Optimal Transport with Tempered Exponential Measures

Ehsan Amid[1]*, Frank Nielsen[2], Richard Nock[3], Manfred K. Warmuth[3]

[1] Google DeepMind
[2] Sony Computer Science Laboratories Inc.
[3] Google Research
eamid@google.com, frank.nielsen@acm.org, {richardnock, manfred}@google.com

Abstract

In the field of optimal transport, two prominent subfields face each other: (i) unregularized optimal transport, "à-la-Kantorovich", which leads to extremely sparse plans but with algorithms that scale poorly, and (ii) entropic-regularized optimal transport, "à-la-Sinkhorn-Cuturi", which gets near-linear approximation algorithms but leads to maximally un-sparse plans. In this paper, we show that an extension of the latter to tempered exponential measures, a generalization of exponential families with indirect measure normalization, gets to a very convenient middle ground, with both very fast approximation algorithms and sparsity, which is under control up to sparsity patterns. In addition, our formulation fits naturally in the unbalanced optimal transport problem setting.

Introduction

Most loss functions used in machine learning (ML) can be related, directly or indirectly, to a comparison of positive measures (in general, probability distributions). Historically, two broad families of distortions were mainly used: f-divergences (Ali and Silvey 1966; Csiszár 1963) and Bregman divergences (Bregman 1967). Among other properties, the former are appealing because they encapsulate the notion of monotonicity of information (Amari 2016), while the latter are convenient because they axiomatize the expectation as a maximum likelihood estimator (Banerjee et al. 2004). Those properties, however, put constraints on the distributions, either on their support for the former or their analytical form for the latter.

A third class of distortion measures has progressively emerged later on, alleviating those constraints and with the appealing property to meet distance axioms: Optimal Transport distances (Peyré and Cuturi 2019; Villani 2009). Those can be interesting in wide ML fields (Peyré and Cuturi 2019), but they suffer from poor scalability. A trick of balancing the metric cost with an entropic regularizer (Cuturi 2013) substantially improves scalability to near-optimality but blurs the frontiers with other distortion measures (Cuturi 2013; Muzellec et al. 2017). Most importantly, the structure of the unregularized OT plan is substantially altered through regularization: its sparsity is reduced by a factor $\Omega(n)$, n being the dimension of the marginals (we consider discrete optimal transport). At the expense of an increase in complexity, getting back to a user-constrained sparse solution can be done by switching to a quadratic regularizer (Liu, Puigcerver, and Blondel 2023), but loses an appealing structure of the entropic-regularized OT (EOT) solution, a discrete exponential family with very specific features. Sparsity is an important topic in optimal transport: both unregularized and EOT plans are extremal in the sparsity scale, which does not necessarily fit in observed patterns (Peyré and Cuturi 2019).

Finally and most importantly, optimal transport, regularized or not, does not require normalized measures; in fact, it can be extended to the unbalanced problem where marginals' total masses do not even match (Janati et al. 2020). In that last, very general case, the problem is usually cast with approximate marginal constraints and without any constraint whatsoever on the transport plan's total mass.

In this context, our paper introduces OT on tempered exponential measures (TEMs, a generalization of exponential families), with a generalization of the EOT. Notable structural properties of the problem include training as fast as Sinkhorn balancing *and* with guarantees on the solution's sparsity, also including the possibility of unbalanced optimal transport *but* with tight control over total masses via their *co-densities*, distributions that are used to indirectly normalize TEMs (see Figure 1). We characterize sparsity up to sparsity patterns in the optimal solution and show that sparsity with TEMs can be interpreted as balancing the classical OT cost with an interaction term interpretable in the popular gravity model for spatial interactions (Haynes and Fotheringham 1984). Interestingly, this interpretation cannot hold anymore for the particular case of exponential families and thus breaks for EOT.

To maximize readability, all proofs are deferred to an appendix.[1]

Definitions

Optimal Transport in the Simplex

In classical discrete optimal transport (OT), we are given a cost matrix $\mathrm{M} \in \mathbb{R}^{n\times n}$ ($n > 1$) and two probability vectors $\boldsymbol{r}$ and column $\boldsymbol{c}$ in the simplex $\Delta_n \doteq \{\boldsymbol{p} \in \mathbb{R}^n :$

*Authors listed alphabetically.

[1] The full article is available at https://arxiv.org/abs/2309.04015.

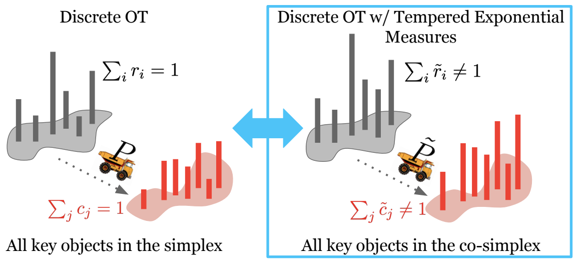


Figure 1: In classical optimal transport (OT, left), regularized or not, marginals and the OT plan sought are in the probability simplex; the optimal solution solely depends on the metric properties of the supports. Entropic regularization balances the metric cost with an entropic cost, and the optimal solution has remarkable properties related to exponential families. In this paper (right), we lift the whole setting to families of measures generalizing exponential families: tempered exponential measures (TEMs). Specific properties that appear include unbalancedness and sparsity of the optimal solution (see text).

$\boldsymbol{p} \geqslant \mathbf{0} \wedge \mathbf{1}^\top \boldsymbol{p} = 1\}$. Usually, $\mathbf{M}$ satisfies the axioms of a distance, though only non-negativity is really important for all the results that we state. The OT problem seeks to find $d_{\mathbf{M}}(\boldsymbol{r}, \boldsymbol{c}) \doteq \min_{\mathbf{P} \in U_n(\boldsymbol{r},\boldsymbol{c})} \langle \mathbf{P}, \mathbf{M} \rangle$, where $U_n(\boldsymbol{r}, \boldsymbol{c}) \doteq \{\mathbf{P} \in \mathbb{R}_+^{n \times n} | \, \mathbf{P}\mathbf{1}_n = \boldsymbol{r}, \, \mathbf{P}^\top \mathbf{1}_n = \boldsymbol{c}\}$ is the *transport polytope* and $\langle \cdot, \cdot \rangle$ stands for the Frobenius dot-product. In the *entropic-regularized OT* problem (Cuturi 2013), we rather seek

$$d_{\mathbf{M}}^{\lambda}(\boldsymbol{r}, \boldsymbol{c}) \doteq \min_{\mathbf{P} \in U_n(\boldsymbol{r},\boldsymbol{c})} \langle \mathbf{P}, \mathbf{M} \rangle + \frac{1}{\lambda} \cdot \langle \mathbf{P}, \log \mathbf{P} \rangle \, , \lambda > 0 \, . \quad (1)$$

