation "$\diamond$" following [3, 7] and define for any $\boldsymbol{y}\in\mathbb{R}^n, P\in\mathbb{R}^{n\times n}$ matrix $\boldsymbol{y}\diamond_q P\in\mathbb{R}^{n\times n}$ such that

$$\begin{aligned}
&(\boldsymbol{y}\diamond_q P)_{ij} \\
&\doteq \frac{\exp_q^{-1}(y_i)p_{ij}}{\exp_q\left[(1-q)y_i\exp_q^{1-q}(y_i)\log_q(p_{ij})\right]} \ .
\end{aligned} \tag{45}$$

We also define key matrix $\tilde{P}\in\mathbb{R}^{n\times n}$ with:

$$\tilde{P} \doteq \boldsymbol{r}\boldsymbol{c}^\top \ . \tag{46}$$

Let us denote

$$\begin{aligned}
\mathcal{Q} &\doteq \left\{\begin{array}{l} Q\in\mathbb{R}^{n\times n}: \\ Q=\exp_q(-1)\exp_q^{-1}(\boldsymbol{\alpha}^\top\boldsymbol{1}+\lambda M+\boldsymbol{1}^\top\boldsymbol{\beta}) \ , \text{ with } \boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{R}^n \end{array}\right\} \ . \\
\mathcal{P} &\doteq \{P\in\triangle_{n^2}: P\boldsymbol{1}=\tilde{P}\boldsymbol{1}=\boldsymbol{r}\} \ .
\end{aligned}$$

One function will be key.

Definition 5 *We define* $A(P, \boldsymbol{y}) \doteq \sum_i A_i(P, \boldsymbol{y})$*, with:*

$$\begin{aligned} & A_i(P, \boldsymbol{y}) \\ \doteq \quad & y_i r_i + \sum_j (p_{ij}^q - \exp_q^q(-1) \exp_q^{-q}(a_{ij} - y_i)) \ . \end{aligned} \tag{47}$$

Here a_{ij} *is defined in eq. (40),* r_i *is the* i*-th coordinate in* $\boldsymbol{r}$ *(the row marginal constraint), and* $\boldsymbol{y} \in \mathbb{R}^n$*.*

Lemma 6 *For any* $\boldsymbol{y}$*,*

$$A(P, \boldsymbol{y}) \quad = \quad D_\phi(\tilde{P} \| P) - D_\phi(\tilde{P} \| \boldsymbol{y} \diamond_q P) \ . \tag{48}$$

Furthermore, $A(P, \mathbf{0}) = 0$*.*

Proof We have

$$\begin{aligned} & D_\phi(\tilde{P} \| P) - D_\phi(\tilde{P} \| \boldsymbol{y} \diamond_q P) \\ = \quad & -D_\phi(P \| \boldsymbol{y} \diamond_q P) \\ & + \langle \tilde{P} - P, \nabla\phi(\boldsymbol{y} \diamond_q P) - \nabla\phi(P) \rangle \ . \end{aligned}$$

■

Because a Bregman divergence is non-negative and $A(P, \mathbf{0}) = 0$, if, as long as there exists some $\boldsymbol{y}$ for which $A(P, \boldsymbol{y}) > 0$ we keep on updating P by replacing it by $\boldsymbol{y}^* \diamond_q P$ such that $A(P, \boldsymbol{y}^*) > 0$, then the sequence

$$P_0 = \tilde{U} \to P_1 \doteq \boldsymbol{y}_0^* \diamond_q P_0 \to P_2 \doteq \boldsymbol{y}_1^* \diamond_q P_1 \cdots \tag{49}$$

will converge to a limit matrix in the sequence,

$$\lim_j P_j \doteq \boldsymbol{y}_{j-1}^* \diamond_q P_{j-1} \ . \tag{50}$$

This matrix turns out to be the one we seek.

