Clustering above Exponential Families with Tempered Exponential Measures

Ehsan Amid **Richard Nock** **Manfred K. Warmuth**
Google Research

Abstract

The link with exponential families has allowed k-means clustering to be generalized to a wide variety of data-generating distributions in exponential families and clustering distortions among Bregman divergences. Getting the framework to go beyond exponential families is important to lift roadblocks like the lack of robustness of some population minimizers, which is carved into their axiomatization. Current generalizations of exponential families like the q-exponential families or even the deformed exponential families fail at achieving the goal. In this paper, we provide a new attempt at getting a complete framework, grounded in a new generalization of exponential families that we introduce, called *tempered* exponential *measures* (TEM). TEMs keep the maximum entropy axiomatization framework of q-exponential families, but instead of normalizing the measure, normalize a dual called a co-distribution. Numerous interesting properties arise for clustering, such as improved and controllable robustness for population minimizers, that keep a simple analytic form.

1 INTRODUCTION

Compared to supervised learning, clustering is a loosely formulated problem. It is not clear which objects to cluster (Bonnier, 1887) and which function to optimize (von Luxburg et al., 2012). Among techniques, k-means is conceivably the most popular clustering algorithm. Decades after its introduction (Lloyd, 1982; Steinhaus, 1956), k-means remains hugely popular (Flach, 2012; Hastie et al., 2002) and still has a very active research agenda (Paul et al., 2021; Vellal et al., 2022). k-means has a comparative advantage over other techniques from two standpoints: the objects clustered are equivalently the expectation parameters of Gaussians with identity covariance (Nielsen and Garcia, 2009, pp 17). Thus, there is a sound statistical interpretation of the objects being clustered (or the parameters learned), and the eventual generative process of the training data (Vellal et al., 2022). Also, the loss optimized has solid information theoretic grounds: it is a Bregman divergence known as squared Mahalanobis distance, which stems from the KL divergence between two such Gaussians. This very elegant property can be extended "above" the Gaussian distribution to any *exponential family* (Banerjee et al., 2005b), generalizing the clustering losses used to general Bregman divergences, which leads to improved designs in specific application areas with practical implications (Févotte et al., 2009).

In fact, this property can be extended further *above* exponential families, towards q- and deformed- exponential families using escort distributions (Nock et al., 2017, Theorem 3), (Amari et al., 2012; Vigelis and Cavalcante, 2011), *but* there is no more "novelty" on the parameters' side as Bregman divergences are kept as distortions between parameters.

Getting such novelty would be crucial for clustering: the cluster centers, also called *population minimizers* that elicit the most general clustering algorithms belong to a small set from the analytic standpoint, with one, the arithmetic average, being ubiquitous for all Bregman divergences (Banerjee et al., 2005b, Proposition 1). This is an issue for clustering in terms of robustness to outliers (Amari, 2016, Section 11.1.6). For example, the arithmetic average lacks robustness: adding a single point that progressively drifts away will drag a cluster center arbitrarily far away from its initial value, bringing considerable instability to clustering. A solution to this problem cannot easily arise within exponential families, nor q-exponential nor deformed exponential families because the arithmetic average as maximum likelihood estimator is carved in their axiomatization (Barndorff-Nielsen, 1979, pp 137). Adding robustness is not necessarily an issue by going "above" Bregman divergences (Nock et al., 2016; Vemuri et al., 2010), *but* either the connection with distributions is lost or substantially departs from exponential families. Establishing a new generalization is not trivial since it has to go through extending all key objects at play, including (i) the distributions (*i.e.*, generalizing exponential

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206. Copyright 2023 by the author(s).

families), (ii) information-theoretic distortions between distributions (KL divergence), (iii) parameter-based clustering distortions (Bregman divergences), (iv) an actionable identity between the distortions in (ii) and (iii), and of course (v) population minimizers (f-means).

We know of no approach that gets above exponential families and covers (i) through (v) while conveniently expanding the realm of distortions beyond Bregman divergences.

Our paper is a proposal that achieves this goal. While our contributions thus span all steps from (i) to (v) (Table 1 indicates where to find our contribution on each), the benefit for downstream clustering is simple: it provides improved robustness for population minimizers. Technically speaking, our key thread is close to Tsallis' nonextensive statistics framework (Tsallis, 2009), inclusive of the specific arithmetic developed in its context (Nivanen et al., 2003), *but* with an early tweak: we do not normalize the solution of maximum entropy to a distribution. Instead, we normalize a dual, which we call the *co-(tempered exponential) distribution* (COD). The unnormalized solution of the maximum entropy formulation is called *tempered exponential measure* (TEM). This is a major difference with work that followed the Amari-Naudts-Tsallis q-exponential families, the deformed exponential families, and their escort distributions, which are all normalized (Amari, 2016; Naudts, 2011; Tsallis, 2009). TEM/COD depend on a parameter t and as $t \rightarrow 1$, both the tempered exponential measure and its co-distribution converge to the same exponential family. Maintaining unnormalized measures is the trick that brings improved robustness for clustering; it creates an *unbalanced* clustering problem whose parameter distortions, generalizing Bregman divergences, belong to a broad subset known as conformal Bregman divergences (Nock et al., 2016).

Our results in (i) to (v) have wider interest than clustering; we establish additional results, such as simple and elegant closed forms for key functions including the cumulant (Theorem 3.2, unlike, *e.g.*, q-exponential families) and the total mass of the TEM (Lemma 3.3), etc. To ease reading, all proofs and additional experiments are given in an Appendix, denoted for short as "App".

2 PROBLEM AND RELATED WORK

For space constraints, we shall reduce technicalities and jargon related to exponential families to their minimum. We refer to textbooks in mathematical statistics (Barndorff-Nielsen, 1979, Chapter 8) or information geometry (Amari and Nagaoka, 2000, Section 4.2) for extensive coverage. An exponential family can be obtained by maximizing Shannon's entropy subject to normalization and conditions on the arithmetic average being the maximum likelihood estimator (Barndorff-Nielsen, 1979); its density has the general form

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) \quad \propto \quad \exp(\boldsymbol{\theta}^\top \boldsymbol{\varphi}(\boldsymbol{x}) - G(\boldsymbol{\theta})), \qquad (1)$$

Item	Brief description	Where?
(i)	(tempered exponential) measures	Thm. 3.2
(ii)	IT distortion	Eq. (19)
(iii)	parameter distortion	Eq. (20)
(iv)	relationship (ii) vs (iii)	Thm. 4.1
(v)	population minimizers	Lem. 5.1, 5.2

Table 1: Correspondence between the key items in the Introduction and where to find them in this paper.

where $\boldsymbol{\varphi}$ is the sufficient statistic, $\boldsymbol{\theta}$ is the natural parameter, and G, the cumulant or partition function, ensures normalization (the "$\propto$" symbol simplifies the carrier or base measure). The natural parameter holds the information about the "individual" distribution inside its family, which is encoded in the cumulant function G. The connection between exponential families and clustering *à-la-k*-means is simple to state and enlightening on what such clustering achieves. Given any two distributions P_i, P_j with densities p_i, p_j, a popular information-theoretic distortion measure for their comparison is an f-divergence (Ali and Silvey, 1966; Csiszár, 1963). In our context, the case of the reverse KL divergence is especially important:

$$F(P_i \| P_j) \quad \doteq \quad \int f(\mathrm{d}p_i/\mathrm{d}p_j)\mathrm{d}p_j, \;\; f \doteq -\log . \quad (2)$$

Suppose then we have a set of distributions $\{P_i\}_{i=1}^m$ and wish to find a set of distributions $\{Q_j\}_{j=1}^k$, k being user-fixed, minimizing the following loss function:

$$F(\{P_i\}_{i=1}^m, \{Q_j\}_{j=1}^k) \quad \doteq \quad \mathbb{E}_i[\min_j F(P_i \| Q_j)]. \quad (3)$$

Without any further assumption, this well-founded formulation of the clustering problem fails at two hurdles: (i) the potential intractability of the integrals to compute (2) and (ii) the formulation and/or computation of the so-called population minimizers $Q_.$ in (3). A simple assumption solves both problems simultaneously: if all distributions are assumed to belong to the *same* exponential family (characterized by a strictly convex differentiable cumulant function G), then

$$F(P_i \| Q_j) \quad = \quad D_G(\boldsymbol{\theta}_i \| \boldsymbol{\vartheta}_j), \qquad (4)$$

the Bregman divergence between the natural parameters and with generator G:

$$D_G(\boldsymbol{\theta}_i \| \boldsymbol{\vartheta}_j) \doteq G(\boldsymbol{\theta}_i) - G(\boldsymbol{\vartheta}_j) - (\boldsymbol{\theta}_i - \boldsymbol{\vartheta}_j)^\top \nabla G(\boldsymbol{\vartheta}_j). \quad (5)$$

The original k-means clustering is obtained for D_G being squared Mahalanobis distance, which corresponds to distributions being Gaussians with identity covariance. For any Bregman divergence, the *right* population minimizer in (3) is *always* the average (Banerjee et al., 2005b). This allows generalizing the k-means algorithm to all Bregman divergences by repeatedly allocating points to their closest center Bregman-wise and updating cluster centers with their cluster's average. A Bregman divergence being asymmetric in general, one can choose to flip arguments in (4): the *left* population minimizer is then an f-mean of the form $\nabla G^{-1}\mathbb{E}\nabla G(.)$.

To summarize, k-means clustering operates in disguise on *parameters* of distributions using distortions that can be understood from both the information geometric (4) and information-theoretic (2) standpoints. Such distributions can naturally be related to a generative process for the observed data and the whole algorithm can also be understood from a Bayesian standpoint (Neal, 2004) where priors and posteriors are modeled with the initial "guess" of an exponential family. All key steps to get the complete characterization consolidate steps (i) to (v), sketched in the introduction.

In the context of clustering, a relevant question is to get this scheme to work beyond its restriction of the "same exponential family" assumption. Alleviating the assumption of the *same* exponential family is not straightforward: doing so decomposes the KL divergence into a sum of two Bregman divergences, one between the cumulants (Nock et al., 2017, Theorem 24). More important is, in fact, getting above the "exponential family" assumption because the population minimizers – in particular, the average – can suffer from a lack of robustness. Unfortunately, this lack of robustness is to some extent carved into the axiomatic definition of exponential families (Barndorff-Nielsen, 1979, pp 137), (Amari, 2016, Section 2.8.1) and Bregman divergences (Banerjee et al., 2005a).