Any discrete distribution is an exponential family (Amari 2016) but the OT plan solution to (1), say $\mathbf{P}^*$, has a special form. Denote $\boldsymbol{\mu}, \boldsymbol{\xi} \in \mathbb{R}^n$ the vectors of dual variables, corresponding to the row and column (respectively) marginalization constraints in (1). The support of $\mathbf{P}^*$ is $[n]^2$, where $[n] \doteq \{1, 2, ..., n\}$. We need to express $\mathbf{P}^*$ as

$$P_{ij}^* = \exp(\langle \mathbf{\Theta}, \mathbf{E}_{ij} \rangle - G(\mathbf{\Theta})), \quad (2)$$

with $\mathbf{E}_{ij}$ the matrix with general entry $(\delta_{ik}\delta_{jl})_{kl}$ ("δ" being Kronecker symbol). Let $\mathbf{S} \doteq \mathbf{M} - \boldsymbol{\mu}\mathbf{1}^\top - \mathbf{1}\boldsymbol{\xi}^\top$, which is a strictly positive matrix. In the unregularized case, $\mathbf{S}$ encodes the slack of the constraints over the dual variables (Peyré and Cuturi 2019, Section 2.5). It follows from Cuturi (2013) that the natural parameter of the exponential family is defined from those slack variables:

$$\mathbf{\Theta} = -\lambda \cdot \mathbf{S},$$

while the cumulant or log-partition function G in (2) is, in fact, 0 because normalization is implicitly ensured in $U_n(\boldsymbol{r}, \boldsymbol{c})$ (otherwise, the cumulant would depend on the Lagrange multiplier of the normalization constraint).

Tempered Exponential Measures

Any exponential family is a probability distribution that maximizes Shannon's entropy subject to a constraint on its expectation (Amari 2016). A *tempered exponential measure* (TEM) adopts a similar axiomatization *but* via a generalization of Shannon's entropy (Tsallis entropy) and normalization imposed not on the TEM itself but on a so-called *co-distribution* (Amid, Nock, and Warmuth 2023). This last constraint is a fundamental difference from previous generalizations of exponential families, q-exponential families, and deformed exponential families (Amari 2016). Compared to those, TEMs also have the analytical advantage of getting a closed-form solution for the cumulant, a key ML function. A TEM has the general form (with $[z]_+ \doteq \max\{0, z\}$):

$$\tilde{p}(\boldsymbol{x}) \doteq \frac{\exp_t(\langle \boldsymbol{\theta}, \boldsymbol{\varphi}(\boldsymbol{x}) \rangle)}{\exp_t(G_t(\boldsymbol{\theta}))}, \quad \exp_t(z) \doteq [1 + (1-t)z]_+^{\frac{1}{1-t}} \, ,$$

where G_t is the cumulant and $\boldsymbol{\theta}$ denotes the natural parameter. The inverse of $\exp_t$ is $\log_t(z) \doteq \left(z^{1-t} - 1\right)/(1-t)$, both being continuous generalizations of $\exp$ and $\log$ for $t = 1$. Both functions keep their $t = 1$ convexity/concavity properties for $t \geqslant 0$. The tilde notation above p indicates that normalization does not occur on the TEM, but on a co-density defined as

$$p \doteq \tilde{p}^{2-t} \quad \left(= \tilde{p}^{1/t^*}, \text{ with } t^* \doteq 1/(2-t) \right) . \quad (3)$$

Remark 1. *For a given vector $\tilde{\mathbf{p}}$ (or a matrix $\tilde{\mathbf{P}}$) with the tilde notation, whenever convenient, we will use the convention $\boldsymbol{p} = \tilde{\boldsymbol{p}}^{1/t^*}$ (correspondingly, $\mathbf{P} = \tilde{\mathbf{P}}^{1/t^*}$, the exponent being coordinate-wise) whenever the tilde sign is removed.*

Hence, a TEM satisfies the indirect normalization $\int \tilde{p}^{2-t} \mathrm{d}\xi = \int p \mathrm{d}\xi = 1$.

Remark 2. *In this paper, we assume $t \in [0, 1]$, though some of our results are valid for a broader range (discussed in context).*

In the same way, as KL divergence is the canonical divergence for exponential families (Amari and Nagaoka 2000), the same happens for a generalization in TEMs. Given two non-negative vectors $\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}} \in \mathbb{R}^m$, we define the generalized tempered relative entropy as (Amid et al. 2019)

$$D_t(\tilde{\boldsymbol{u}} \| \tilde{\boldsymbol{v}}) \doteq \sum_{i \in [n]} \tilde{u}_i \left(\log_t \tilde{u}_i - \log_t \tilde{v}_i \right) - \log_{t-1} \tilde{u}_i + \log_{t-1} \tilde{v}_i .$$

Just like the KL divergence ($t \to 1$), the tempered relative entropy is a Bregman divergence, induced by the generator $\varphi_t(z) \doteq z \log_t z - \log_{t-1}(z)$, which is convex for $t \in \mathbb{R}$. We also have $\varphi_t'(z) = \log_t(z)$. We define the following extension of the probability simplex Δ_n in $\mathbb{R}^n$.

Definition 1. *The co-simplex of $\mathbb{R}^n$, $\tilde{\Delta}_n$ is defined as $\tilde{\Delta}_n \doteq \{\tilde{\boldsymbol{p}} \in \mathbb{R}^n : \tilde{\boldsymbol{p}} \geqslant \mathbf{0} \wedge \mathbf{1}^\top \tilde{\boldsymbol{p}}^{1/t^*} = 1\}$.*

Note that $\tilde{\boldsymbol{p}}^{1/t^*} \doteq \boldsymbol{p} \in \Delta_n$ iff $\tilde{\boldsymbol{p}} \in \tilde{\Delta}_n$ and $\tilde{\Delta}_n \to \Delta_n$ when $t \to 1$. Similarly, given $\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{c}} \in \tilde{\Delta}_n$, we define their corresponding co-polytope in $\mathbb{R}_+^{n \times n}$.

Definition 2. *The co-polyhedral set of $n \times n$ non-negative matrices with co-marginals $\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{c}} \in \tilde{\Delta}_n$ is defined as $\tilde{U}_n(\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{c}}) \doteq \{\tilde{\mathbf{P}} \in \mathbb{R}_+^{n \times n} | \, \tilde{\mathbf{P}}^{1/t^*}\mathbf{1} = \tilde{\boldsymbol{r}}^{1/t^*}, \, \tilde{\mathbf{P}}^{1/t^* \top}\mathbf{1} = \tilde{\boldsymbol{c}}^{1/t^*}\}$.*

Likewise, $\tilde{U}_n(\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{c}}) \rightarrow U_n(\boldsymbol{r}, \boldsymbol{c})$ (the transport polytope) in the limit $t \rightarrow 1$. More importantly, using our notation convention,

$$\mathbf{P} \in U_n(\boldsymbol{r}, \boldsymbol{c}) \text{ iff } \tilde{\mathbf{P}} \in \tilde{U}_n(\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{c}}).$$

Related Work

From an ML standpoint, there are two key components to optimal transport (OT): the problem structure and its solving algorithms. While historically focused on the former (Monge 1781; Kantorovich 1958), the field then became substantially "algorithm-aware", indirectly first via linear programming (Dantzig 1949) and then specifically because of its wide applicability in ML (Cuturi 2013). The entropic-regularized OT (EOT) mixes metric and entropic terms in the cost function but can also be viewed as an approximation of OT in a Kullback-Leibler ball centered at the independence plan, which is, in fact, a metric (Cuturi 2013). The resolution of the EOT problem can be obtained via Sinkhorn's algorithm (Sinkhorn and Knopp 1967; Franklin and Lorenz 1989; Knight 2008) (see Algorithm 1), which corresponds to iterative Bregman projections onto the affine constraint sets (one for the rows and another for the columns). The algorithm requires matrix-vector multiplication and can be easily implemented in a few lines of code, making it ideal for a wide range of ML applications. However, alternative implementations of the algorithm via the dual formulation prove to be more numerically stable and better suited for high-dimensional settings (Peyré and Cuturi 2019).

