Optimal Transport with Tempered Exponential Measures

Ehsan Amid\textsuperscript{1*}, Frank Nielsen\textsuperscript{2}, Richard Nock\textsuperscript{3}, Manfred K. Warmuth\textsuperscript{3}

\textsuperscript{1} Google DeepMind
\hspace{1em} \textsuperscript{2} Sony Computer Science Laboratories Inc.
\hspace{1em} \textsuperscript{3} Google Research

eamid@google.com, frank.nielsen@acm.org, \{richardnock, manfred\}@google.com

\section*{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.

\section*{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 (Peyrè and Cuturi 2019; Villani 2009). Those are appealing property to meet distance axioms: Optimal Transport distances (Peyré and Cuturi 2019; Villani 2009). Those can be interesting in wide ML fields (Peyrè and Cuturi 2019; Villani 2009). Those can be interesting in wide ML fields (Peyrè and Cuturi 2019; Villani 2009).

In classical discrete optimal transport (OT), we are given a cost matrix $P \in \mathbb{R}^{n \times n}$ and two probability vectors $\mathbf{r}$ and column $\mathbf{c}$ in the simplex $\Delta_n = \{p \in \mathbb{R}^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 it is not 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.\footnote{Authors listed alphabetically. Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.}

\section*{Definitions}

\subsection*{Optimal Transport in the Simplex}
In classical discrete optimal transport (OT), we are given a cost matrix $M \in \mathbb{R}^{n \times n}$ and two probability vectors $\mathbf{r}$ and column $\mathbf{c}$ in the simplex $\Delta_n = \{p \in \mathbb{R}^n : \}$

\footnote{The full article is available at https://arxiv.org/abs/2309.04015.}
There are several key objects in the simplex $\sum_{i \in [n]} \rho_i = 1$.

The optimal transport problem seeks to find a distance, though only non-negativity is really important for the optimization constraints in (1). The support of the slack of the constraints over the dual variables (Peyré and Cuturi 2016) but the OT plan solution to (1), say $\tilde{\rho}(x) = \exp\left(\langle \theta, \varphi(x) \rangle \right) / \exp\left(\langle G_t(\theta) \rangle \right)$, 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 $\max(0, z)$):

$$\tilde{\rho}(x) = \frac{\exp\left(\langle \theta, \varphi(x) \rangle \right)}{\exp\left(\langle G_t(\theta) \rangle \right)}, \quad \exp_t(z) = \left[1 + (1-t)z\right]^{\frac{1}{t-1}},$$

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

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

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

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

**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 $\tilde{u}, \tilde{v} \in \mathbb{R}^{m}$, we define the generalized tempered relative entropy as (Amid et al. 2019)

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

Just like the KL divergence ($t \rightarrow 1$), the tempered relative entropy is a Bregman divergence, induced by the generator $\varphi_t(z) = \log_t z - \log_{t-1} (z)$, 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 $\Delta_n$ in $\mathbb{R}^n$.

**Definition 1.** The co-simplex of $\mathbb{R}^n$, $\tilde{\Delta}_n$ is defined as $\tilde{\Delta}_n = \{\tilde{p} \in \mathbb{R}^n: \tilde{p} \geq 0 \land 1^\top \tilde{p}^{1/t^*} = 1\}$. Note that $\tilde{p}^{1/t^*} = p \in \Delta_n$ iff $\tilde{p} \in \tilde{\Delta}_n$ and $\tilde{\Delta}_n \rightarrow \Delta_n$ when $t \rightarrow 1$. Similarly, given $\tilde{r}, \tilde{c} \in \tilde{\Delta}_n$, 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 $\tilde{r}, \tilde{c} \in \tilde{\Delta}_n$ is defined as $\tilde{U}_n(\tilde{r}, \tilde{c}) = \{\tilde{P} \in \mathbb{R}_+^{n \times n} | \tilde{P}^{1/t^*} 1 = \tilde{p}^{1/t^*}, \tilde{P}^{1/t^*} \tilde{1} = \tilde{c}^{1/t^*}\}$. 

10839
Likewise, $\tilde{U}_n(\tilde{r}, \tilde{c}) \to U_n(r, c)$ (the transport polytope) in the limit $t \to 1$. More importantly, using our notation convention,
\[
P \in U_n(r, c) \iff \tilde{P} \in \tilde{U}_n(\tilde{r}, \tilde{c}).
\]

**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 iterative Bregman projections onto the affine constraint (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 Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)

---

\textsuperscript{2}Nonquam ponenda est pluralitas sine necessitate, “plurality should never be imposed without necessity”, William of Ockham, XIV\textsuperscript{th} century.