Natural candidates to rise above exponential families are q-exponential families and deformed exponential families (Amari, 2016; Amari et al., 2012; Naudts, 2011). q-exponential families essentially replace the $\exp$ in (1) by a generalization, the q-exponential:

$$\exp_q(z) \quad \doteq \quad [1+(1-q)z]_+^{1/(1-q)}, \tag{6}$$

with $[z]_+ \doteq \max\{0, z\}$ and $q \geqslant 0$ ($q > 0$) guarantees the (strict) convexity of the function. Deformed exponential families go further in the generalization by replacing the q-exponential by a χ-exponential for some χ positive non-decreasing, where its reciprocal defines the χ-logarithm:

$$\log_\chi(z) \quad \doteq \quad \int_1^z \frac{1}{\chi(t)}\mathrm{d}t, \tag{7}$$

the q-exponential being derived for $\chi(z) \doteq z^q$. An *escort* distribution can be defined, whose density has the general form $\tilde{p}_{\boldsymbol{\theta}}(\boldsymbol{x}) \propto \chi(p_{\boldsymbol{\theta}}(\boldsymbol{x}))$.

It turns out neither q-exponential nor deformed exponential families allow to generalize the Bregman divergence part in the RHS of (4), see for example Nock et al. (2017, Theorem 3). To get robustness, one previous work departs from both Bregman divergences and exponential families, Liu et al. (2012): in this case, the Bregman divergence, which computes the difference between a convex function and a tangent plane, is replaced by the distance to the projection on a tangent plane, called a total Bregman divergence. A link is established with distributions, but these are substantially different from exponential families as their natural parameters belong to a submanifold defining a curved family of distributions. Our objective is rather to go above exponential families with a sufficient broadening of the Bregman divergence part. Ideally, the divergence part would pave the way for new properties such as improved robustness for clustering, and the distribution part, beyond generalizing exponential families, would include guarantees of "proximity" to exponential families as new properties on the parameters' side appear. This is important given the ubiquitous nature of exponential families as a tool in ML.

Finally, we also note a recent breakthrough tied to exponential families: instead of an information theoretic / information geometric link as in (4), Janati et al. (2020) establish a regularized optimal transport / information geometric link. However, this was only done for Gaussian measures (not necessarily normalized).

3 TEMPERED EXPONENTIAL MEASURES AND CO-DENSITIES

We make extensive use of the q-exponential function defined in (6); in our context, parameter q is renamed t to make a clear distinction of the notations we use. We define the inverse of the t-exponential (Naudts, 2011):

$$\log_t(z) \doteq \frac{1}{1-t}\left(z^{1-t}-1\right) \qquad \left(\lim_{t\rightarrow 1}\log_t = \log\right). \tag{8}$$

We introduce notions of duality using t.

Definition 3.1. *The dual t^* of t is $t^* \doteq 1/(2-t)$; the dual $(\exp_t)^*$ of $\exp_t$ is the perspective transform:*

$$\left(\exp_t\right)^*(z) \quad \doteq \quad t^* \exp_{t^*}\left(\frac{z}{t^*}\right). \tag{9}$$

Last, we define in the same way the dual $(\log_t)^*$ of $\log_t$.

We remark that if $t \in [0,1]$ then $t^* \in [1/2, 1]$. As already outlined in the introduction, we shall make use of unnormalized measures – and by extension, unnormalized densities – when dealing with such objects, a *tilda* shall indicate it is *not necessarily normalized.* The following gives the first example, where $\boldsymbol{\varphi} : \mathcal{X} \rightarrow \mathbb{R}^d$ denotes a sufficient statistics and $\boldsymbol{\hbar}$ an expectation parameter (boldfaces are used for vector notations).

$$\tilde{\mathcal{P}}_{t|\boldsymbol{\hbar}} \doteq \left\{ \tilde{p} \;\middle|\; \begin{array}{l} \mathbb{E}_{\tilde{P}}[\boldsymbol{\varphi}] \doteq \int \boldsymbol{\varphi}(\boldsymbol{x})\,\tilde{p}(\boldsymbol{x})\,\mathrm{d}\xi = \boldsymbol{\hbar}, \\ \int \tilde{p}(\boldsymbol{x})^{1/t^*}\,\mathrm{d}\xi = 1, \\ \tilde{p}(\boldsymbol{x}) \geqslant 0, \forall \boldsymbol{x} \in \mathcal{X}. \end{array} \right\} \tag{10}$$

denotes a set of unnormalized densities.[1] Following the classical approach, we elicit the element(s) of $\tilde{\mathcal{P}}_{t|\boldsymbol{\hbar}}$ whose maximizing a generalized notion of the Tsallis entropy (Capital $\tilde{P}$ denotes the measure of density $\tilde{p}$ wrt ξ):

$$H_t(\tilde{P}) \doteq -\int \psi_t(\tilde{p}(\boldsymbol{x}))\,\mathrm{d}\xi, \tag{11}$$

$$\psi_t(z) \doteq z\log_t z - \log_{t-1} z. \tag{12}$$

With Tsallis entropy[2] and replacing in (10) $1/t*$ by constant 1, $\tilde{\mathcal{P}}_{t|\boldsymbol{\hbar}}$ would cover *probability* density functions related to t=q-exponential families (Naudts, 2004) (in fact, their escorts). The change $1 \rightarrow 1/t*$ may look cosmetic in the definition but has dramatic consequences in the whole chain of results that leads from $\tilde{\mathcal{P}}_{t|\boldsymbol{\hbar}}$ to clustering. The first major difference is that q-exponential families do not admit a closed form expression for the cumulant G in (1) (Naudts, 2004, p 12). Our solution *does* and leads to an elegant generalization of the cumulant function for exponential families. The theorem makes use of a generalization of the subtraction, $\ominus_t$, in the *t-arithmetic* introduced in Nivanen et al. (2003):

$$z \ominus_t x \doteq \frac{z-x}{1+(1-t)x}. \tag{13}$$

Theorem 3.2. *For any $t \in [0,1]$ and $\boldsymbol{\hbar} \in \mathbb{R}^d$, the solution $\arg\max_{\tilde{\mathcal{P}}_{t|\boldsymbol{\hbar}}} H_t$ has the non-normalized density*

$$\tilde{p}_{t|\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{\exp_t(\boldsymbol{\theta}^\top\boldsymbol{\varphi}(\boldsymbol{x}))}{\exp_t(G_t(\boldsymbol{\theta}))} = \exp_t(\boldsymbol{\theta}^\top\boldsymbol{\varphi}(\boldsymbol{x}) \ominus_t G_t(\boldsymbol{\theta})), \tag{14}$$

where

$$G_t(\boldsymbol{\theta}) = (\log_t)^* \int (\exp_t)^*(\boldsymbol{\theta}^\top\boldsymbol{\varphi}(\boldsymbol{x}))\mathrm{d}\xi \tag{15}$$

*is the (convex) cumulant ensuring the normalization of the **dual** $\tilde{\mathrm{p}}^{1/t^*}$; assuming G_t differentiable, the correspondence $\boldsymbol{\theta} = \nabla G_t^{-1}(\boldsymbol{\hbar})$ also holds and $\boldsymbol{\theta}$ is called a natural parameter.*

(Proof in App, Section II) Hereafter, we assume $t \in [0,1]$, which is technically convenient in our context, but note that Theorem 3.2 operates on a wider range of $t < 2$ values. We introduce the nomenclature of *tempered exponential measures* (TEM) whose (non-normalized) densities are given by (14) and their *co-densities* (COD) which are the (normalized) "duals" defined by $(\tilde{p}_{t|\boldsymbol{\theta}})^{1/t^*}$. Note that since $\lim_{t\rightarrow 1}(\log_t)^* = \log$ and $\lim_{t\rightarrow 1}(\exp_t)^* = \exp$, the cumulant function G_t in (14) is indeed a generalization of the well-known log partition function G for the standard exponential families, which normalizes the density (1). Such a closed form expression for the cumulant is not known for q-exponential families. To properly define TEMs like exponential families (Nielsen and Garcia, 2009), we have to include an eventual carrier measure: we thus let $\kappa(\boldsymbol{x})$ denote the carrier measure and let a general TEM be defined from the unnormalized density:

$$\tilde{p}_{t|\boldsymbol{\theta}}(\boldsymbol{x}) = \exp_t(\boldsymbol{\theta}^\top\boldsymbol{\varphi}(\boldsymbol{x}) \ominus_t G_t(\boldsymbol{\theta}) \oplus_t \kappa(\boldsymbol{x})) \tag{16}$$

where $\oplus_t$ was also introduced in Nivanen et al. (2003):

$$z \oplus_t x \doteq z + t + (1-t)zx. \tag{17}$$

In addition to having a cumulant in nice form, TEMs have another key property: the total mass due to $\tilde{p}_{t|\boldsymbol{\theta}}$, $\mathrm{M}_t(\boldsymbol{\theta}) \doteq \int \tilde{p}_{t|\boldsymbol{\theta}}(\boldsymbol{x})\mathrm{d}\xi$, is *also* available in an elegant closed form. Hereafter, $G_t^\star$ denotes the convex conjugate of G_t.