The structure of the solution – the transportation plan – is also important, and some features have become prominent in ML, like the sparsity of the solution (Liu, Puigcerver, and Blondel 2023). Sinkhorn iteration can be fine-tuned to lead to near-optimal complexity (Altschuler, Weed, and Rigollet 2017), but entropic regularization suffers a substantial structural downside: the solution is maximally un-sparse, which contrasts with the sparsity of the unregularized solution (Peyré and Cuturi 2019). Sparsity is a modern instantiation of ML's early constraint on the model's simplicity, otherwise known as Ockham's razor (Blumer et al. 1987).[2] It has been known for a long time that "extreme" simplicity constraints lead to intractability for linear programming (Karp 1972), so sparsity in OT is desirable – and not just for the sake of Ockham's razor (Blondel, Seguy, and Rolet 2018) – but it is non-trivial and various notions of tractable sparsity can be sought, from a general objective (Blondel, Seguy, and Rolet 2018; Muzellec et al. 2017) down to *ex-ante* node specifics like transport obstruction (Dessein, Papadakis, and Rouas 2018) or limiting the transport degree (Liu, Puigcerver, and Blondel 2023). Sparsity makes it convenient to train from general optimizers (Liu, Puigcerver, and Blondel 2023). This comes, however, at the expense of losing an appealing probabilistic structure of the EOT solution, a discrete exponential family with very specific features, and eventually loses as well the near-optimal algorithmic convenience that fine-tuning Sinkhorn offers for training (Altschuler, Weed, and Rigollet 2017).

Taking EOT as a starting point, two different directions can be sought for generalization. The first consists in replacing the entropic term with a more general one, such as Tsallis entropy (still on the simplex), which was introduced in Muzellec et al. (2017); the second consists in alleviating the condition of identical marginal masses, which is touched upon in Janati et al. (2020) and was initially proposed without regularization by Benamou (2003).

In work predating the focus on optimal transport (Helmbold and Warmuth 2009), the same relative entropy regularized optimal transport problem was used to develop online algorithms for learning permutations that predict close to the best permutation chosen in hindsight. See a recent result in Ballu and Berthet (2023) on mirror Sinkhorn algorithms.

Beyond Sinkhorn Distances With TEMs

OT Costs With TEMs

Since tempered exponential measures involve two distinct sets (the probability simplex and the co-simplex), we can naturally define two unregularized OT objectives given a cost matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$. The first is the classical OT cost; we denote it as the *expected cost*,

$$d^t_{\mathbf{M}}(\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{c}}) \doteq \min_{\tilde{\mathbf{P}} \in \tilde{U}_n(\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{c}})} \langle \mathbf{P}, \mathbf{M} \rangle \quad (4)$$

(with our notations, note that the constraint is equivalent to $\mathbf{P} \in U_n(\boldsymbol{r}, \boldsymbol{c})$). Instead of embedding the cost matrix on the probability simplex, we can put it directly on the co-simplex, which leads to the *measured cost*:

$$\tilde{d}^t_{\mathbf{M}}(\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{c}}) \doteq \min_{\tilde{\mathbf{P}} \in \tilde{U}_n(\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{c}})} \langle \tilde{\mathbf{P}}, \mathbf{M} \rangle. \quad (5)$$

$d^t_{\mathbf{M}}$ is a distance if $\mathbf{M}$ is a metric matrix. $\tilde{d}^t_{\mathbf{M}}$ is trivially non-negative, symmetric, and meets the identity of indiscernibles. However, it seems to only satisfy a slightly different version of the triangle inequality, which converges to the triangle inequality as $t \rightarrow 1$.

Proposition 1. *If* $\mathbf{M}$ *is a distance matrix and* $t \leqslant 1$, $(\tilde{d}^t_{\mathbf{M}}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{z}}))^{2-t} \leqslant M^{1-t} \cdot \left(\tilde{d}^t_{\mathbf{M}}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) + \tilde{d}^t_{\mathbf{M}}(\tilde{\boldsymbol{y}}, \tilde{\boldsymbol{z}})\right)$, $\forall \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}, \tilde{\boldsymbol{z}} \in \tilde{\Delta}_n$, *where* $M \doteq \sum_{ij} M_{ij}$.

Factor M^{1-t} is somehow necessary to prevent vacuity of the inequality: scaling a cost matrix by a constant $\kappa > 0$ does not change the OT optimal plan, but scales the OT cost by κ; in this case, the LHS scales by κ^{2-t} and the RHS scales by $\kappa^{1-t} \cdot \kappa = \kappa^{2-t}$ as well. Note that it can be the case that $M < 1$ so the RHS can be smaller than the triangle inequality's counterpart – yet we would not necessarily get an inequality tighter than the triangle inequality because, in this case, it is easy to show that the LHS would also be smaller than the triangle inequality's counterpart.

OT Costs in a Ball

This problem is an intermediary that grounds a particular metric structure of EOT. This constrained problem seeks the

[2] *Numquam ponenda est pluralitas sine necessitate*, "plurality should never be imposed without necessity", William of Ockham, XIVth century.

optimal transport plan in an information ball – a KL ball – centered at the independence plan. Using our generalized tempered relative entropy, this set can be generalized as:

$$\tilde{U}_n^{\varepsilon}(\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{c}}) \doteq \{\tilde{\mathbf{P}} \in \tilde{U}_n(\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{c}}) | \, D_t\big(\tilde{\mathbf{P}} \| \tilde{\boldsymbol{r}}\tilde{\boldsymbol{c}}^\top\big) \leqslant \varepsilon)\}\,, \quad (6)$$

where ε is the radius of this ball. It turns out that when $t = 1$, minimizing the OT cost subject to being in this ball also yields a distance, called a Sinkhorn distance – if, of course, $\mathbf{M}$ is a metric matrix (Cuturi 2013). For a more general t, we can first remark that

$$D_t\big(\tilde{\mathbf{P}} \| \tilde{\boldsymbol{r}}\tilde{\boldsymbol{c}}^\top\big) \leqslant \frac{1}{1-t}, \forall \tilde{\mathbf{P}} \in \tilde{\Delta}_{n\times n}, \forall \tilde{\boldsymbol{r}}, \tilde{\boldsymbol{c}} \in \tilde{\Delta}_n,$$

so that we can consider that

$$\varepsilon \quad < \quad \frac{1}{1-t} \quad (7)$$

for the ball constraint not to be vacuous. $\tilde{\boldsymbol{r}}\tilde{\boldsymbol{c}}^\top \in \tilde{U}_n(\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{c}})$ is the *independence table* with co-marginals $\tilde{\boldsymbol{r}}$ and $\tilde{\boldsymbol{c}}$. When $\varepsilon \to \infty$, we have $\tilde{U}_n^{\varepsilon}(\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{c}}) \to \tilde{U}_n(\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{c}})$. When $t \to 1$, $\tilde{U}_n^{\varepsilon}(\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{c}}) \to U_n^{\varepsilon}(\boldsymbol{r}, \boldsymbol{c})$, the subset of the transport polytope with bounded KL divergence to the independence table (Cuturi 2013).