Theorem 7 *Let* $P_{j+1} \doteq \boldsymbol{y}_j \diamond_q P_j$ *(with* $P_0 \doteq \tilde{U}$*) be such that* $A(P_j, \boldsymbol{y}_j) > 0, \forall j \geq 0$*, and the sequence ends when no such* $\boldsymbol{y}_j$ *exists. Then* $\mathcal{S} \doteq \{P_j\}_{j \geq 0} \subset \bar{\mathcal{Q}}$*. If furthermore* $\mathcal{S}$ *lies in a compact of* $\bar{\mathcal{Q}}$*, then it satisfies*

$$P^\star \doteq \lim_j P_j \quad = \quad \arg\min_{P \in \mathcal{P}} D_q(P \| \tilde{U}) \ . \tag{51}$$

Proof sketch: The proof relies on two steps, first that

$$P^\star \doteq \lim_j P_j \quad = \quad \arg\min_{P \in \mathcal{P}} D_\phi(P \| \tilde{U}) \ , \tag{52}$$

and then the fact that (51) holds as well, which "amounts" to replacing D_ϕ, which is Bregman, by D_q, which is *not*. Because it is standard in Bregman divergences, we sketch the first step. The fundamental result we use is adapted from [5] (see also [3, Theorem 1]).

Theorem 8 *Suppose that $D_\phi(\tilde{P}, \tilde{U}) < \infty$. Then there exists a unique $P^\star$ satisfying the following four properties:*

1. $P^\star \in \mathcal{P} \cap \bar{\mathcal{Q}}$
2. $\forall P \in \mathcal{P}, \forall R \in \bar{\mathcal{Q}}, D_\phi(P\|R) = D_\phi(P\|P^\star) + D_\phi(P^\star\|R)$
3. $P^\star = \arg\min_{P\in\mathcal{P}} D_\phi(P\|\tilde{U})$
4. $P^\star = \arg\min_{R\in\bar{\mathcal{Q}}} D_\phi(\tilde{P}\|R)$

Moreover, any of these four properties determines $P^\star$ uniquely.

It is not hard to check that $\tilde{U} \in \bar{\mathcal{Q}}$ and whenever $P_j \in \bar{\mathcal{Q}}$, then $\boldsymbol{y} \diamond_q P_j \in \bar{\mathcal{Q}}, \forall \boldsymbol{y}$, so we indeed have $\mathcal{S} \subset \bar{\mathcal{Q}}$. With the constraint that $A(P_j, \boldsymbol{y}_j) > 0, \forall j \geq 0$, it follows from Lemma 6 that $A(P, \boldsymbol{y})$ is an auxiliary function for $\mathcal{S}$ [3] *if* we can show in addition that if $\boldsymbol{y} = \mathbf{0}$ is a maximum of $A(P, \boldsymbol{y})$, then $P \in \mathcal{P}$. To remark that this is true, we have

$$\nabla A(P, \boldsymbol{y})_{\boldsymbol{y}} = \boldsymbol{r} - P\mathbf{1} \ , \tag{53}$$

so whenever $A(P, \boldsymbol{y})$ reaches a maximum in $\boldsymbol{y}$, we indeed have $P\mathbf{1} = \boldsymbol{r}$ and so $P \in \mathcal{P}$, and if $\boldsymbol{y} = \mathbf{0}$ then because a Bregman divergence satisfies the identity of the indiscernibles, if $\boldsymbol{y} = \mathbf{0}$ is the maximum, then $\mathcal{S}$ has converged to some $P^\star$. From 4. above, we get

$$P^\star = \arg\min_{R\in\bar{\mathcal{Q}}} D_\phi(\tilde{P}\|R) \ , \tag{54}$$

and so from 3. above, we also get

$$P^\star = \arg\min_{P\in\mathcal{P}} D_\phi(P\|\tilde{U}) \ . \tag{55}$$

To "transfer" this result to D_q, we just need to remark that there is one remarkable trivial equality:

$$D_\phi(P\|R) = D_q(P\|R) - \sum_{i,j}(p_{ij}^q - r_{ij}^q) \ , \tag{56}$$

so that even when $K_{1/q}$ is *not* a Bregman divergence for a general q, it still meets the *Bregman triangle equality* [1].