**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 $M \in \mathbb{R}^{n \times n}$. The first is the classical OT cost; we denote it as the expected cost,
\[
\bar{d}_M^t(\tilde{r}, \tilde{c}) = \min_{P \in \tilde{U}_n(\tilde{r}, \tilde{c})} \langle P, M \rangle
\]

with our notations, note that the constraint is equivalent to $P \in \tilde{U}_n(r, c))$. 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:
\[
\hat{d}_M^t(\tilde{r}, \tilde{c}) = \min_{\tilde{P} \in \tilde{U}_n(\tilde{r}, \tilde{c})} \langle \tilde{P}, M \rangle.
\]

$d_M^t$ is a distance if $M$ is a metric matrix. $\hat{d}_M^t$ 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 $t \to 1$.

**Proposition 1.** If $M$ is a distance matrix and $t \leq 1$, $(\hat{d}_M^t(\tilde{x}, \tilde{z}))^{2-t} \leq M^{1-t} \cdot \left( d_M^t(\tilde{x}, \tilde{y}) + d_M^t(\tilde{y}, \tilde{z}) \right)$, $\forall \tilde{x}, \tilde{y}, \tilde{z} \in \tilde{\Delta}_n$, where $M = \sum_{ij} M_{ij}$.

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 $\kappa$; in this case, the RHS scales by $\kappa^{2-t}$ and the LHS 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 mic 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 (Helmibold 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 Balu and Berthet (2023) on mirror Sinkhorn algorithms.
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:

\[ \tilde{U}_n^\varepsilon(\tilde{r}, \tilde{c}) = \{ \tilde{P} \in \tilde{U}_n(\tilde{r}, \tilde{c}) | D_t(\tilde{P}||\tilde{r}\tilde{c}^\top) \leq \varepsilon \}, \]  

where \( \varepsilon \) 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, \( M \) is a metric matrix (Cuturi 2013). For a more general \( t \), we can first remark that

\[ D_t(\tilde{P}||\tilde{r}\tilde{c}^\top) \leq \frac{1}{1-t}, \forall \tilde{P} \in \tilde{D}_{n \times n}, \forall \tilde{r}, \tilde{c} \in \tilde{D}_n, \]

so that we can consider that

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

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

Notably, the generalization of the ball \( \tilde{U}_\varepsilon^\top(\tilde{r}, \tilde{c}) \) 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 \( \tilde{P}, \tilde{Q} \in \tilde{U}_n^\varepsilon(\tilde{r}, \tilde{c}) \) and any \( t \in \mathbb{R} \setminus \{2\} \),

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

**Regularized OT Costs**

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

\[ d_M^{\lambda}(\tilde{r}, \tilde{c}) = \min_{P \in \tilde{U}_n(\tilde{r}, \tilde{c})} \langle P, M \rangle + \frac{1}{\lambda} \cdot D_t(\tilde{P}||\tilde{r}\tilde{c}^\top), \quad (8) \]

for \( \lambda > 0 \) and the regularized measured cost

\[ d_M^{\lambda}(\tilde{r}, \tilde{c}) = \min_{P \in \tilde{U}_n(\tilde{r}, \tilde{c})} \langle P, M \rangle + \frac{1}{\lambda} \cdot D_t(\tilde{P}||\tilde{r}\tilde{c}^\top). \quad (9) \]

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 \( \lambda \) 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 \( M \) 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 \( \tilde{P} \in \tilde{D}_{n \times n} \), any \( t \in [0,1) \) and \( 0 \leq \varepsilon \leq 1/(1-t) \), letting \( M_t = (\tilde{r}\tilde{c}^\top)^{1-t} \), we have

\[ D_t(\tilde{P}||\tilde{r}\tilde{c}^\top) \leq \varepsilon \iff \langle \tilde{P}, M_t \rangle \geq \exp^{1-t}(\varepsilon). \quad (10) \]

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

\[ D_t(\tilde{P}||\tilde{r}\tilde{c}^\top) = \frac{1}{1-t} \cdot (1 - \langle \tilde{P}, M_t \rangle). \quad (11) \]

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 \( M_t \) 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

\[ \lambda \cdot d_M^{\lambda}(\tilde{r}, \tilde{c}) = \frac{1}{1-t} + \min_{P \in \tilde{U}_n(\tilde{r}, \tilde{c})} \langle \tilde{P}, M' \rangle, \quad (12) \]

\[ M' = \lambda \cdot M - \frac{1}{1-t} \cdot M_t, \quad (13) \]

with \( M_t \) 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 \( j \neq \bar{k} \) containing two distinct paths of negative costs and at least one of positive cost.