Notably, the generalization of the ball $\tilde{U}_n^{\varepsilon}(\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{c}})$ for $t \neq 1$ loses the convexity of the ball itself – while the divergence D_t remains convex. However, the domain keeps an important property: it is $1/t^*$*-power convex*.

Proposition 2. *For any* $\tilde{\mathbf{P}}, \tilde{\mathbf{Q}} \in \tilde{U}_n^{\varepsilon}(\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{c}})$ *and any* $t \in \mathbb{R}\text{-}\{2\}$,

$$(\beta\, \tilde{\mathbf{P}}^{1/t^*} + (1-\beta)\, \tilde{\mathbf{Q}}^{1/t^*})^{t^*} \in \tilde{U}_n^{\varepsilon}(\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{c}})\,, \forall \beta \in [0,1].$$

Regularized OT Costs

in the case of entropic regularization, the OT cost is replaced by $\langle \mathbf{P}, \mathbf{M} \rangle + (1/\lambda) \cdot D_1\big(\mathbf{P} \| \boldsymbol{r}\boldsymbol{c}^\top\big)$. In the case of TEMs, we can formulate two types of regularized OT costs generalizing this expression, the *regularized expected cost*

$$d_{\mathbf{M}}^{t,\lambda}(\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{c}}) \doteq \min_{\tilde{\mathbf{P}} \in \tilde{U}_n(\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{c}})} \langle \mathbf{P}, \mathbf{M} \rangle + \frac{1}{\lambda} \cdot D_t(\tilde{\mathbf{P}} \| \tilde{\boldsymbol{r}}\tilde{\boldsymbol{c}}^\top)\,, \quad (8)$$

for $\lambda > 0$ and the *regularized measured cost*

$$\tilde{d}_{\mathbf{M}}^{t,\lambda}(\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{c}}) \doteq \min_{\tilde{\mathbf{P}} \in \tilde{U}_n(\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{c}})} \langle \tilde{\mathbf{P}}, \mathbf{M} \rangle + \frac{1}{\lambda} \cdot D_t(\tilde{\mathbf{P}} \| \tilde{\boldsymbol{r}}\tilde{\boldsymbol{c}}^\top)\,. \quad (9)$$

The *raison d'être* of entropic regularization is the algorithmic efficiency of its approximation. As we shall see, this stands for TEMs as well. In the case of TEMs, the question remains: what is the structure of the regularized problem? In the case of EOT ($t = 1$), the answer is simple, as the OT costs in a ball bring the metric foundation of regularized OT, since the regularized cost is just the Lagrangian of the OT in a ball problem. Of course, the downside is that parameter λ in the regularized cost comes from a Lagrange multiplier, which is unknown in general, but at least a connection does exist with the metric structure of the unregularized OT problem – assuming, again, that $\mathbf{M}$ is a metric matrix.

As we highlight in Proposition 1, the measured cost only meets an approximate version of the triangle inequality, so a metric connection holds only in a weaker sense for $t \neq 1$. However, as we now show, when $t \neq 1$, there happens to be a *direct* connection with the unregularized OT costs themselves (expected and measured), the connection to which is blurred when $t = 1$ and sheds light on the algorithms we use.

Proposition 3. *For any TEM* $\tilde{\mathbf{P}} \in \tilde{\Delta}_{n\times n}$*, any* $t \in [0,1)$ *and* $0 \leqslant \varepsilon \leqslant 1/(1-t)$*, letting* $\mathbf{M}_t \doteq (\tilde{\boldsymbol{r}}\tilde{\boldsymbol{c}}^\top)^{1-t}$*, we have*

$$D_t\big(\tilde{\mathbf{P}} \| \tilde{\boldsymbol{r}}\tilde{\boldsymbol{c}}^\top\big) \leqslant \varepsilon \quad \Leftrightarrow \quad \langle \tilde{\mathbf{P}}, \mathbf{M}_t \rangle \geqslant \exp_t^{1-t}(\varepsilon). \quad (10)$$

The proof is immediate once we remark that on the co-simplex, the generalized tempered relative entropy simplifies for $t \neq 1$ as:

$$D_t\big(\tilde{\mathbf{P}} \| \tilde{\boldsymbol{r}}\tilde{\boldsymbol{c}}^\top\big) \quad = \quad \frac{1}{1-t} \cdot \Big(1 - \langle \tilde{\mathbf{P}}, \mathbf{M}_t \rangle\Big)\,. \quad (11)$$

Interestingly, this simplification does not happen for $t = 1$, a case for which we keep Shannon's entropy in the equation and thus get an expression not as "clean" as (11). Though $\mathbf{M}_t$ does not define a metric, it is useful to think of (10) as giving an equivalence of being in the generalized tempered relative entropy ball to the independence plan (a fact relevant to information theory) and having a large OT cost with respect to a cost matrix defined from the independence plan (a fact relevant to OT). For $t \neq 1$, the constrained OT problem becomes solving one OT problem subject to a constraint on another one.

Regularized OT Costs With TEMs Implies Sparsity

The regularized problem becomes even "cleaner" for the regularized measured cost (9) as it becomes an unregularized measured cost[3] (5) over a *fixed* cost matrix.

Proposition 4. *For any* $t \in [0,1)$*, the regularized measured cost* (9) *can be written as*

$$\lambda \cdot \tilde{d}_{\mathbf{M}}^{t,\lambda}(\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{c}}) \quad = \quad \frac{1}{1-t} + \min_{\tilde{\mathbf{P}} \in \tilde{U}_n(\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{c}})} \langle \tilde{\mathbf{P}}, \mathbf{M}' \rangle, \quad (12)$$

$$\mathbf{M}' \quad \doteq \quad \lambda \cdot \mathbf{M} - \frac{1}{1-t} \cdot \mathbf{M}_t, \quad (13)$$

with $\mathbf{M}_t$ *defined in Proposition 3.*

(Proof straightforward) This formulation shows that regularized OT with TEMs for $t \neq 1$ can achieve something that classical EOT ($t = 1$) cannot: getting sparse OT plans. Indeed, as the next theorem shows, specific sparsity patterns happen in any configuration of two sources $i \neq k$ and two destinations $l \neq j$ containing two distinct paths of negative costs and at least one of positive cost.

Theorem 1. *Let* $\tilde{\mathbf{P}}$ *be the optimal solution to the regularized measured cost* (9)*. Let* $\mathbf{S}$ *be the support indicator matrix of* $\tilde{\mathbf{P}}$*, defined by the general term* $S_{ij} = 1$ *if* $\tilde{P}_{ij} > 0$ *(and 0 otherwise). The following properties hold:*

1. *For any coordinates* (i,j)*, if* $M'_{ij} < 0$ *then* $S_{ij} = 1$*;*

[3] A similar discussion, albeit more involved, holds for the expected cost. We omit it due to the lack of space.