Lemma 9 *We have;*

$$\begin{aligned} & D_q(P\|R) + D_q(R\|S) - D_q(P\|S) \\ & = D_\phi(P\|R) + D_\phi(R\|S) - D_\phi(P\|S) \\ & = \langle P - R, \nabla\phi(S) - \nabla\phi(R)\rangle \ . \end{aligned} \tag{57}$$

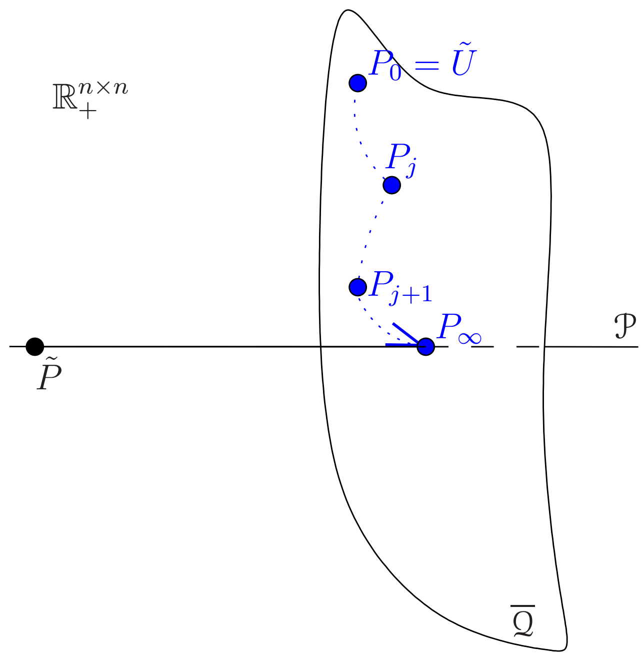

Figure 1: High level overview of the proof of Theorem 7 (see text for details).

Hence, point 2. implies as well

$$D_q(P\|R) = D_q(P\|P^\star) + D_q(P^\star\|R) \ , \tag{58}$$

$\forall P \in \mathcal{P}, \forall R \in \bar{\mathcal{Q}}$, and so $D_q(P\|\tilde{U}) = D_q(P\|P^\star) + D_q(P^\star\|\tilde{U}), \forall P \in \mathcal{P}$, so that we also have (since D_q is non negative and satisfies $D_q(P\|P) = 0$)

$$P^\star = \arg\min_{P\in\mathcal{P}} D_q(P\|\tilde{U}) \ ,$$

as claimed (end of the proof of Theorem 7). ☐

Figure 1 summarizes Theorem 7. We are left with the problem of finding an auxiliary function for the sequence $\mathcal{S}$, which we recall boils down to finding, whenever it exists, some $\boldsymbol{y}$ such that $A(P,\boldsymbol{y}) > 0$.

Theorem 10 *$A(P,\boldsymbol{y})$ is an auxiliary function for $\mathcal{S}$ for the sequence of updates $\boldsymbol{y}$ given as in steps 6-11 of* SO–TROT *(Algorithm 1).*

Proof We shall need the complete Taylor expansion of $A(P,\boldsymbol{y})$.

Lemma 11 *Let us denote for short $\gamma \doteq 1-q$. The Taylor series expansion of $A_i(P,\boldsymbol{y})$*

(as defined in Definition 5) is:

$$\begin{aligned}
&A_i(P, \boldsymbol{y}) \\
&= \quad y_i(r_i - \sum_j p_{ij}) \\
&\qquad -\sum_j p_{ij} \sum_{k=2}^{\infty} \left[\frac{1}{k}\prod_{l=1}^{k-1}(\gamma + q/l)\right] y_i^k \left(\frac{p_{ij}^{\gamma}}{q}\right)^{k-1} \quad . \qquad (59)
\end{aligned}$$

Proof Let us denote $f(x) = \exp_q^{-q}(x)$. We have:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}x} f(x) &= q \exp_q^{1-q}(x) \frac{\mathrm{d}}{\mathrm{d}x} \exp_q^{-1}(x) \\
&= -q \exp_q^{-1}(x) \quad . \qquad (60)
\end{aligned}$$