Lemma 3.3. $\mathrm{M}_t(\boldsymbol{\theta}) = 1 + (1-t)(G_t(\boldsymbol{\theta}) - \boldsymbol{\theta}^\top\boldsymbol{\hbar})$. *If G_t is strictly convex differentiable,*

$$\mathrm{M}_t(\boldsymbol{\theta}) = 1 + (1-t)(-G_t^\star(\boldsymbol{\hbar}))\ (= \exp_t^{1-t}(-G_t^\star(\boldsymbol{\hbar}))). \tag{18}$$

Proof sketch Surprisingly perhaps, it takes two equations to sketch its proof (more in App, Section III), both expressing $Q \doteq \mathbb{E}_{\tilde{p}_{t|\boldsymbol{\theta}}}\left[\log_t \tilde{p}_{t|\boldsymbol{\theta}}\right]$ in two different ways. First, using the definition of $\mathrm{M}_t(\boldsymbol{\theta})$ and the fact that $\tilde{p}_{t|\boldsymbol{\theta}}^{2-t}$ sums to 1, we get $Q = \int \tilde{p}_{t|\boldsymbol{\theta}}\,(1/(1-t)) \cdot \left(\tilde{p}_{t|\boldsymbol{\theta}}^{1-t} - 1\right)\mathrm{d}\xi = (1/(1-t)) \cdot (1 - \mathrm{M}_t(\boldsymbol{\theta}))$. Second, using (10) and (14), we also get $Q = \int \tilde{p}_{t|\boldsymbol{\theta}}\left(\boldsymbol{\theta}^\top\boldsymbol{\varphi} - \tilde{p}_{t|\hat{\boldsymbol{\theta}}}^{1-t}\,G_t(\boldsymbol{\theta})\right)\mathrm{d}\xi = \boldsymbol{\theta}^\top\boldsymbol{\hbar} - G_t(\boldsymbol{\theta})$. There remains to identify $\mathrm{M}_t(\boldsymbol{\theta})$ from the two expressions for Q. ■

Naturally, we recover $\lim_{t\rightarrow 1}\mathrm{M}_t(\boldsymbol{\theta}) = 1$ for exponential families. Since the total mass is positive by definition, we get two nontrivial bounds on the cumulant and its convex conjugate: $G_t(\boldsymbol{\theta}) \geqslant -1/(1-t) + \boldsymbol{\theta}^\top\boldsymbol{\hbar}$ and $G_t^\star(\boldsymbol{\hbar}) \leqslant 1/(1-t)$, both of which become vacuous when $t \rightarrow 1^-$. Table 2 presents a few examples of TEMs and the related parameters useful in our clustering context (see Sections 4 and 5). Hereafter, we assume that G_t is strictly convex and differentiable.

4 AN INFORMATION THEORETIC/GEOMETRIC RESULT

TEMs being a generalization of exponential families, one would expect that the key information theoretic / information geometric identity (4) does admit a generalization to our context. We will now show that this indeed is the case. We

[1] We omitted the domination condition of $\tilde{p}$ wrt ξ for simplicity; importantly, the *expectation* $\mathbb{E}$ also uses the *unnormalized* measure $\tilde{P}$. Checking the argument of an expectation allows to infer whether the inner density is normalized.

[2] Amid et al. (2019) introduce this slightly different notion of entropy which recovers the Itakura-Saito convex generator $\psi_2(z) = z - \log z - 1$ at $t \rightarrow 2$. Following the standard definition of the Tsallis entropy and discarding the last term in ψ_t does not affect our construction. Also note that we handle the constant term in (11) by subtracting from the integrand and adding back outside of the integral. Thus, we assume that the constant t^* inside the second term is unaffected by the integral.

TEM	Support	$\boldsymbol{\lambda}$	$\boldsymbol{\theta}$	$\boldsymbol{\hbar}$	$G_t^\star(\boldsymbol{\hbar})$
1D t-exponential	$\left[0, \frac{3-2t}{(1-t)\lambda}\right]$	λ	$\frac{-\lambda}{3-2t}$	$t^* \left(\frac{3-2t}{\lambda}\right)^{2-t^*}$	$-t^* \cdot \left(\log_{\frac{1}{2-t^*}} \left(\frac{\hbar}{t^*}\right) - 1\right)$
1D t-Gaussian ($\mu = 0$)	$\left[-\frac{1}{\sqrt{1-t}}, \frac{1}{\sqrt{1-t}}\right]$	σ^2	$-\frac{t^*}{2\sigma^2}$	$(c_{t^*}\sqrt{2})^{1-t^*}\sigma^{3-t^*}$	$-\frac{t^*}{2} \cdot \left(\log_{t^{**}}(2c_{t^*}^2\hbar) - 1\right)$

TEM	$G_t(\boldsymbol{\theta})$	$B_{G_t}(\hat{\boldsymbol{\theta}}\Vert\boldsymbol{\theta})$
1D t-exponential	$-\log_{2-t}\left((-\theta)^{\frac{1}{2-t}}\right)$	$t^* \cdot \left(\left(\frac{\hat{\theta}}{\theta}\right)^{2-t^*} - (2-t^*) \cdot \log_{t^*}\left(\frac{\hat{\theta}}{\theta}\right) - 1\right)$
1D t-Gaussian ($\mu = 0$)	$(\log_t)^*\left(\frac{c_{t^*}}{\sqrt{-\theta}}\right)$	$\frac{t^*}{2} \cdot \left(\left(\sqrt{\frac{\hat{\theta}}{\theta}}\right)^{3-t^*} - (3-t^*) \cdot \log_{t^*}\sqrt{\frac{\hat{\theta}}{\theta}} - 1\right)$

TEM	$\boldsymbol{\theta}_{\mathrm{l}}$	$\boldsymbol{\theta}_{\mathrm{r}}$
1D t-exponential	$-\mathbb{E}_i\left[\frac{1}{(-\theta_i)^{1-t^*}}\right] / \mathbb{E}_i\left[\frac{1}{(-\theta_i)^{2-t^*}}\right]$	$-\mathbb{E}_i\left[(-\theta_i)^{2-t^*}\right]$
1D t-Gaussian ($\mu = 0$)	$-\mathbb{E}_i\left[\frac{1}{(-\theta_i)^{\frac{1-t^*}{2}}}\right] / \mathbb{E}_i\left[\frac{1}{(-\theta_i)^{\frac{3-t^*}{2}}}\right]$	$-\frac{1}{(c_{t^*}\sqrt{t^*})^{1-t^*}} \cdot \mathbb{E}_i\left[(-\theta_i)^{\frac{3-t^*}{2}}\right]$

Table 2: Functions of key interest related to some TEM families, mentioning the source ($\boldsymbol{\lambda}$), natural ($\boldsymbol{\theta}$) and expectation ($\boldsymbol{\hbar}$) parameters, the cumulant $G_t(\boldsymbol{\theta})$ and its convex dual $G_t^\star(\boldsymbol{\hbar})$, the corresponding divergence on natural parameters $B_{G_t}(\hat{\boldsymbol{\theta}}\|\boldsymbol{\theta})$ (20) and its two population minimizers. Remark that for each of them α_* in Lemma 5.2 has a closed form and we obtain two *different* generalizations of Itakura-Saito divergence with $B_{G_t}(\hat{\boldsymbol{\theta}}\|\boldsymbol{\theta})$. We let $t^{**} \doteq 2/(3-t^*)$, $c_t \doteq \sqrt{\frac{\pi}{1-t}\frac{\Gamma\left(1+\frac{1}{1-t}\right)}{\Gamma\left(\frac{3}{2}+\frac{1}{1-t}\right)}}$.

first need a generalization of the KL divergence used in (2) and thus define

$$F_t(\tilde{P}_{t|\hat{\boldsymbol{\theta}}}\|\tilde{P}_{t|\boldsymbol{\theta}}) \doteq \int f\left(\frac{\mathrm{d}\tilde{p}_{t|\hat{\boldsymbol{\theta}}}}{\mathrm{d}\xi} \oslash_t \frac{\mathrm{d}\tilde{p}_{t|\boldsymbol{\theta}}}{\mathrm{d}\xi}\right) \cdot \mathrm{d}\tilde{p}_{t|\boldsymbol{\theta}}, \qquad f \doteq -\log_t, \tag{19}$$

and finally $x \oslash_t y \doteq (x^{1-t} - y^{1-t} + 1)_+^{\frac{1}{1-t}}$ if $x, y \geqslant 0$ (else it is undefined). We recover (2) as $t \rightarrow 1$. In the case of TEMs, (19) is equivalent to the *tempered KL divergence* induced by the convex function (11) that was introduced in Amid et al. (2019). We now state our generalization of (4) with F as in (2). The Theorem is important in the context of the well-known formulation of Bregman divergences using the KL divergence between exponential families (Amari, 2016, Section 2.7).

Theorem 4.1. *For any two members of the same* TEM *family,*

$$F_t(\tilde{P}_{t|\hat{\boldsymbol{\theta}}}\|\tilde{P}_{t|\boldsymbol{\theta}}) \quad = \quad B_{G_t}(\hat{\boldsymbol{\theta}}\|\boldsymbol{\theta}),$$

where

$$B_{G_t}(\hat{\boldsymbol{\theta}}\|\boldsymbol{\theta}) \doteq \frac{G_t(\hat{\boldsymbol{\theta}}) - G_t(\boldsymbol{\theta}) - (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top \nabla G_t(\boldsymbol{\theta})}{1 + (1-t)G_t(\hat{\boldsymbol{\theta}})}. \tag{20}$$

(proof in App, Section IV) One can see that the numerator in (20) is in fact the Bregman divergence with generator G_t. The whole construct B_{G_t} belongs to a generalization of Bregman divergences known as conformal Bregman divergences (Nock et al., 2016) and we recover Bregman divergences as $t \rightarrow 1$. Clustering with exponential families relies on the Bregman divergence as a distortion measure between parameters. In our case, the presence of the denominator $D_t(\hat{\boldsymbol{\theta}}) \doteq 1 + (1-t)G_t(\hat{\boldsymbol{\theta}}) = \exp_t^{1-t} G_t(\hat{\boldsymbol{\theta}})$ is crucial for clustering if $\hat{\boldsymbol{\theta}}$ is an outlier, and some algebra allows to see that $D_t(\hat{\boldsymbol{\theta}})$ is a function (increasing) proportional to the *total mass* of a TEM since this denominator also meets:

$$D_t(\hat{\boldsymbol{\theta}})^{\frac{1}{1-t^*}} \quad = \quad \int \exp_{t^*}\left(\frac{\hat{\boldsymbol{\theta}}^\top \boldsymbol{\varphi}(\boldsymbol{x})}{t^*}\right) \mathrm{d}\xi, \tag{21}$$

and the RHS is indeed proportional to $\mathrm{M}_{t^*}((1/t^*) \cdot \boldsymbol{\theta})$. In short, when $\hat{\boldsymbol{\theta}}$ in (20) is a data point, choosing a "heavy" enough TEM in $D_t(\hat{\boldsymbol{\theta}})$ can have it grow sufficiently fast as $\hat{\boldsymbol{\theta}}$ moves far away and eventually reduce its influence on the cluster centroids. We now study clustering more formally.