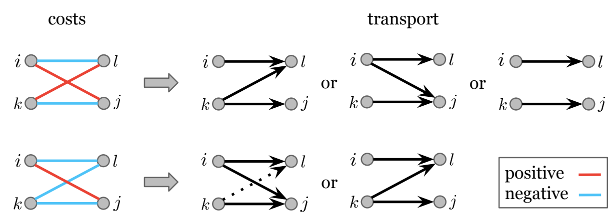


Figure 2: Illustration of Theorem 1. Negative costs of matrix $\mathbf{M}'$ in the regularized measured cost (13) are in blue, and those positive are in red. The sparsity results of the theorem are shown, where no arrow means no transport and a dashed arrow means a transport necessarily "small" (see text).

2. *For any coordinates $i \neq k$ and $j \neq l$, suppose we have the following configuration (Figure 2): $M'_{ij} > 0, M'_{il} < 0, M'_{kj} < 0$. Then we have the following*
   - *if $M'_{kl} > 0$, then $S_{ij} = 0$ or (non exclusive) $S_{kl} = 0$;*
   - *if $M'_{kl} < 0$ and $S_{ij} = 1$ then necessarily*

$$\tilde{P}_{kl}^{1-t} \leqslant \frac{|M'_{kl}|}{|M'_{ij}| + |M'_{il}| + |M'_{kj}|} \cdot \max\{\tilde{P}_{ij}, \tilde{P}_{il}, \tilde{P}_{kj}\}^{1-t}.$$

Theorem 1 is illustrated in Figure 2. Interpretations of the result in terms of transport follow: if we have a negative cost between i, l and j, k, then "some" transport necessarily happens in both directions; furthermore, if the cost between i, j is positive, then sparsity patterns happen:

- if the other cost k, l is also positive, then we do not transport between i, j or (non exclusive) k, l;
- if the other cost k, l is negative, then either we do not transport between i, j, or the transport between k, l is "small".

What is most interesting is that negative coordinates in $\mathbf{M}'$ are under tight control by the user, and they enforce non-zero transport, so the flexibility of the design of $\mathbf{M}'$, via tuning the strength of the regularization λ *and* the TEMs family via t, can allow to tightly design transport patterns.

Interpretation of Matrix $\mathbf{M}'$

Coordinate (i, j) is $M'_{ij} = \lambda M_{ij} - (\tilde{r}_i\tilde{c}_j)^{1-t}/(1-t)$. Quite remarkably, the term $(\tilde{r}_i\tilde{c}_j)^{1-t}/(1-t)$ happens to be equivalent, up to the exponent itself, to an interaction in the gravity model if distances are constant (Haynes and Fotheringham 1984). If the original cost $\mathbf{M}$ factors the distance, then we can get the full interaction term with the distance via its factorization. Hence, we can abstract any coordinate in $\mathbf{M}'$ as:

$$M'_{ij} \quad \propto \quad \text{original cost}(i, j) - \text{interaction}(i, j). \quad (14)$$

One would then expect that OT with TEMs reflects the mixture of both terms: having a large cost wrt interaction should not encourage transport, while having a large interaction wrt cost might just encourage transport. This is, in essence, the basis of Theorem 1.

Algorithm 1: $\texttt{Sinkhorn}(\mathbf{K}, \boldsymbol{r}, \boldsymbol{c})$

Input: $\mathbf{K} \in \mathbb{R}_+^{n\times n}$, $\boldsymbol{r}, \boldsymbol{c} \in \Delta_n$
Output: $\mathbf{P} \in U_n(\boldsymbol{r}, \boldsymbol{c})$

$\boldsymbol{\mu} \leftarrow \mathbf{1}_n,\ \boldsymbol{\xi} \leftarrow \mathbf{1}_n$
while not converged **do**
  $\boldsymbol{\mu} \leftarrow \boldsymbol{r}/(\mathbf{K}\boldsymbol{\xi})$
  $\boldsymbol{\xi} \leftarrow \boldsymbol{c}/(\mathbf{K}^\top\boldsymbol{\mu})$
end while
return $\operatorname{diag}(\boldsymbol{\mu})\,\mathbf{K}\,\operatorname{diag}(\boldsymbol{\xi})$

Algorithm 2: Regularized Optimal Transport with TEMs

Input: Cost matrix $\mathbf{M} \in \mathbb{R}_+^{n\times n}$, $\boldsymbol{r}, \boldsymbol{c} \in \Delta_n$, Regularizer λ
Output: $\tilde{\mathbf{P}} \in \tilde{U}_n(\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{c}})$

$\tilde{\mathbf{K}} \leftarrow \tilde{\mathbf{K}}_e$ or $\tilde{\mathbf{K}}_m$
$\mathbf{P} \leftarrow \texttt{Sinkhorn}(\mathbf{K}, \boldsymbol{r}, \boldsymbol{c}) \quad (\mathbf{K} = \tilde{\mathbf{K}}^{1/t^*})$
return $\mathbf{P}^{t^*}$

Algorithms for Regularized OT With TEMs

We show the analytic form of the solutions to (8) and (9) and explain how to solve the corresponding problems using an iterative procedure based on alternating Bregman projections. We then show the reduction of the iterative process to the Sinkhorn algorithm via a simple reparameterization.

Regularized Expected Cost

The following theorem characterizes the form of the solution, i.e., the transport plan for the regularized expected cost OT with TEMs.

Theorem 2. *A solution of* (8) *can be written in the form*

$$\tilde{P}_{ij} = \frac{\tilde{r}_i\tilde{c}_j}{\exp_t(\nu_i + \gamma_j + \lambda M_{ij})}, \quad (15)$$

where ν_i and γ_j are chosen s.t. $\tilde{\mathbf{P}} \in \tilde{U}_n(\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{c}})$.

Regularized Measured Cost

The solution of the regularized measured cost OT with TEMs is characterized next.

Theorem 3. *A solution of* (9) *can be written in the form*

$$\tilde{P}_{ij} = \exp_t\left((\log_t(\tilde{r}_i\tilde{c}_j) - \lambda M_{ij}) \ominus_t (\nu_i + \gamma_j)\right), \quad (16)$$

where $a \ominus_t b \doteq (a-b)/(1+(1-t)b)$ and ν_i and γ_j are chosen s.t. $\tilde{\mathbf{P}} \in \tilde{U}_n(\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{c}})$.

Approximation via Alternating Projections

The Lagrange multipliers ν_i and γ_j in the solutions (15) and (16) no longer act as separate scaling factors for the rows and the columns because $\exp_t(a+b) \neq \exp_t(a) \cdot \exp_t(b)$ (yet, an efficient approximation is possible, *Cf* below). Consequently, the solutions are not *diagonally equivalent* to their corresponding seed matrices (19) and (21). However, keeping just one marginal constraint leads to a solution that bears the analytical shape of Sinkhorn balancing.

Letting $\tilde{\mathbf{P}}_\circ \in \mathbb{R}_+^{n\times n}$, the *row projection* and *column projection* to $\tilde{U}_n(\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{c}})$ correspond to

$$\min_{\tilde{\mathbf{P}}:\, \mathbf{P}\mathbf{1}_n = \boldsymbol{r}} D_t(\tilde{\mathbf{P}} \| \tilde{\mathbf{P}}_\circ)\,, \qquad \text{(row projection)}$$

$$\min_{\tilde{\mathbf{P}}:\, \mathbf{P}^\top\mathbf{1}_n = \boldsymbol{c}} D_t(\tilde{\mathbf{P}} \| \tilde{\mathbf{P}}_\circ)\,, \qquad \text{(column projection)}$$

in which, we use the shorthand notation $\mathbf{P} = \tilde{\mathbf{P}}^{1/t^*}$ (similarly for $\boldsymbol{r}$ and $\boldsymbol{c}$).