A simple recursion also shows ($\forall k \geq 2$):

$$\begin{aligned}
&\frac{\mathrm{d}^k}{\mathrm{d}x^k} \exp_q^{-1}(x) \\
&= \quad (-1)^k \left[\prod_{i=1}^{k}(i - (i-1)q)\right] \exp_q^{kq-(k+1)}(x) \quad ,
\end{aligned}$$

which yields $\forall k \geq 1$,

$$\begin{aligned}
\frac{\mathrm{d}^k}{\mathrm{d}x^k} f(x) &= -q \frac{\mathrm{d}^{k-1}}{\mathrm{d}x^{k-1}} \exp_q^{-1}(x) \\
&= (-1)^k q \left[\prod_{i=1}^{k-1}(i\gamma + q)\right] \exp_q^{-(k-1)\gamma - 1}(x) \quad .
\end{aligned}$$

Since $\exp_q^q(-1) = \exp_q(-1)/q$ and $\forall i, j, p_{ij} = \exp_q(-1)\exp_q^{-1}(a_{ij})$, writing the Taylor development of f at point a_{ij} evaluated at y_i, and adding the $y_i r_i + \sum_j p_{ij}^q$ term, we obtain the desired result. ■

We have two special reals to define, t_i and z_i. If $r_i \leq \sum_j p_{ij}$, we let t_i denote the maximum of the second order approximation of $A_i(P, \boldsymbol{y})$,

$$T_i^{(2)}(y_i) \quad \doteq \quad y_i(r_i - \sum_j p_{ij}) - \frac{y_i^2}{2}\sum_j \frac{p_{ij}^{1+\gamma}}{q} \quad , \qquad (61)$$

i.e. the root of

$$\frac{\mathrm{d}}{\mathrm{d}y} T^{(2)}(y_i) \quad = \quad (r_i - \sum_j p_{ij}) - y_i \sum_j \frac{p_{ij}^{1+\gamma}}{q} \quad .$$

If $\sum_j p_{ij} \leq r_i$, we let z_i be the the largest root of

$$\begin{aligned} R_i &\doteq (r_i - \sum_j p_{ij}) \\ &\quad -y_i \sum_j \frac{p_{ij}^{1+\gamma}}{q} - y_i^2(2-q)\sum_j \frac{p_{ij}^{1+2\gamma}}{q^2} \ . \end{aligned} \tag{62}$$

We shall see that z_i is positive. Let $y_i^* \doteq t_i$ if $r_i \leq \sum_j p_{ij}$, and $y_i^* \doteq z_i$ otherwise. We first make the assumption that

$$\left|\frac{y_i^* p_{ij}^{\gamma}}{q} \cdot \left(\gamma + \frac{q}{3}\right)\right| \leq \frac{1}{2} \ , \forall i, j \ . \tag{63}$$

Under this assumption, we have two cases.
($\star$) Case $r_i \leq \sum_j p_{ij}$. By definition, we have in this case that $y_i = t_i \leq 0$ in SO–TROT (Step 10). We also have

$$\begin{aligned} & A_i(P, \boldsymbol{y}) \\ &= T^{(2)}(y_i) \\ &\quad \underbrace{-\sum_j p_{ij} \sum_{k=3}^{\infty} \left[\frac{1}{k}\prod_{l=1}^{k-1}(\gamma + q/l)\right] y_i^k \left(\frac{p_{ij}^{\gamma}}{q}\right)^{k-1}}_{\doteq S_3} \ . \end{aligned} \tag{64}$$

Since $y_i = t_i \leq 0$, S_3 is an alternating series, that is a series whose general term is alternatively positive and negative. Under assumption (63), the module of its general term is decreasing. A classic result on series allows us to deduce from this fact that (a) $S_3 \ll \infty$ and (b) the sign of S_3 is that of its first term, *i.e.*, it is negative. Since $A_i(P, \boldsymbol{y}) = T^{(2)}(y_i) - S_3$, we have that

$$A_i(P, \boldsymbol{y}) \geq T^{(2)}(y_i) = 0 \ . \tag{65}$$

Note also that $A_i(P, \boldsymbol{y}) = 0$ iff $\sum_j p_{ij} = r_i$ as $T^{(2)}(y_i)$ is decreasing on $[t_i, 0]$ and $T^{(2)}(0) = 0$. Hence, for the choice in Step 10, $A_i(P, \boldsymbol{y})$ is an auxiliary function for variable i.