5 CLUSTERING AND POPULATION MINIMIZERS

Let $\{\boldsymbol{\theta}_i\}_{i=1}^m$ be a training set of parameters endowed with an implicit (*e.g.*, uniform) distribution. We define two losses for the so-called left and right population minimizers:

$$L_{\mathrm{l}}(\boldsymbol{\theta}) \doteq \mathbb{E}_i[B_{G_t}(\boldsymbol{\theta}\|\boldsymbol{\theta}_i)] \ ; \ L_{\mathrm{r}}(\boldsymbol{\theta}) \doteq \mathbb{E}_i[B_{G_t}(\boldsymbol{\theta}_i\|\boldsymbol{\theta})], \tag{22}$$

where $\mathbb{E}_i[.]$ denotes an average over the training sample. The left and right population minimizers, respectively $\boldsymbol{\theta}_{\mathrm{l}}$ and $\boldsymbol{\theta}_{\mathrm{r}}$, are then defined as

$$\boldsymbol{\theta}_{\mathrm{l}} \doteq \arg\min_{\boldsymbol{\theta}} L_{\mathrm{l}}(\boldsymbol{\theta}) \quad ; \quad \boldsymbol{\theta}_{\mathrm{r}} \doteq \arg\min_{\boldsymbol{\theta}} L_{\mathrm{r}}(\boldsymbol{\theta}). \tag{23}$$

The left and right population minimizers are the parameters whose corresponding losses are called Bregman information (Banerjee et al., 2005b, Section 3.1). We elaborate on clustering in two directions. The first is the elicitation of population minimizers and the second is their *robustness* to outliers (Amari, 2016; Fujisawa and Eguchi, 2008; Vemuri et al., 2010). To evaluate robustness, we add a new element $\boldsymbol{\theta}_*$ with weight ε in the new loss. The initial loss is scaled by $(1-\varepsilon)$. The population minimizer is said to be robust to outliers if the new population minimizer satisfies $\boldsymbol{\theta}_{\mathrm{l/r}}^{\mathrm{new}} - \boldsymbol{\theta}_{\mathrm{l/r}}^{\mathrm{old}} = \varepsilon \cdot \boldsymbol{z}(\boldsymbol{\theta}_*)$, where $\boldsymbol{z}(.)$, the influence function, has bounded norm.

Population minimizers elicited We first provide both population minimizers in (23), reminding we assume G_t strictly convex and differentiable. The simplest one is the right population minimizer.

Lemma 5.1. *The right population minimizer* (23) *is given by*

$$\boldsymbol{\theta}_{\mathrm{r}} \quad = \quad \mathbb{E}_i\left[\frac{1}{\exp_t^{1-t}(G_t(\boldsymbol{\theta}_i))} \cdot \boldsymbol{\theta}_i\right]. \qquad (24)$$

The proof of this Lemma trivially comes from (Banerjee et al., 2005b, Proposition 1), and it also recovers their result for Bregman divergences as $\lim_{t\rightarrow 1} \boldsymbol{\theta}_{\mathrm{r}} = \mathbb{E}_i[\boldsymbol{\theta}_i]$. We turn to the left population minimizer and let $\mathsf{T}_i(\boldsymbol{\theta}) \doteq G_t(\boldsymbol{\theta}_i) + (\boldsymbol{\theta}-\boldsymbol{\theta}_i)^\top \nabla G_t(\boldsymbol{\theta}_i)$ the value at $\boldsymbol{\theta}$ of the tangent hyperplane to G_t at $\boldsymbol{\theta}_i$. We also let $N(\boldsymbol{\theta}) \doteq 1 + (1-t)\mathbb{E}_i[\mathsf{T}_i(\boldsymbol{\theta})]$.

Lemma 5.2. *The critical point of* $L_{\mathrm{l}}(\boldsymbol{\theta})$ *satisfies* $\nabla G_t(\boldsymbol{\theta}_{\mathrm{l}}) = \alpha_* \cdot \mathbb{E}_i \nabla G_t(\boldsymbol{\theta}_i)$ *for some* $\alpha_* > 0$*. It is the left population minimizer if* $N(\boldsymbol{\theta}_{\mathrm{l}}) > 0$*.*

The proof, in App, Section V, also shows that since $t \leqslant 1$,

$$\alpha_* \quad \in \quad \left[1, \min_i \frac{1+(1-t)G_t(\boldsymbol{\theta}_i)}{N(\boldsymbol{\theta}_i)}\right], \qquad (25)$$

which provides a convenient initialization interval for a line search of α_*. Table 2 shows that it is also possible to get the left population minimizer in closed form for specific choices of TEM. One also sees that $\alpha \geqslant 1$ and $\lim_{t\rightarrow 1}\alpha = 1$, which gives us back the f-mean left population minimizer of Bregman divergences, also noting that $\lim_{t\rightarrow 1} N(\boldsymbol{\theta}) = 1$ so the condition $N(\boldsymbol{\theta}) > 0$ vanishes, and can in fact always be satisfied by choosing t close enough to 1.[3] Notice also that the left population minimizer is unique.

Robustness of population minimizers We first tackle the right population minimizer: the average is notoriously not robust and so in the case of Bregman divergences, this population minimizer can never be robust, regardless of the divergence. In the case of TEMs, however, the partition function gives a direct handle for robustness as the following simple Lemma shows, $\|.\|$ being any norm.

Lemma 5.3. *If* $G_t(\boldsymbol{\theta}) = \Omega(\|\boldsymbol{\theta}\|)$ *and* $t \neq 1$*, the right population minimizer* (23) *is robust.*

(proof in App, Section VI) Obviously, this robustness property vanishes as $t \rightarrow 1$. Since the denominators in (24) are an increasing function of (21), one roughly gets that robustness is achieved by picking a "heavy" enough TEM. The case of the left population minimizer is treated in the following Lemma. For any strictly convex G, the "f-mean generated by G" refers to $\nabla G^{-1}(\mathbb{E}_i \nabla G(\boldsymbol{\theta}_i))$, which is the left population minimizer for exponential families (Banerjee et al., 2005b).

Lemma 5.4. *Suppose* G_t *strongly convex differentiable. Then the left population minimizer* (23) *is robust iff the* f*-mean generated by* G_t *is robust.*

(proof in App, Section VII) A technical advantage of this Lemma is that to show the robustness of our left population minimizer, it is necessary and sufficient to investigate that of the f-mean, which can be simple to establish. As an example, the harmonic mean is robust, and it is the left population minimizer associated to the (1D) exponential distribution. In Table 2 for the 1D t-exponential TEM, one can check that θ_{l} is also robust: suppose θ_j is the outlier. When $\theta_j \rightarrow -\infty$, its influence vanishes in θ_{l} and when $\theta_j \rightarrow 0$, $\theta_{\mathrm{l}} \sim \theta_j \rightarrow 0$. Formal robustness is a binary notion but the experiments shall unveil that *improved* robustness can also be achieved for $t < 1$ when the case $t = 1$ is already robust.

Finally, there is an interesting parallel on robustness to be made between the left and right population minimizers. We have seen that the right population minimizer is robust if G_t is chosen "large enough". One can remark that $1+(1-t)G_t(\boldsymbol{\theta}_i) - N(\boldsymbol{\theta}_i) = (1-t)D_{G_t}(\boldsymbol{\theta}_i\|\boldsymbol{\theta}) \geqslant 0$, D_{G_t} being a Bregman divergence. If $\boldsymbol{\theta}_i$ is an outlier, it may well be the case that $(1+(1-t)G_t(\boldsymbol{\theta}_i))/N(\boldsymbol{\theta}_i)$ becomes huge, but the right bound in (25) depends on the $\min$ of the training sample's ratios and thus is that of a non-outlier. Thus, picking G_t to get a robust right population minimizer does not *a priori* prevent the left population minimizer from being robust *as well*, a property that cannot hold for exponential, q-exponential nor deformed exponential families.

6 EXPERIMENTS

We report experiments on simulated data on four topics related to clustering: (a) the shape of the balls whose associated distortion is B_{G_t} in (20), (b) Voronoi diagrams associated to the cluster centers, (c) robustness, and (d) clustering with or without noise. For (a) through (d), the key parameter used from our theory is the divergence considered (Eq. (20)); we focus on the divergence associated to the 1D t-exponential measure in Table 2, which is a generalization of the Itakura-Saito divergence. In a domain of dimension > 1, the divergence we compute is just a sum of coordinate-wise 1D, scalar divergences, thereby mimicking a separable divergence, which is a common approach in ML. For (d), the clustering algorithm built on top of our distortions follows

[3] In fact, Lemma B provided in App shows it is a weak assumption to directly assume $N(\boldsymbol{\theta}) > 0$.

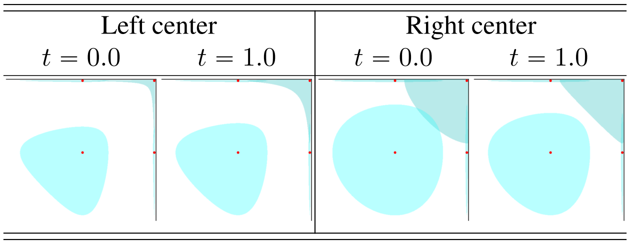


Figure 1: Information geometric balls for the 1D t-exponential (domain = $\mathbb{R}^2_{<0}$); each plot displays four balls whose centers are the same among plots. Balls are computed so that the *on-screen pixel* radius is fixed, so as not to get disproportionate balls between plots (see text for details).

the k-means blueprint, which consists, after initializing the cluster centers, in repeating the two steps until convergence:
[A] allocate points to clusters using distortions to centers (20);
[B] update cluster centers by computing the related population minimizers (Lemmata 5.1, 5.2).
Thus, as $t \rightarrow 1$, our clustering converges to the classical k-means with Bregman divergences (Banerjee et al., 2005b).