Theorem 4. *Given $\tilde{\mathbf{P}}_\circ \in \mathbb{R}_+^{n\times n}$, the row and column projections to $\tilde{U}_n(\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{c}})$ can be performed respectively via*

$$\tilde{\mathbf{P}} = \operatorname{diag}\big(\tilde{\boldsymbol{r}}/\tilde{\boldsymbol{\mu}}\big)\, \tilde{\mathbf{P}}_\circ\,, \;\; \textit{where } \tilde{\boldsymbol{\mu}} = \big(\tilde{\mathbf{P}}_\circ^{1/t^*\top}\mathbf{1}_n\big)^{t^*}, \quad (17)$$

$$\tilde{\mathbf{P}} = \tilde{\mathbf{P}}_\circ \operatorname{diag}\big(\tilde{\boldsymbol{c}}/\tilde{\boldsymbol{\xi}}\big)\,, \;\; \textit{where } \tilde{\boldsymbol{\xi}} = \big(\tilde{\mathbf{P}}_\circ^{1/t^*\top}\mathbf{1}_n\big)^{t^*}. \quad (18)$$

It is imperative to understand the set of solutions of the alternating Bregman projections are of the form $\tilde{\mathbf{P}} = \operatorname{diag}(\boldsymbol{\nu})\tilde{\mathbf{P}}_\circ \operatorname{diag}(\boldsymbol{\gamma})$, whose analytical shape is different from that required by Theorems 2 and 3. The primary reason for the solution being an approximation is the fact that the solution set by definition is non-convex for $t \neq 1$. We empirically evaluate the quality of this approximation.

Sinkhorn Balancing for Approximating (15) & (16)

Regularized Expected Cost

We have, in general, the possible simplification

$$\tilde{P}_{ij} = \frac{\tilde{r}_i \tilde{c}_j}{\exp_t(\nu_i) \otimes_t \exp_t(1/t^* \lambda M_{ij}) \otimes_t \exp_t \gamma_j}\,.$$

Conditions for these simplifications to be valid include t close enough to 1. We can define an *expected seed matrix* for the problem (8) as

$$\tilde{\mathbf{K}}_e \doteq \frac{1}{\exp_t\big(1/t^* \lambda\, \mathbf{M}\big)} = \exp_t\big(\ominus_t\, 1/t^* \lambda\, \mathbf{M}\big)\,, \qquad (19)$$

where $\ominus_t a \doteq \frac{-a}{1+(1-t)a}$ (and $\lim_{t\to 1} \ominus_t a = -a$), and the solution (15) can this time be approximated through a diagonally equivalent matrix of $\tilde{\mathbf{K}}_e$ (See the preceding section with $\tilde{\mathbf{P}}_\circ \doteq \tilde{\mathbf{K}}_e$).

Regularized Measured Cost

We can also simplify (16) as

$$\tilde{P}_{ij} = \frac{\exp_t\big(\log_t(\tilde{r}_i\tilde{c}_j) - \lambda M_{ij}\big)}{\exp_t(\nu_i) \otimes_t \exp_t(\gamma_j)} \qquad (20)$$

(see the appendix) where $a \otimes_t b \doteq \big[a^{1-t} + b^{1-t} - 1\big]_+^{\frac{1}{1-t}}$ (and $\lim_{t\to 1} a \otimes_t b = a \cdot b$). The regularizer in (9) is the Bregman divergence to the independence table, rather than the tempered entropy function $\varphi_t(\tilde{\boldsymbol{P}})$; the reason being that cross terms in the numerator of the solution (20) can no longer be combined with the denominator, primarily because $\exp_t(\log_t(a) - b) \neq a\, \exp_t(-b)$ for $t \neq 1$. Furthermore, due to the normalization in (20), the solution is not a diagonally equivalent scaling of the *measured seed matrix*

$$\tilde{\mathbf{K}}_m \doteq \exp_t\big(\log_t(\tilde{\boldsymbol{r}}\tilde{\boldsymbol{c}}^\top) - \lambda\, \mathbf{M}\big)\,, \qquad (21)$$

but it can be approximated by a diagonally equivalent matrix of $\tilde{\mathbf{K}}_m$ (See the preceding section with $\tilde{\mathbf{P}}_\circ \doteq \tilde{\mathbf{K}}_m$). In terms of sparsity patterns, the simplification (20) has the direct effect of constraining the transportation plan to the coordinates $\log_t(\tilde{r}_i\tilde{c}_j) - \lambda M_{ij} > -1/(1-t)$ (otherwise, $(\tilde{K}_m)_{ij} = 0$). Remark the link with $\mathbf{M}'$ in (13): $\tilde{\mathbf{K}}_m = [-\mathbf{M}']_+$. So, the simplification (20) prevents coordinates $\log_t(\tilde{r}_i\tilde{c}_j) - \lambda M_{ij}$ too small so that they are $< -1/(1-t)$ to yield non-zero transport, as would Theorem 1 authorize. This is not an issue because, as the theorem shows, only a subset of such coordinates would yield non-zero transport anyway, and the approximation brings the benefit of being in position to design in a tighter way the desired sparsity patterns. One needs to make sure that the support of $\tilde{\mathbf{K}}_m$ is big enough to allow for feasible solutions — which is not also a real issue, granted that any optimal solution to the unregularized expected cost is so sparse that it has at most $2n-1$ non-zero values (Peyré and Cuturi 2019). Both cases (19) and (21) reduce to the entropic regularized seed $\mathbf{K} = \exp(-\lambda \mathbf{M})$ (up to a diagonal scaling) when $t \to 1$. We then get the general approach to approximating (15) and (16), which consists in a *reduction* to Sinkhorn balancing with specific initializations.

Solution by Reduction to Sinkhorn Balancing

It can be simply verified that the projection steps in (17) and (18) can be written in terms of the transport polytope of $\boldsymbol{r}$ and $\boldsymbol{c}$, when working directly with the transport plan $\mathbf{P} = \tilde{\mathbf{P}}^{1/t^*} \in U_n(\boldsymbol{r}, \boldsymbol{c})$. Notably, the steps are identical to the standard Sinkhorn's iterations (i.e., scaling of the rows and columns), which can be computed efficiently via Algorithm 1. The main alterations to carry out the iterations are: i) form the seed matrix $\tilde{\mathbf{K}}$ via (19) or (21), ii) apply Sinkhorn's iterations to $\tilde{\mathbf{K}}^{1/t^*}$, iii) map the solution back to the co-polyhedral by computing its t^*-th power. See Algorithm 2 for the steps.

Sparsity of Approximate Solutions

Although the sparsity result of Theorem 1 is for the closed-form solution of the regularized OT plan, the approximate solutions via Sinkhorn may result in a sparse solution for an appropriate choice of t and for sufficiently large λ.