($\star$) Case $\sum_j p_{ij} \leq r_i$: we still have $A_i(P, \boldsymbol{y}) = T^{(2)}(y_i) - S_3$, but this time y_i will be positive, ensuring $y_i(r_i - \sum_j p_{ij}) \geq 0$. We first show that S_3 is upperbounded by a geometric series under assumption (63):

$$
\begin{aligned}
& S_3 \\
&= \sum_j p_{ij} y_i^3 \left(\frac{p_{ij}^\gamma}{q}\right)^2 \sum_{k=0}^{\infty} \frac{y_i^k}{k+3} \left[\prod_{l=1}^{k+2} (\gamma + q/l)\right] \left(\frac{p_{ij}^\gamma}{q}\right)^k \\
&\leq \sum_j p_{ij}(1-q/2) \frac{y_i^3}{3} \left(\frac{p_{ij}^\gamma}{q}\right)^2 \sum_{k=0}^{\infty} \left(\frac{y_i p_{ij}^\gamma}{q}(\gamma + q/3)\right)^k \\
&= \sum_j p_{ij}(1-q/2) \frac{y_i^3}{3} \left(\frac{p_{ij}^\gamma}{q}\right)^2 \times \frac{1}{1 - \frac{y_i p_{ij}^\gamma}{q}(\gamma + q/3)} \\
&\leq (2-q) \sum_j p_{ij} \frac{y_i^3}{3} \left(\frac{p_{ij}^\gamma}{q}\right)^2 \quad ,
\end{aligned}
$$

which conveniently yields

$$
A_i(P, \boldsymbol{y}) \quad \geq \quad T^{(2)}(y_i) - (2-q) \sum_j p_{ij} \frac{y_i^3}{3} \left(\frac{p_{ij}^\gamma}{q}\right)^2 \quad . \tag{66}
$$

The derivative of the right-hand term of (66) is R_i defined in eq. (62) above. Let us define:

$$
a \quad \doteq \quad (2-q) \sum_j \frac{p_{ij}^{1+2\gamma}}{q^2} \quad , \tag{67}
$$

$$
b \quad \doteq \quad \sum_j \frac{p_{ij}^{1+\gamma}}{q} \quad , \tag{68}
$$

$$
c \quad \doteq \quad -(r_i - \sum_j p_{ij}) \quad . \tag{69}
$$

We have $ac < 0$ and consequently the discriminant $\Delta \doteq b^2 - 4ac > b^2$, implying R_i has a positive root $z_i \doteq (-b + \sqrt{\Delta})/(2a)$ which maximises the right-hand term of 66, and is such that this right-hand term is positive. Further, we again have that $z_i = 0$ iff $\sum_j p_{ij} = r_i$. It is easy to check that $z_i = y_i$ in Step 8 of SO–TROT, for which we check that $A_i(P, \boldsymbol{y}) \geq 0$, wich equality iff $\sum_j p_{ij} = r_i$. Hence, for the choice in Step 8, $A_i(P, \boldsymbol{y})$ is an auxiliary function for variable i.

We can now conclude that under assumption (63), $A(P, \boldsymbol{y})$ is an auxiliary function.

If assumption (63) does not hold, then notice that this cannot not hold at convergence for coordinate i. For this reason, $r_i \neq \sum_j p_{ij}$ and the sign $\mathrm{sign}(r_i - \sum_j p_{ij})$ is also well defined. Therefore, we just need to pick a value for $y_i \neq 0$ which guarantees

$A_i(P, \boldsymbol{y}) > 0$. To do so, we pick

$$y_i \quad = \quad \frac{q \cdot \mathrm{sign}(r_i - \sum_j p_{ij})}{(6-4q) \cdot \max_j p_{ij}^{1-q}} \ , \tag{70}$$

remarking that this y_i indeed violates (63) (recalling $\gamma \doteq 1 - q$). We also have $|y_i| \in (0, n^{2(1-q)}/2]$. Notice that this choice guarantees $A_i(P, \boldsymbol{y}) > 0$. (end of the proof of Theorem 10) ■

Theorems 7 and 10 altogether prove Theorem 7.