**Theorem 1.** Let \( \tilde{P} \) be the optimal solution to the regularized measured cost (9). Let \( \tilde{S} \) be the support indicator matrix of \( \tilde{P} \), 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.
For any coordinates \( i \neq k \) and \( j \neq l \), suppose we have the following configuration (Figure 2): \( M_{ij}^{t} > 0, M_{il}^{t} < 0, M_{kl}^{t} < 0 \). Then we have the following

- if \( M_{kl}^{t} > 0 \), then \( S_{ij} = 0 \) or (non exclusive) \( S_{kl} = 0 \);
- if \( M_{kl}^{t} < 0 \) and \( S_{ij} = 1 \) then necessarily

\[
P_{kl}^{1-t} \leq \frac{|M_{kl}^{t}|}{|M_{ij}^{t}| + |M_{il}^{t}|} \cdot \max\{P_{k,l}, \tilde{P}_{k,l}, \hat{P}_{k,l}\}^{1-t}.
\]

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 \( M' \) are under tight control by the user, and they enforce non-zero transport, so the flexibility of the design of \( M' \), via tuning the strength of the regularization \( \lambda \) and the TEMs family via \( t \), can allow to tightly design transport patterns.

**Interpretation of Matrix \( M' \)**

Coordinate \((i, j)\) is \( M_{ij}^{t} = \lambda M_{ij} - (\bar{r}_{i} \bar{c}_{j})^{1-t}/(1-t) \). Quite remarkably, the term \((\bar{r}_{i} \bar{c}_{j})^{1-t}/(1-t)\) 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 \( M \) factors the distance, then we can get the full interaction term with the distance via its factorization. Hence, we can abstract any coordinate in \( M' \) as:

\[
M_{ij}^{t} \propto \text{original cost}(i, j) - \text{interaction}(i, j).
\]

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:** Sinkhorn(\( K, r, c \))

**Input:** Cost matrix \( K \in \mathbb{R}_{++}^{n \times n} \), \( r, c \in \Delta_{n} \)

**Output:** \( P \in \hat{U}_{n}(r, c) \)

1: \( \mu \leftarrow \mathbf{1}_{n}, \xi \leftarrow \mathbf{1}_{n} \)
2: while not converged do
3: \( \mu \leftarrow r/(K \xi) \)
4: \( \xi \leftarrow c/(K' \mu) \)
5: end while
6: return diag(\( \mu \)) \( K \) diag(\( \xi \))

**Algorithm 2:** Regularized Optimal Transport with TEMs

**Input:** Cost matrix \( M \in \mathbb{R}_{++}^{n \times n}, r, c \in \Delta_{n} \), Regularizer \( \lambda \)

**Output:** \( \hat{P} \in \tilde{U}_{n}(\hat{r}, \hat{c}) \)

1: \( \tilde{K} \leftarrow \tilde{K}_{n} \) or \( \tilde{K}_{m} \)
2: \( P \leftarrow \text{Sinkhorn}(\tilde{K}, r, c) \) \( (K = \tilde{K}^{1/n}) \)
3: return \( P^{t} \)

**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

\[
\hat{P}_{ij} = \frac{\hat{r}_{i} \hat{c}_{j}}{\exp(t(\nu_{i} + \gamma_{j} + \lambda M_{ij}))},
\]

where \( \nu_{i} \) and \( \gamma_{j} \) are chosen s.t. \( \hat{P} \in \hat{U}_{n}(\hat{r}, \hat{c}) \).

**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

\[
\hat{P}_{ij} = \exp(t((\log_{s}(\hat{r}_{i} \hat{c}_{j}) - \lambda M_{ij}) \Theta_{t}(\nu_{i} + \gamma_{j}))),
\]

where \( a \oplus b = (a - b)/(1 + (1 - t)b) \) and \( \nu_{i} \) and \( \gamma_{j} \) are chosen s.t. \( P \in \hat{U}_{n}(\hat{r}, \hat{c}) \).

**Approximation via Alternating Projections**

The Lagrange multipliers \( \nu_{i} \) and \( \gamma_{j} \) 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 \( \hat{P}_c \in \mathbb{R}^{n \times n} \), the row projection and column projection to \( \hat{U}_n(\hat{r}, \hat{c}) \) correspond to
\[
\min_{\hat{P} \cdot \hat{P}^+} D_l(\hat{P}_c | \hat{P}_c), \quad \text{(row projection)}
\]
\[
\min_{\hat{P} \cdot \hat{P}^+} D_l(\hat{P}_c | \hat{P}_c), \quad \text{(column projection)}
\]
in which, we use the shorthand notation \( \hat{P} = \hat{P}_c^{1/t^*} \) (similarly for \( r \) and \( c \)).