Proposition 5. *For $\mathbf{M} \in \mathbb{R}_+^{n\times n}$ (assuming $t < 2$), the expected cost seed matrix (19) contains zero elements for $t > 1$ and sufficiently large λ. Similarly, the measured cost seed matrix (21) includes zero elements for $t < 1$ when λ is large enough. Both matrices are positive otherwise for any $\lambda > 0$. Additionally, in both cases, for $\lambda_1 < \lambda_2$, the zero elements of the seed matrix induced by λ_1 are a subset of the zero elements induced by λ_2.*

The level of the sparsity of the solution monotonically increases with λ. Nonetheless, when the sparsity level is too high (*e.g.*, for $\lambda \to \infty$, $\tilde{\mathbf{K}} \to \mathbf{0}_{n\times n}$) the resulting seed matrix may no longer induce a feasible solution that is diagonally equivalent to a transport plan $\tilde{\mathbf{P}} \in \tilde{U}_n(\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{c}})$, as stated next.

Convergence and Remarks on Feasibility

Franklin and Lorenz (1989) show the linear convergence of

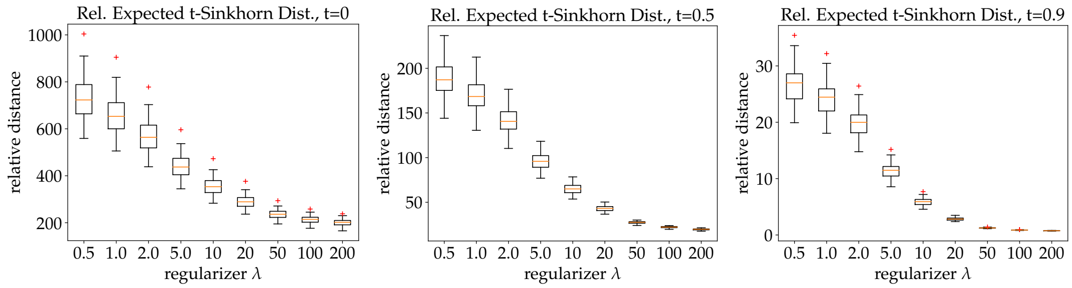
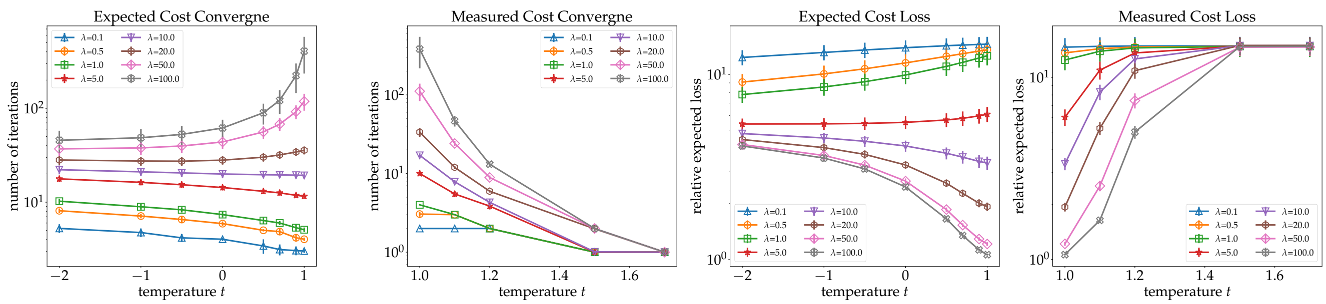


Figure 3: The expected t-Sinkhorn distance relative to the Sinkhorn distance for different values of t. As $t \to 1$, the approximation error due to solving the unconstrained problem via alternating Bregman projections becomes smaller and the expected t-Sinkhorn distance converges to the Sinkhorn distance when $\lambda \to \infty$.



(a) (b) (c) (d)

Figure 4: Number of iterations to converge for (a) expected and (b) measured cost OT for different values of λ. Relative (to the OT solution) expected cost for the two cases, respectively, shown in (c) and (d).

the both scaling factors $\boldsymbol{\mu}$ and $\boldsymbol{\xi}$ of Sinkhorn's algorithm for *positive matrices*. Specifically, the convergence rate is proportional to the square of the *contraction coefficient* $\kappa(\mathbf{K}) \doteq \tanh(\delta(\mathbf{K})/4)$ where $\delta(\mathbf{K}) \doteq \log \max_{i,j,k,\ell} \frac{K_{i\ell}K_{jk}}{K_{j\ell}K_{ik}}$ is called the *projective diameter* of the linear map $\mathbf{K}$.

Remark 3. *When the seed matrix $\tilde{\mathbf{K}}$ in Algorithm 2 is positive, the linear convergence is then an immediate consequence of the convergence of Sinkhorn's iteration. The range of t for which $\tilde{\mathbf{K}}$ is a positive matrix is characterized by Proposition 5 and the convergence rate is thus proportional to $\kappa(\tilde{\mathbf{K}}^{1/t^*})^2$. Note that for $t = 1$, $\kappa(\operatorname{diag}(\boldsymbol{r})\exp(-\lambda\,\mathbf{M})\operatorname{diag}(\boldsymbol{c})) = \kappa(\exp(-\lambda\,\mathbf{M}))$ and both seeds* (21) *and* (19) *recover the convergence rate of the EOT.*

Although the convergence of Algorithm 2 is guaranteed for positive $\tilde{\mathbf{K}}$, we still need to specify when a solution exists for non-negative $\tilde{\mathbf{K}}$ (see the appendix for remarks on the feasibility). Nonetheless, *if a solution exists*, we have the following result in terms of the seed and transport plan.

Remark 4. *The non-negative matrix $\tilde{\mathbf{K}}$ is diagonally equivalent to $\tilde{\mathbf{P}} \in \tilde{U}_n(\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{c}})$ if an only if $\tilde{\mathbf{K}}^{1/t^*}$ with $t^* > 0$ is diagonally equivalent to a matrix $\mathbf{P} \in U_n(\boldsymbol{r}, \boldsymbol{c})$.*

Experiments

We provide experimental evidence to validate the results in the paper. For each case, we sample $\mathbf{M}$ uniformly between $[0, 1]$ and also sample $\boldsymbol{r}$ and $\boldsymbol{c}$ randomly. Due to limited space, we defer some of the results to the appendix.

t-Sinkhorn Distances

We plot the relative cost of the tempered entropic regularized OT to the value of the unregularized (measured or expected) cost. For the experiment, we set $n = 64$ and average over 20 trials. Figure 3 shows the relative expected cost for different t and λ. The relative cost decreases with larger λ, and the asymptotic value is closer to zero for t is closer to one, which is the case for the EOT.