Information-geometric Balls An important question for clustering, especially when it comes to generalizing approaches based on Bregman divergences, is the shape of the corresponding information-geometric balls. Such balls are defined by a radius, a distortion and a center, generalizing the classical Euclidean balls whose distortion, the squared Euclidean distance, is a particular case of the Mahalanobis divergence. Generalized to Bregman divergences, the balls can adopt a variety of shapes, even becoming eventually non-convex when the center is on the left position of the Bregman divergence (Nock et al., 2008). In our case, Figure 1 shows examples of balls for the 1D t-exponential TEM, thus generalizing the Itakura-Saito balls (they appear for $t = 1$). One can remark that extending $t < 1$ allows for more "extreme" shapes, where balls are more "flattened", in particular when they are close to the quadrant's border (left center) or more "round" for the right center. Having increased diversity in ball shapes is good for clustering.

Voronoi diagrams An important structure for clustering is the Voronoi diagrams that partition the space in cells associated to a training data point being the closest center. Since the information geometric divergences (Bregman divergences or our B_{G_t} in (20)) are not symmetric in general, we have two types of Voronoi diagrams, a left and a right one depending on whether the cell's center is put in the left or right position in the corresponding divergence. There is a big difference between Voronoi diagrams associated to Bregman divergences (Boissonnat et al., 2010) and those associated with B_{G_t} in (20): the right Voronoi diagram is always affine with convex polyhedral cells for all Bregman divergences. In our case, this does not hold anymore and thus, we end up with two curved Voronoi diagrams (Figure 2 for $t < 1.0$).

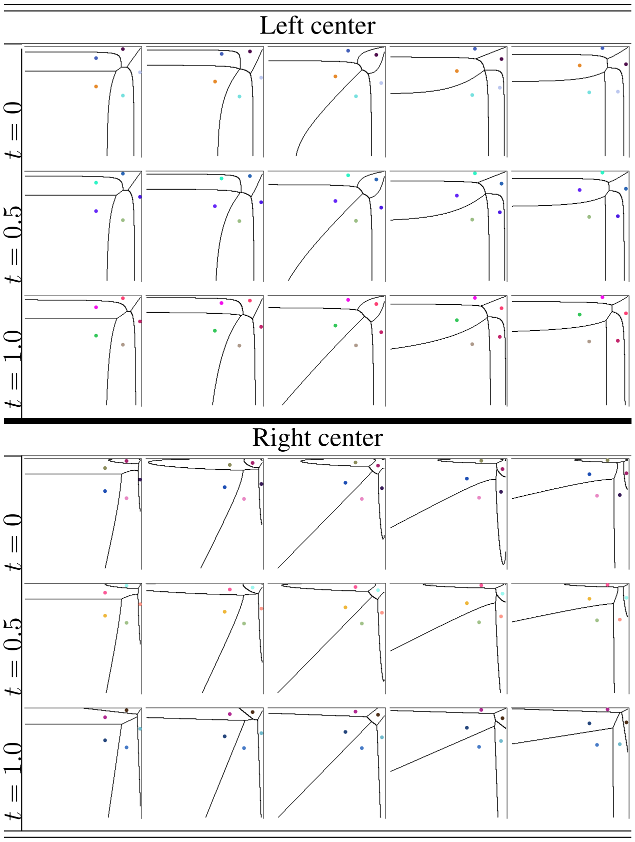


Figure 2: Voronoi diagrams associated to the left (top) and right (bottom) center of the B_{G_t} divergence of the 1D t-exponential (in $\mathbb{R}^2_{<0}$), where cell centers are the vertices of a rotating regular pentagon, for $t \in \{0, 0.5, 1\}$. For $t = 1$, the right Voronoi diagram is *affine*, but not for $t < 1$.

Robustness To analyze whether we can indeed observe improved robustness for $t \neq 1$ vs. $t = 1$, we have used the 1D t-exponential's left population minimizer. It is an interesting case because for $t = 1$, the divergence is Itakura-Saito divergence and its left population minimizer, the harmonic mean, is robust to outliers (See Section 5). Whether we can get improved robustness for $t \neq 1$ is displayed in Figure 3. Here, we choose a point close to the average, that we associated with a very heavy weight and then move away progressively the point by a constant vector in $\mathbb{R}^2$. The resulting trajectory of the outlier, in green, is picked at random. We then compute the trajectory of the population minimizer, in blue. One can observe that for $t = 1$, the center moves away with a segment length slowly decreasing, whereas, for $t = 0$, this length quickly decreases as the outlier moves far away, displaying improved robustness. Note that the robustness for $t = 1$ appears more clearly for a displacement of the outlier further away, which is not shown to keep the pictures readable. Another interesting phenomenon appears, not just from the standpoint of the distance of the new center to its original position, but also from the standpoint of the

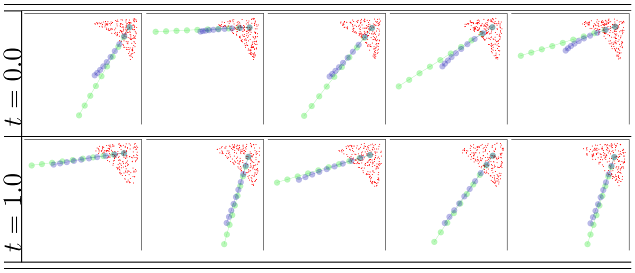


Figure 3: Outlier effect on the left population minimizer of the 1D t-exponential (domain = $\mathbb{R}^2_{<0}$) for two values ot t, on 5 random trials (columns). Clusters are generated by uniformly sampling 200 points in a B_F-ball whose *on-screen pixel* radius is fixed. A point close to the left population minimizer is chosen and treated as an outlier (green), with weight $5000\times$ that of non-outliers. We then move away the outlier with a fixed step size (in green) and compute the resulting cluster center (in blue, see text for details).

angle to its original position, as measured by a cone whose half lines cross the origin and go through the two centers (before and after max displacement): one can check that this angle is smaller for the $t = 0$ case. Equivalently, for $t = 1$, the center is not just dragged away according to a distance that is larger than for $t = 0$: it also follows more closely the trajectory of the outlier compared to $t = 0$.

Clustering with and without noise In this experiment, we test whether improved robustness can be translated to a better handling of noise. We treat noise as follows. We generate a fixed number of k clusters (and keep this value for clustering). The clusters are generated by uniform sampling in information-geometric balls with the left center, for $t \in \{0, 1\}$. To make the clusters unbalanced, one cluster has $20\times$ more points than the others. We then cluster using the left population minimizer using k-means type iterations (computing centers, reallocating points to clusters), and measure several metrics to assess the quality of clustering (see below). When there is noise, we generate it uniformly on the picture as an additional cluster. Noise thus biases clustering results *but* it is not taken into account for the measurement of the metrics. To explain it better, we compute three metrics: (I) at the end of clustering, we compute a distortion between the true clusters centers and those found, *excluding* the center of the noise cluster, using the following algorithm: we repeatedly compute the couple (theoretical center, learned center) that minimizes the average B_{G_t} divergence (where we permute the roles of the centers), and remove the theoretical center from the list – and eventually remove the learned center if there still exist learned centers (sometimes, clustering comes up with less than k clusters). We finally compute the average of those distances and report it as "$\overline{B}_F$"; (II) we compute the proportion of true clusters being split among learned clusters in such a way that less than $2/3$rd of the cluster belongs to a single learned cluster (we call these "true clusters that are split"). We do not use a larger proportion than .67 to authorize some of the learned clusters to scrap a minor proportion of the true clusters; we report this proportion as "p_{split}". Of course, we do not count the noise cluster in this computation; finally (III) for each true cluster, we compute the learned cluster with the largest fraction of the true cluster and count the remaining proportion of the true cluster as an error term; we compute the average of those errors over true clusters and denominate it as the "p_{err}". Table 3 summarizes the results obtained, where each statistic is computed over 50 runs, along with example clusterings. Modulo the fact that we treat our theoretical clusters as the ground truth for clustering (there could be some slight changes in optimal clusterings, especially in the "close" configuration), Table 3 confirms that choices $t \neq 1$ can improve clustering from the standpoint of all metrics, in particular when there is noise.

7 DISCUSSION

We split this discussion into three parts, from a focus on clustering to more general considerations on TEMs.

On clustering, one may remark that (scalar) f-means have intuitive properties, such as monotonicity (increasing an argument cannot decrease the mean), idempotence (the mean of the same repeated value is the value itself) and bounding (the mean is in between the min and max argument values). Our population minimizers *can* break these properties (unless $t = 1$): for example, the left population average of the 1D t-exponential TEM in Table 2 is monotonic and idempotent but does not meet the bounding constraint. Relaxing the constraints of the population minimizers outside those met by traditional means is not a bad thing, as ultimately the properties of a population minimizer depend on the distortion it is supposed to minimize in expectation. Also, as exemplified by our experiments, relaxing those properties can be beneficial. Ultimately, it can be a design choice to consider or tune *ex ante*: for example, assuming $t \in [0, 1]$, one needs $G_t(\min_i \theta_i) \leqslant 0, G_t(\max_i \theta_i) \geqslant 0$ to get bounding. One also has to keep in mind that clustering faces substantial impediments in terms of design choices (Kleinberg, 2002). Third, our experiments have made use of simple random (Forgy) initialization for the cluster centers. A better initialization with guarantees has been designed for clustering with Gaussians (Arthur and Vassilvitskii, 2007), extended to exponential families (Nock et al., 2008), and even to distortion classes without closed form for the population minimizers (Nielsen and Nock, 2015). Applying it to our setting is a promising direction.