**Theorem 4.** Given \( \hat{P}_c \in \mathbb{R}^{n \times n} \), the row and column projections to \( \hat{U}_n(\hat{r}, \hat{c}) \) can be performed respectively via
\[
\hat{P} = \text{diag}(\hat{r}/\hat{\mu}) \hat{P}_c, \quad \text{where} \quad \hat{\mu} = (\hat{P}_c^{1/t^*} \otimes 1_n)^{t^*}, \quad (17)
\]
\[
\hat{P} = \hat{P}_c \text{diag}(\hat{c}/\hat{\xi}), \quad \text{where} \quad \hat{\xi} = (\hat{P}_c^{1/t^*} \otimes 1_n)^{t^*}. \quad (18)
\]

It is imperative to understand the set of solutions of the alternating Bregman projections are of the form \( \hat{P} = \text{diag}(\nu) \hat{P}_c \text{diag}(\gamma) \), 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
\[
\hat{P}_{ij} = \frac{\hat{r}_i \hat{c}_j}{\exp_\nu(v_i) \otimes_t \exp_\nu(1/\lambda M_{ij}) \otimes_t \exp_\gamma \gamma_j}.
\]
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
\[
\hat{K}_c = \frac{1}{\exp_t(1/\lambda M)} = \exp_t \left( \otimes_t 1/\lambda M \right), \quad (19)
\]
where \( \otimes_a = \frac{a}{1 + \alpha(1-\gamma)a} \) (and \( \lim_{t \to 1} \otimes_t a = -a \)), and the solution (15) can this time be approximated through a diagonally equivalent matrix of \( \hat{K}_c \) (See the preceding section with \( \hat{P}_c = \hat{K}_c \)).

**Regularized Measured Cost**

We can also simplify (16) as
\[
\hat{P}_{ij} = \frac{\exp_t \left( \log_t(\hat{r}_i \hat{c}_j) - \lambda M_{ij} \right)}{\exp_\nu(v_i) \otimes_t \exp_\gamma \gamma_j} \quad (20)
\]
(see the appendix) where \( a \otimes_t b = [a^{1-t} + b^{1-t} - 1]^{1/t} \) (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 \( \varphi_t(P) \); the reason being that cross terms in the numerator of the solution (20) can no longer be combined with the denominator, primarily because \( \exp_\nu(\log_\nu(a) - b) \neq \exp_\nu(-b) \) for \( \nu \neq 1 \). Furthermore, due to the normalization in (20), the solution is not a diagonally equivalent scaling of the measured seed matrix
\[
\hat{K}_m = \exp_t \left( \log_t(\hat{P}^\top) - \lambda M \right), \quad (21)
\]
but it can be approximated by a diagonally equivalent matrix of \( K_m \) (See the preceding section with \( \hat{P}_c = K_m \)). In terms of sparsity patterns, the simplification (20) has the direct effect of constraining the transportation plan to the coordinates \( \log_t(\hat{r}_i \hat{c}_j) - \lambda M_{ij} < -1/(1-t) \) (otherwise, \( (\hat{K}_{m})_{ij} = 0 \)).

Remark the link with \( M' \) in (13): \( \hat{K}_m = [-M']_\perp \). So, the simplification (20) prevents coordinates \( \log_t(\hat{r}_i \hat{c}_j) - \lambda M_{ij} \) 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 \( K_m \) 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 \( K = \exp(-\lambda M) \) (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 \( r \) and \( c \), when working directly with the transport plan \( \hat{P} = \hat{P}_c^{1/t^*} \in \hat{U}_n(\hat{r}, \hat{c}) \). 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 \( \hat{K} \) via (19) or (21), ii) apply Sinkhorn’s iterations to \( \hat{K}^{1/t^*} \), 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 \( \lambda \).

**Proposition 5.** For \( M \in \mathbb{R}^{n \times n} \) (assuming \( t < 2 \)), the expected cost seed matrix (19) contains zero elements for \( t > 1 \) and sufficiently large \( \lambda \). Similarly, the measured cost seed matrix (21) includes zero elements for \( t < 1 \) when \( \lambda \) 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 \( \lambda_1 \) are a subset of the zero elements induced by \( \lambda_2 \).

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

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 μ and ξ of Sinkhorn’s algorithm for positive matrices. Specifically, the convergence rate is proportional to the square of the contraction coefficient κ(K) = tanh(δ(K)/α) where δ(K) = log max_{i,j,k,l} K_{il}K_{jk}/K_{ij}K_{lk} is called the projective diameter of the linear map K.

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

Although the convergence of Algorithm 2 is guaranteed for positive K, we still need to specify when a solution exists for non-negative 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 K is diagonally equivalent to P ∈ U_n(\bar{r}, \bar{c}) if and only if K^{1/t^*} with t^* > 0 is diagonally equivalent to a matrix P ∈ U_n(\bar{r}, \bar{c}).

Experiments
We provide experimental evidence to validate the results in the paper. For each case, we sample M uniformly between [0, 1] and also sample r and 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 ξ is less than 1e−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
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 $\lambda = 6.0/t^8$ 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 $1 \times 10^{-5}$). 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