Convergence of Tempered OT

We measure the number of steps to converge for the tempered entropic regularized OT problem using Sinkhorn's iterations for different values of t and λ. We stop Sinkhorn's iterations when the maximum absolute change in each coordinate of $\boldsymbol{\xi}$ is less than $1\mathrm{e}{-10}$. For the experiment, we set $n = 64$ and average over 100 trials. Figure 4 shows the number of iterations to converge along with the relative *expected* cost to the solution of the unregularized OT problem. The number of iterations to converge follows a similar pattern to the contraction ratios of the seed matrices, shown in

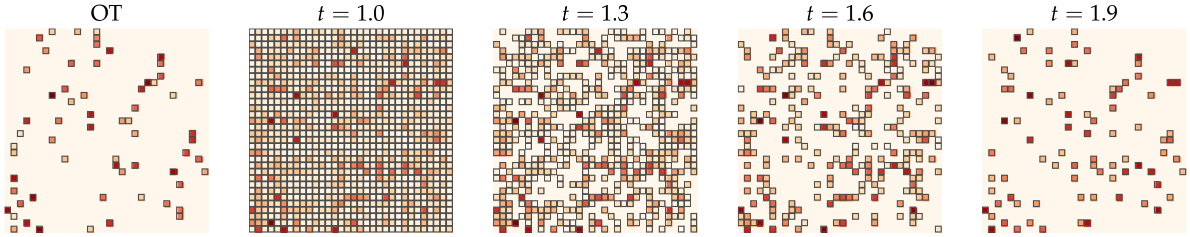


Figure 5: Transport plans induced by OT and the expected cost formulation for different values of t. The non-zero values are marked by a square. The EOT ($t = 1$) induces a fully-dense plan. The sparsity of the solution increases by increasing $1 < t < 2$.

Figure 7 in the appendix, while the relative expected cost is inversely proportional to the number of iterations. This result highlights the trade-off between the convergence and the (expected) transport cost.

Sparse Solutions

We analyze the sparsity of the solution of the (unregularized) OT problem as well as the solutions of the regularized expected cost problem (8) for $t \in [1, 2)$. Note that $t = 1$ is equal to the EOT problem (1). We set $n = 32$ and, for each case, set $\lambda = 6.0/t^*$ to offset the scaling factor in (19). In Figure 5, we show the non-zero values of the transport plans (more precisely, values larger than $1\mathrm{e}{-25}$). OT induces a sparse solution with $2n - 1 = 63$ non-zero components. On the other hand, the EOT ($t = 1$) solution is fully dense, with 1024 non-zero components. The sparsity increased by increasing $t \in [1, 2)$. In this case, the transport plan with $t = 1.9$ has only 83 non-zero values. More results for the regularized measured cost are given in the appendix.

Conclusions

We investigated the regularized version of the optimal transport problem with tempered exponential measures. The regularizations are Bregman divergences induced by the negative tempered Tsallis entropy. We studied how regularization affects the sparsity pattern of the solution and adapted Sinkhorn balancing to quickly approximate the solution.

Acknowledgments

The authors warmly thank Mathieu Blondel for remarks and discussions around the material presented.

References

Ali, S.-M.; and Silvey, S.-D.-S. 1966. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society B*, 28: 131–142.

Altschuler, J.-M.; Weed, J.; and Rigollet, P. 2017. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *NeurIPS'17*, 1964–1974.

Amari, S.-I. 2016. *Information Geometry and Its Applications*. Springer-Verlag, Berlin.

Amari, S.-I.; and Nagaoka, H. 2000. *Methods of Information Geometry*. Oxford University Press.

Amid, E.; Nock, R.; and Warmuth, M. K. 2023. Clustering above Exponential Families with Tempered Exponential Measures. In *International Conference on Artificial Intelligence and Statistics*, 2994–3017. PMLR.

Amid, E.; Warmuth, M. K. K.; Anil, R.; and Koren, T. 2019. Robust Bi-Tempered Logistic Loss Based on Bregman Divergences. In *Advances in Neural Information Processing Systems*.

Ballu, M.; and Berthet, Q. 2023. Mirror Sinkhorn: Fast Online Optimization on Transport Polytopes. In *Proceedings of the 40th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR.

Banerjee, A.; Merugu, S.; Dhillon, I.; and Ghosh, J. 2004. Clustering with Bregman Divergences. In *Proc. of the 4^{th} SIAM International Conference on Data Mining*, 234–245.

Benamou, J.-D. 2003. Numerical resolution of an "unbalanced" mass transport problem. *ESAIM: Mathematical Modelling and Numerical Analysis*, 37: 851–868.

Blondel, M.; Seguy, V.; and Rolet, A. 2018. Smooth and Sparse Optimal Transport. In *AISTATS'18*, volume 84 of *Proceedings of Machine Learning Research*, 880–889.

Blumer, A.; Ehrenfeucht, A.; Haussler, D.; and Warmuth, M. K. 1987. Occam's razor. *Information Processing Letters*, 377–380.

Bregman, L. M. 1967. 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.

Brualdi, R. A. 1968. Convex sets of non-negative matrices. *Canadian Journal of Mathematics*, 20: 144–157.

Csiszár, I. 1963. Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizitat von Markoffschen Ketten. *Magyar. Tud. Akad. Mat. Kutato Int. Kozl.*, 8: 85–108.

Cuturi, M. 2013. Sinkhorn distances: lightspeed computation of optimal transport. In *Advances in neural information processing systems (NeurIPS)*, volume 26, 2292–2300.

Dantzig, G. 1949. Programming of interdependent activities: II mathematical model. *Econometrica*, 17: 200–211.

Dessein, A.; Papadakis, N.; and Rouas, J.-L. 2018. Regularized Optimal Transport and the Rot Mover's Distance. *Journal of Machine Learning Research*, 19(15): 1–53.

Fertin, G.; Rusu, I.; and Vialette, S. 2015. Obtaining a triangular matrix by independent row-column permutations. In *Algorithms and Computation: 26th International Symposium, ISAAC 2015, Nagoya, Japan, December 9-11, 2015, Proceedings 26*, 165–175. Springer.

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

Haynes, K.-E.; and Fotheringham, A.-S. 1984. *Gravity and spatial interaction models*. Berkeley Hills, CA.

Helmbold, D. P.; and Warmuth, M. K. 2009. Learning Permutations with Exponential Weights. *Journal of Machine Learning Research*, 10: 1705–1736.

Horn, R. A.; and Johnson, C. R. 1990. *Matrix Analysis*. Cambridge University Press.

Janati, H.; Muzellec, B.; Peyré, G.; and Cuturi, M. 2020. Entropic Optimal Transport between Unbalanced Gaussian Measures has a Closed Form. In *NeurIPS*33*.

Kantorovich, L. 1958. O peremeshchenii mass. *Doklady Akademii Nauk SSSR*, 37: 227–230.

Karp, R. 1972. Reducibility among combinatorial problems. In Miller, R.; and Thatcher, J., eds., *Complexity of Computer Computations*, 85–103. Plenum Press.

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

Liu, T.; Puigcerver, J.; and Blondel, M. 2023. Sparsity-Constrained Optimal Transport. In *ICLR'23*.

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

Muzellec, B.; Nock, R.; Patrini, G.; and Nielsen, F. 2017. Tsallis Regularized Optimal Transport and Ecological Inference. In *AAAI'17*, 2387–2393.

Nivanen, L.; Le Méhauté, A.; and Wang, Q.-A. 2003. Generalized algebra within a nonextensive statistics. *Reports on Mathematical Physics*, 52: 437–444.

Peyré, G.; and Cuturi, M. 2019. Computational Optimal Transport. *Found. Trends Mach. Learn.*, 11(5-6): 355–607.

Sinkhorn, R.; and Knopp, P. 1967. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2): 343–348.

Tarjan, R. 1972. Depth-first search and linear graph algorithms. *SIAM journal on computing*, 1(2): 146–160.

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