Supplementary Material: experiments

6 Per county error distribution, TROT survey vs Florida average

Figure 2 displays the empirical distribution of the errors for TROT vs Florida average. While not being a true distribution of the solution error of TROT — in a Bayesian sense —, the graph should convey the intuition that algorithms with a distribution that shrinks around zero provide better inference.

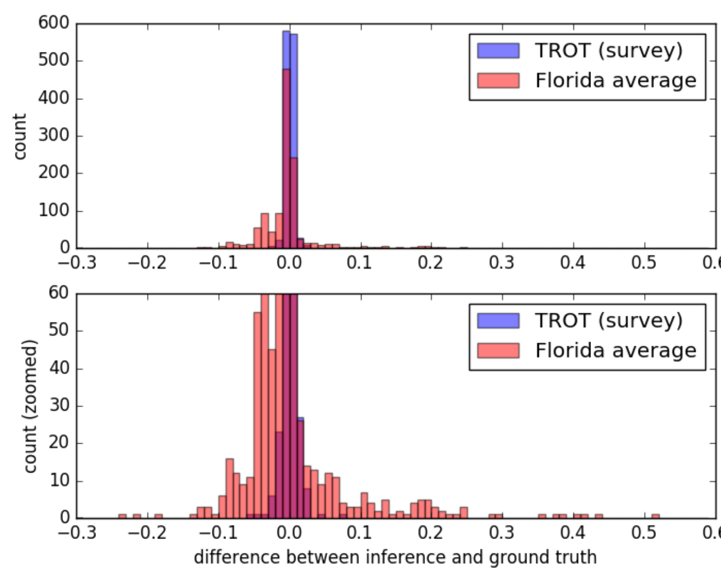

Figure 2: (Signed) error distribution of TROT compared to Florida-average.

7 Per county errors, TROT survey vs TROT $\mathbf{11}^\top$

Figure 3 confronts the prediction errors by county of TROT when we use $M = M^{\text{sur}}$ (survey) and $M = M^{\text{no}} (= \mathbf{11}^\top)$ as cost matrix: while the overall performance of the two algorithms is very close, the graph demonstrates that TROT optimized with M^{sur} achieves very often smaller error, although the average error is worsen by few particularly bad counties.

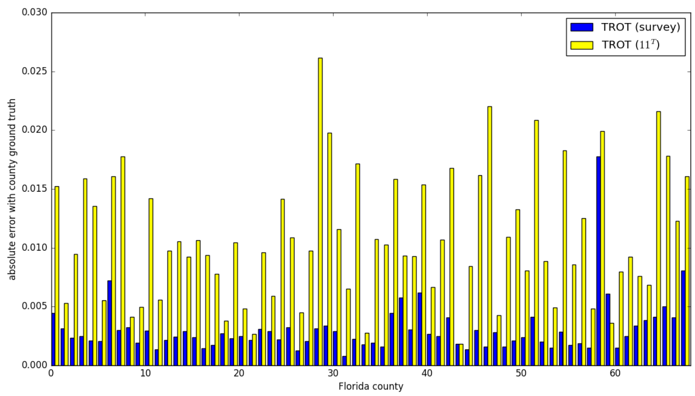

Figure 3: Absolute error of TROT optimized with M compared to with no prior.

References

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

[2] L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Comp. Math. and Math. Phys.*, 7:200–217, 1967.

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

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

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

[6] S. Furuichi. Information theoretical properties of tsallis entropies. *Journal of Mathematical Physics*, 47(2), 2006.

[7] R. Nock and F. Nielsen. On the efficient minimization of classification-calibrated surrogates. In *NIPS*21*, pages 1201–1208, 2008.