Second, as we noted in Section 2, some previous work related to robust clustering has also put a focus on links with distributions, departing from both Bregman divergences and exponential families (Liu et al., 2012; Vemuri et al., 2010). In our case, our generalizations of exponential families to TEMs, which allows for improved robustness as $t \neq 1$, still

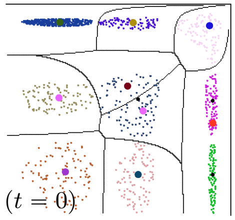
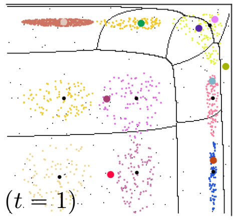
	close- 3×3		Clustering $t = 0$	 $t = 0.5$	 $t = 1.0$	far- 3×3	Clustering $t = 0$	 $t = 0.5$	 $t = 1.0$	
$p_{\text{noise}} = 0$	$\mathcal{B}_{t=0}$	p_{err}	0.43±0.29	0.43±0.29	$\underline{0.35\pm0.24}$	$\mathcal{B}_{t=0}$	0.57±0.29	$\underline{0.52\pm0.32}$	0.61±0.30	
		p_{split}	0.22±0.09	0.21±0.09	$\mathbf{0.17}\pm\mathbf{0.08}$		0.17±0.07	$\mathbf{0.13}\pm\mathbf{0.06}$	0.17±0.06	
		$\overline{B}_F$	2.00±0.82	$\underline{1.95\pm1.03}$	2.29±1.11		3.22±1.80	$\mathbf{2.93}\pm\mathbf{1.53}$	3.77±1.76	
	$\mathcal{B}_{t=1}$	p_{err}	$\underline{0.40\pm0.25}$	0.44±0.29	0.46±0.28	$\mathcal{B}_{t=1}$	0.62±0.30	0.52±0.30	$\underline{0.49\pm0.33}$	
		p_{split}	$\underline{0.19\pm0.09}$	0.22±0.10	0.21±0.09		0.16±0.07	0.16±0.06	0.16±0.08	
		$\overline{B}_F$	2.03±1.21	2.11±1.10	$\underline{1.86\pm1.19}$		2.42±1.15	$\underline{2.21\pm1.08}$	2.25±0.95	
	close- 3×3		Clustering $t = 0$	 $t = 0.5$	 $t = 1.0$	far- 3×3	Clustering $t = 0$	 $t = 0.5$	 $t = 1.0$	
$p_{\text{noise}} = 0.1$	$\mathcal{B}_{t=0}$	p_{err}	$\mathbf{0.28}\pm\mathbf{0.11}$	0.30±0.12	0.32±0.13	$\mathcal{B}_{t=0}$	$\mathbf{0.24}\pm\mathbf{0.16}$	0.29±0.22	0.37±0.25	
		p_{split}	$\mathbf{0.10}\pm\mathbf{0.08}$	0.14±0.05	0.15±0.07		0.06±0.06	$\mathbf{0.05}\pm\mathbf{0.06}$	0.08±0.07	
		$\overline{B}_F$	14.86±11.30	8.76±5.35	$\mathbf{4.77}\pm\mathbf{4.12}$		$\mathbf{6.57}\pm\mathbf{5.61}$	10.89±5.36	12.18±9.52	
	$\mathcal{B}_{t=1}$	p_{err}	0.35±0.19	0.29±0.11	$\underline{0.28\pm0.12}$	$\mathcal{B}_{t=1}$	$\underline{0.17\pm0.11}$	0.38±0.19	0.37±0.20	
		p_{split}	0.15±0.09	$\mathbf{0.09}\pm\mathbf{0.07}$	0.13±0.07		$\mathbf{0.03}\pm\mathbf{0.06}$	0.06±0.06	0.13±0.06	
		$\overline{B}_F$	$\underline{2.05\pm1.20}$	5.04±2.37	2.26±1.84		$\underline{4.26\pm2.15}$	5.27±1.96	$\underline{4.26\pm1.76}$	

Table 3: Clustering with the left population minimizer of 1D t-exponential distributions and the results of the corresponding clusterings for $t \in \{0, 0.5, 1\}$ in the form average±std-dev (average over 50 runs), without (top table) and with noise (10%, bottom table). Underlined values are the best among the three t choices and **bold faces** denote a significant winner in $t \in \{0, 0.5\}$ (best result) vs. $t = 1.0$ using a Gaussian test, p-val=.05. Pictures on the right give an example result on far-3×3 for $\mathcal{B}_{t=0}$, (clustering's t value indicated, true clusters shown using random colors with bigger black dots as their centers, Voronoi diagram displayed; learned centers in big coloured dots, see text).

come with a guarantee of "closeness" to exponential families. We provide a proof on a key parameter, the cumulant (15), and show that one can always come "as close as desired" to the exponential family case with $t \neq 1$. Such a result is relevant not just to numerical analysis at large: the cumulant is indeed the ID of a family of distributions in exponential families and it is not available in closed form for classical generalizations of exponential families that are q-exponential families or deformed exponential families. We let Θ denote the (open) set of natural parameters.

Theorem 7.1. *$\forall \boldsymbol{\theta} \in \Theta, \forall \varepsilon > 0, \exists t < 1 : |G_t(\boldsymbol{\theta}) - G_1(\boldsymbol{\theta})| \leqslant \varepsilon$.*

(Proof in App, Section VIII) As a consequence, we also get continuity in the neighborhood of the exponential family's case of the total mass of the TEM (Lemma 3.3) and of the convex conjugate of the cumulant.

Last, from a more general standpoint on TEMs, our general approach may seem close to the design of q-exponential families and even deformed exponential families – the knowledgeable reader will notice that our CODs technically look similar to escort distributions in the way we design them through (10), despite a normalization which belongs to the *divisive* normalization of distribution rather than the subtractive normalization of q-exponential families and deformed exponential families (Zhang and Wong, 2022). Classical escort distributions, however, appear independently of the q-exponential families or deformed exponential families: they do not belong to their axiomatization. In our case, they do, and the fact that we chose to somehow "mix" TEM and COD in the axiomatization of the TEM, by constraining the normalization of the COD, seems to yield technical conveniences not known for q-exponential families or even deformed exponential families, the first of which is the elegant closed form of the cumulant in (15). Beyond such technical conveniences appear some concrete advantages for clustering. Given the ubiquity of exponential families and Bregman divergences in ML, those advantages could bear fruitful applications in other ML areas.

8 CONCLUSION

In this paper, we introduce a new generalization of exponential families named tempered exponential measures, whose constrained maximum entropy design involves normalizing a dual instead of the measure itself as in the state-of-the-art generalization of exponential families (q-exponential families and deformed exponential families). Tempered exponential measures provide a generalization of Bregman divergences in the parameter space, which allows designing clustering with improved robustness properties compared to the classical k-means extended to exponential, q-exponential, or deformed exponential families.

Given the wide footprint of exponential families and Bregman divergences in ML and the fact that tempered exponential measures also provide new and general technical conveniences beyond the realm of clustering, more ML applications of this new tool are expected, as well as additional technical insights relevant to ML such as the information geometry of the parameter space.

ACKNOWLEDGMENTS

The authors warmly thank Frank Nielsen for numerous remarks on this material, and the reviewers for their constructive comments.

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.

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.

Amari, S.-I., Ohara, A., and Matsuzoe, H. (2012). Geometry of deformed exponential families: Invariant, dually-flat and conformal geometries. *Physica A: Statistical Mechanics and its Applications*, 391:4308–4319.

Amid, E., Warmuth, M. K., Anil, R., and Koren, T. (2019). Robust bi-tempered logistic loss based on Bregman divergences. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NeurIPS.

Arthur, D. and Vassilvitskii, S. (2007). k-means++ : the advantages of careful seeding. In *19 th SODA*, pages 1027 – 1035.

Banerjee, A., Guo, X., and Wang, H. (2005a). On the optimality of conditional expectation as a Bregman predictor. *IEEE Trans. IT*, 51:2664–2669.

Banerjee, A., Merugu, S., Dhillon, I., and Ghosh, J. (2005b). Clustering with Bregman divergences. *JMLR*, 6:1705–1749.

Barndorff-Nielsen, O. (1979). *Information and Exponential Families in Statistical Theory*. John Wiley.

Boissonnat, J.-D., Nielsen, F., and Nock, R. (2010). Bregman Voronoi diagrams. *DCG*, 44(2):281–307.

Bonnier, G. (1887). *Les Plantes des champs et des bois, excursions botaniques. Printemps, été, automne, hiver (in French)*. Hachette – Bibliothèque Nationale de France.

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.

Duchi, J. (Fall 2021). Lecture notes in information theory and statistics. http://www.snn.ru.nl/∼bertk/machinelearning/exponential_families.pdf.

Févotte, C., Bertin, N., and Durrieu, J.-L. (2009). Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis. *Neural Computation*, 21(3):793–830.

Flach, P.-A. (2012). *Machine Learning - The Art and Science of Algorithms that Make Sense of Data*. Cambridge University Press.

Fujisawa, H. and Eguchi, S. (2008). Robust parameter estimation with a small bias against heavy contamination. *J. of Multivariate Analysis*, 99:2053–2081.

Hastie, T., Tibshirani, R., and Friedman, J. (2002). *The Elements of Statistical Learning*. Springer Series in Statistics.

Janati, H., Muzellec, B., Peyré, G., and Cuturi, M. (2020). Entropic optimal transport between unbalanced Gaussian measures has a closed form. In *NeurIPS'20*.

Kleinberg, J.-M. (2002). An impossibility theorem for clustering. In *Advances in Neural Information Processing Systems*15*, pages 446–453.

Liu, M., Vemuri, B.-C., i. Amari, S., and Nielsen, F. (2012). Shape retrieval using hierarchical total Bregman soft clustering. *IEEE Trans.PAMI*, 34(12):2407–2419.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Trans. IT*, 28:129–136.

Naudts, J. (2004). Estimators, escort probabilities and phi-exponential families in statistical physics. *J. Ineq. Pure Applied Math.*, 5:162–177.

Naudts, J. (2011). *Generalized thermostatistics*. Springer.

Neal, R. (2004). Tutorial: Bayesian methods for machine learning. In *Advances in Neural Information Processing Systems*17*.

Nielsen, F. and Garcia, V. (2009). Statistical exponential families: A digest with flash cards. *CoRR*, abs/0911.4863.

Nielsen, F. and Nock, R. (2015). Total Jensen divergences: definition, properties and clustering. In *ICASSP15*, pages 2016–2020.

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

Nock, R., Cranko, Z., Menon, A.-K., Qu, L., and Williamson, R.-C. (2017). f-GANs in an information geometric nutshell. In *NIPS*30*.

Nock, R., Luosto, P., and Kivinen, J. (2008). Mixed Bregman clustering with approximation guarantees. In *Proc. of the 19 th ECML*, pages 154–169.

Nock, R., Nielsen, F., and Amari, S.-I. (2016). On conformal divergences and their population minimizers. *IEEE Trans. IT*, 62:1–12.

Paul, D., Chakraborty, S., Das, S., and Xu, J.-Q. (2021). Uniform concentration bounds toward a unified framework for robust clustering. In *NeurIPS'21*, pages 8307–8319.

Steinhaus, H. (1956). Sur la division des corps matériels en parties. *Bulletin Acad. Pol. Sc.*, 12:801–804.

Tsallis, C. (2009). *Introduction to nonextensive statistical mechanics*. Springer.

Vellal, A., Chakraborty, S., and Xu, J.-Q. (2022). Bregman power k-means for clustering exponential family data. In *ICML'22*, volume 162 of *Proceedings of Machine Learning Research*, pages 22103–22119. PMLR.

Vemuri, B.-C., Liu, M., i. Amari, S., and Nielsen, F. (2010). Total Bregman divergence and its application to DTI analysis. *IEEE Trans. on Medical Imaging*, 30:475–483.

Vigelis, R.-F. and Cavalcante, C.-C. (2011). On φ-families of probability distributions. *J. Theor. Probab.*, 21:1–15.

von Luxburg, U., Williamson, R.-C., and Guyon, I. (2012). Clustering: Science or art? In *Unsupervised and Transfer Learning - Workshop held at ICML 2011, Bellevue, Washington, USA, July 2, 2011*, volume 27 of *JMLR Proceedings*, pages 65–80. JMLR.org.

Zhang, J. and Wong, T.-K.-L. (2022). λ-deformation: A canonical framework for statistical manifolds of constant curvature. *Entropy*, 24(2):193.

Appendix

This is the appendix for AISTATS'23 paper ”Clustering above Exponential Families with Tempered Exponential Measures”, by E. Amid, R. Nock and M. K. Warmuth. To differentiate with the numberings in the main file, the numbering of Theorems, etc. is letter-based (A, B, ...).

Table of contents

Supplementary material on proofs

I Cheatsheet for t-functions, t-algebra and related functions

t-algebra Following Nivanen et al. (2003), we define

$$\begin{aligned}
x \oplus_t y &\doteq \log_t(\exp_t(x)\exp_t(y)) = x + y + (1-t)xy, && (26)\\
x \ominus_t y &\doteq \log_t \frac{\exp_t(x)}{\exp_t(y)} = \frac{x-y}{1+(1-t)y}, && (27)\\
x \otimes_t y &\doteq \exp_t(\log_t(x) + \log_t(y)) = \left(x^{1-t} + y^{1-t} - 1\right)_+^{\frac{1}{1-t}} \text{ if } x, y \geqslant 0 \text{ else undefined}, && (28)\\
x \oslash_t y &\doteq \exp_t(\log_t(x) - \log_t(y)) = \left(x^{1-t} - y^{1-t} + 1\right)_+^{\frac{1}{1-t}} \text{ if } x, y \geqslant 0 \text{ else undefined}. && (29)
\end{aligned}$$

t-functions $\log_t$ and $\exp_t$ satisfy

$$\begin{aligned}
\log_t'(z) &= z^{-t}, && (30)\\
\exp_t'(z) &= \exp_t^t(z), && (31)\\
(\log_t)^{*\prime}(z) &= \left(\frac{z}{t*}\right)^{-t*}, && (32)\\
(\exp_t)^{*\prime}(z) &= \exp_t(z). && (33)
\end{aligned}$$

For non-negative scalars $x, y \geqslant 0$, we also have

$$\begin{aligned}
\log_t x\, y &= \log_t x + x^{1-t} \log_t y\,,\\
\log_t \frac{x}{y} &= \log_t x - (\frac{x}{y})^{1-t} \log_t y\,.
\end{aligned} \tag{34}$$

General properties The t-functions and t-algebra have the interesting property that properties of the $t = 1$ functions transfer modulo the general rule that "classical arithmetic outside the function becomes t-arithmetic inside and vice-versa". For example:

$$\begin{aligned}
\frac{\exp_t(x)}{\exp_t(y)} &= \exp_t(x \ominus_t y), && (35)\\
\exp_t(x) \oslash_t \exp_t(y) &= \exp_t(x - y). && (36)
\end{aligned}$$

The t-functions also satisfy

$$\begin{aligned}
\exp_{\frac{1}{t*}}(z) &= \frac{1}{\exp_t(-z)}, && (37)\\
\log_{\frac{1}{t*}} z &= -\log_t \frac{1}{z} && (38)
\end{aligned}$$

$(\exp_t)^*(z)$ and $(\log_t)^*(z)$ are inverses of each other.

II Proof of Theorem 3.2

We first show the expression of $\tilde{p}_{t|\theta}$ (in the scalar case for natural parameters for readability); the proof is a generalization of the proof for the exponential family (See, e.g., Duchi (2021)). We first consider the case where $\tilde{p}_{t|\theta} = [\tilde{p}_{t|\theta}(x)]_{x \in \mathcal{X}}$ is a finite-dimensional vector. The solution to this problem can be obtained by introducing Lagrange multipliers $\theta \in \mathbb{R}$, $\lambda \in \mathbb{R}$, and $\nu \geqslant 0$ to enforce the constraints

$$\begin{aligned}
\tilde{p}_{t|\theta}(x) = \operatorname*{argmin}_{\tilde{p}} \Big\{ &- H_t(\tilde{p}) - \theta\big(\int x\, \tilde{p}(x)\, \mathrm{d}\xi(x) - \mu\big)\\
&+ \lambda\,\big(\int \tilde{p}(x)^{2-t}\, \mathrm{d}\xi(x) - 1\big) - \nu\, \tilde{p}(x)\Big\}\,,
\end{aligned} \tag{39}$$

where H_t is Tsallis' entropy, defined in (11) (main file). Setting the functional derivative with respect to $p(x)$ to zero yields

$$\log_t \tilde{p}(x) - \theta x + \lambda'\, \tilde{p}(x)^{1-t} - \nu = 0\,,$$

with $\lambda' = (2-t)\,\lambda$. Expanding the definition of $\log_t$, we can rewrite the equation as

$$(1 + (1-t)\,\lambda')\,\tilde{p}(x)^{1-t} = 1 + (1-t)\,\theta x + \nu'\,.$$

By the KKT conditions, $\nu' = (1-t)\,\nu$ is zero iff $1+(1-t)\,\theta x \geqslant 0$. We remark that for this to hold, we need $1+(1-t)\lambda' \geqslant 0$. We suppose it holds and then we will check that it does indeed hold. Using the definition of $\exp_t$, the equation becomes

$$\exp_t(\lambda')\,\tilde{p}(x) = \exp_t(\theta x), \tag{40}$$

which is thus the general form of a TEM. Denoting λ' by $G_t(\theta)$ yields the form of Eq. (14). Next, we show that the solution holds for any event space $\mathcal{X}$. For $\psi_t \doteq z \log_t z - \log_{t-1} z$ (originally defined in Amid et al. (2019)), we let

$$D_{\psi_t}(u, v) = u \log_t u - u \log_t v - \log_{t-1} u + \log_{t-1} v \tag{41}$$

denote the (scalar) Bregman divergence induced by ψ_t and by extension

$$D_{\psi_t}(\tilde{P}, \tilde{Q}) \doteq \int D_{\psi_t}(\tilde{p}(x), \tilde{q}(x))\,\mathrm{d}\xi.$$

Consider any $\tilde{P} \in \tilde{\mathcal{P}}_{t|\boldsymbol{\hbar}}$ with unnormalized density $\tilde{p}(x)$. We have

$$\begin{aligned}
H_t(\tilde{P}) &= -\int_{\mathcal{X}} \psi_t(\tilde{p}(x))\,\mathrm{d}\xi \\
&= -\int_{\mathcal{X}} \tilde{p}(x) \log_t \tilde{p}(x)\,\mathrm{d}\xi + \int_{\mathcal{X}} \tilde{p}(x) \log_t \tilde{p}_{t|\theta}(x)\,\mathrm{d}\xi - \int_{\mathcal{X}} \tilde{p}(x) \log_t \tilde{p}_{t|\theta}(x)\,\mathrm{d}\xi \\
&= -D_{\psi_t}(\tilde{P}, \tilde{P}_{t|\theta}) - \theta\hbar + \left(\int_{\mathcal{X}} \tilde{p}(x)\tilde{p}_{t|\theta}(x)^{1-t}\mathrm{d}\xi\right) G(\theta)\,.
\end{aligned}$$

By adding and subtracting $G(\theta)$ and refactoring the terms, we have

$$\begin{aligned}
H_t(\tilde{P}) &= -D_{\psi_t}(\tilde{P}, \tilde{P}_{t|\theta})(1 + (1-t)G(\theta)) - (\theta\hbar - G(\theta)) \\
&= -D_{\psi_t}(\tilde{P}, \tilde{P}_{t|\theta}) \exp_t(G_t(\theta))^{1-t} + H_t(\tilde{P}_{t|\theta})\,,
\end{aligned}$$

where we use the fact that $\mathsf{E}_{\tilde{P}}[\varphi(x)] = \mathsf{E}_{\tilde{P}_{t|\theta}}[\varphi(x)] = \hbar$ and $G_t(\theta) \geqslant -1/(1-t)$ by the fact that (40), the denomination $\lambda' \doteq G_t(\theta)$ and the normalization constraint of the dual COD, we obtain

$$G_t(\theta) \quad = \quad \log_t\left(\int \exp_t(\theta\varphi(x))^{2-t}\mathrm{d}\xi\right)^{\frac{1}{2-t}} = (\log_t)^* \int (\exp_t)^*(\theta\varphi(x))\mathrm{d}\xi, \tag{42}$$

which since $\log_t(z) \geqslant -1/(1-t)$, shows $G_t(\theta) \geqslant -1/(1-t)$ and confirms $1 + (1-t)\lambda' \geqslant 0$.

To finish up, we check that

$$\begin{aligned}
\frac{\partial G_t(\theta)}{\partial \theta} &= \left(\frac{\int (\exp_t)^*(\theta\varphi(x))\mathrm{d}\xi}{t*}\right)^{-t*} \cdot \int \varphi(x) \exp_t(\theta\varphi(x))\mathrm{d}\xi && (43) \\
&= \left(\frac{\int (\exp_t)^*(\theta\varphi(x))\mathrm{d}\xi}{t*}\right)^{-t*} \cdot \exp_t G_t(\theta) \cdot \hbar && (44) \\
&= \left(\frac{\int (\exp_t)^*(\theta\varphi(x))\mathrm{d}\xi}{t*}\right)^{-t*} \cdot \left(\frac{\int (\exp_t)^*(\theta\varphi(x))\mathrm{d}\xi}{t*}\right)^{t*} \cdot \hbar && (45) \\
&= \hbar, && (46)
\end{aligned}$$

as claimed. To show convexity of G_t, let us consider the more general case of $\boldsymbol{\theta} \in \mathbb{R}^d$ and define the *score* function as

$$s_{t|\boldsymbol{\theta}}(\boldsymbol{x}) \doteq \nabla \log_t \tilde{p}_{t|\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{\exp_t(G_t(\boldsymbol{\theta}))^{1-t}}\Big(\boldsymbol{\varphi}(\boldsymbol{x}) - \tilde{p}_{t|\boldsymbol{\theta}}(\boldsymbol{x})^{1-t}\,\nabla G_t(\boldsymbol{\theta})\Big). \tag{47}$$

We then have

$$\begin{aligned}
\nabla^2 G_t(\boldsymbol{\theta}) &= \frac{1}{\exp_t(G_t(\theta))^{1-t}}\left(\int \boldsymbol{\varphi}(x)\boldsymbol{\varphi}(x)^\top \tilde{p}_{t|\boldsymbol{\theta}}(x)^t\,\mathrm{d}\xi - \nabla G_t(\boldsymbol{\theta})\nabla G_t(\boldsymbol{\theta})^\top\right) && (48) \\
&= \exp_t(G_t(\theta))^{1-t} \int s_{t|\boldsymbol{\theta}}(x) s_{t|\boldsymbol{\theta}}(x)^\top \tilde{p}_{t|\boldsymbol{\theta}}(x)^t\,\mathrm{d}\xi \succcurlyeq \mathbf{0}, && (49)
\end{aligned}$$

which concludes the proof.

III Proof of Lemma 3.3

We get the result using the $\log_t$ entropy with two different derivations,

$$\begin{aligned}\mathbb{E}_{\tilde{p}_{t|\boldsymbol{\theta}}}\left[\log_t \tilde{p}_{t|\boldsymbol{\theta}}\right] &= \int \tilde{p}_{t|\boldsymbol{\theta}}\, \frac{1}{1-t}\left(\tilde{p}^{1-t}_{t|\boldsymbol{\theta}} - 1\right) \mathrm{d}\xi = \frac{1}{1-t}\left(1 - \mathrm{M}_t(\boldsymbol{\theta})\right)\\ &= \int \tilde{p}_{t|\boldsymbol{\theta}} \left(\boldsymbol{\theta}^\top \boldsymbol{\varphi} - \tilde{p}^{1-t}_{t|\hat{\boldsymbol{\theta}}}\, G_t(\boldsymbol{\theta})\right)\mathrm{d}\xi = \boldsymbol{\theta}^\top \boldsymbol{\hbar} - G_t(\boldsymbol{\theta}),\end{aligned}$$

and we identify the right-hand sides to get the statement of the Lemma. In the upmost derivation, we use the definition of $\mathrm{M}_t(\boldsymbol{\theta})$ and the fact that $\tilde{p}^{2-t}_{t|\boldsymbol{\theta}}$ sums to 1. In the bottommost derivation, we use the expression in (14) (main file) to identify the terms between the integrals and then simplify. We get $\mathrm{M}_t(\boldsymbol{\theta}) = 1 + (1-t)(G_t(\boldsymbol{\theta}) - \boldsymbol{\theta}^\top \boldsymbol{\hbar})$. If G_t is strictly convex differentiable, since by the relationship $\boldsymbol{\theta} = \nabla G_t^{-1}(\boldsymbol{\hbar})$ and convex duality, $G_t^\star(\boldsymbol{\hbar}) = \boldsymbol{\theta}^\top \boldsymbol{\hbar} - G_t(\boldsymbol{\theta})$,

$$\mathrm{M}_t(\boldsymbol{\theta}) = 1 + (1-t)(-G_t^\star(\boldsymbol{\hbar})) \quad (= \exp_t^{1-t}(-G_t^\star(\boldsymbol{\hbar}))). \tag{50}$$

Remark A. *The non-negativity of the total mass* M_t *gives us a non-trivial lowerbound for* G_t *and upperbound for* $G_t^\star$*:*

$$G_t(\boldsymbol{\theta}) \geqslant -\frac{1}{1-t} + \boldsymbol{\theta}^\top \boldsymbol{\hbar}, \tag{51}$$

$$G_t^\star(\boldsymbol{\hbar}) \leqslant \frac{1}{1-t}, \tag{52}$$

both of which become vacuous when $t \to 1$.

IV Proof of Theorem 4.1

Using the t-algebra and the definition of F_t in (19) (main file), we first get an integral-free expression:

$$\begin{aligned}F_t(\tilde{P}_{t|\hat{\boldsymbol{\theta}}} \| \tilde{P}_{t|\boldsymbol{\theta}}) &= \int f\left(\frac{\mathrm{d}\tilde{p}_{t|\hat{\boldsymbol{\theta}}}}{\mathrm{d}\xi} \oslash_t \frac{\mathrm{d}\tilde{p}_{t|\boldsymbol{\theta}}}{\mathrm{d}\xi}\right) \mathrm{d}\tilde{p}_{t|\boldsymbol{\theta}}\\ &= \int -\log_t\left(\exp_t(\hat{\boldsymbol{\theta}}^\top \boldsymbol{\varphi} \ominus_t G_t(\hat{\boldsymbol{\theta}})) \oslash_t \exp_t(\boldsymbol{\theta}^\top \boldsymbol{\varphi} \ominus_t G_t(\boldsymbol{\theta}))\right) \mathrm{d}\tilde{p}_{t|\boldsymbol{\theta}}\\ &= \int \left(\boldsymbol{\theta}^\top \boldsymbol{\varphi} \ominus_t G_t(\boldsymbol{\theta}) - \hat{\boldsymbol{\theta}}^\top \boldsymbol{\varphi} \ominus_t G_t(\hat{\boldsymbol{\theta}})\right) \mathrm{d}\tilde{p}_{t|\boldsymbol{\theta}}\\ &= \frac{\boldsymbol{\theta}^\top \boldsymbol{\mu} - \mathrm{M}_t(\boldsymbol{\theta}) G_t(\boldsymbol{\theta})}{1 + (1-t) G_t(\boldsymbol{\theta})} - \frac{\hat{\boldsymbol{\theta}}^\top \boldsymbol{\mu} - \mathrm{M}_t(\boldsymbol{\theta}) G_t(\hat{\boldsymbol{\theta}})}{1 + (1-t) G_t(\hat{\boldsymbol{\theta}})},\end{aligned}$$

and we then simplify the last expression using Lemma 3.3:

$$\begin{aligned}&\frac{\boldsymbol{\theta}^\top \boldsymbol{\mu} - \mathrm{M}_t(\boldsymbol{\theta}) G_t(\boldsymbol{\theta})}{1 + (1-t) G_t(\boldsymbol{\theta})} - \frac{\hat{\boldsymbol{\theta}}^\top \boldsymbol{\mu} - \mathrm{M}_t(\boldsymbol{\theta}) G_t(\hat{\boldsymbol{\theta}})}{1 + (1-t) G_t(\hat{\boldsymbol{\theta}})}\\ &\quad= \frac{\boldsymbol{\theta}^\top \boldsymbol{\mu} - (1 + (1-t)(G_t(\boldsymbol{\theta}) - \boldsymbol{\theta}^\top \boldsymbol{\mu})) G_t(\boldsymbol{\theta})}{1 + (1-t) G_t(\boldsymbol{\theta})} - \frac{\hat{\boldsymbol{\theta}}^\top \boldsymbol{\mu} - \mathrm{M}_t(\boldsymbol{\theta}) G_t(\hat{\boldsymbol{\theta}})}{1 + (1-t) G_t(\hat{\boldsymbol{\theta}})}\\ &\quad= \boldsymbol{\theta}^\top \boldsymbol{\mu} - G_t(\boldsymbol{\theta}) - \frac{\hat{\boldsymbol{\theta}}^\top \boldsymbol{\mu} - \mathrm{M}_t(\boldsymbol{\theta}) G_t(\hat{\boldsymbol{\theta}})}{1 + (1-t) G_t(\hat{\boldsymbol{\theta}})}\\ &\quad= \frac{\boldsymbol{\theta}^\top \boldsymbol{\mu} - G_t(\boldsymbol{\theta}) - \hat{\boldsymbol{\theta}}^\top \boldsymbol{\mu} + \left((1-t)(\boldsymbol{\theta}^\top \boldsymbol{\mu} - G_t(\boldsymbol{\theta})) + \mathrm{M}_t(\boldsymbol{\theta})\right) G_t(\hat{\boldsymbol{\theta}})}{1 + (1-t) G_t(\hat{\boldsymbol{\theta}})}\\ &\quad= \frac{\boldsymbol{\theta}^\top \boldsymbol{\mu} - G_t(\boldsymbol{\theta}) - \hat{\boldsymbol{\theta}}^\top \boldsymbol{\mu} + G_t(\hat{\boldsymbol{\theta}})}{1 + (1-t) G_t(\hat{\boldsymbol{\theta}})} = \frac{G_t(\hat{\boldsymbol{\theta}}) - G_t(\boldsymbol{\theta}) - (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top \boldsymbol{\mu}}{1 + (1-t) G_t(\hat{\boldsymbol{\theta}})}\\ &\quad\doteq B_{G_t}(\hat{\boldsymbol{\theta}} \| \boldsymbol{\theta}),\end{aligned}$$

which yields the statement of the Theorem.

Remark A. *We remark that* $F_t(\tilde{P}_{t|\hat{\boldsymbol{\theta}}} \| \tilde{P}_{t|\boldsymbol{\theta}})$ *is also equal to the Bregman divergence* $D_{\psi_t}(\tilde{P}_{t|\boldsymbol{\theta}} \| \tilde{P}_{t|\hat{\boldsymbol{\theta}}})$*, a connection also known to hold for exponential families' analysis where the KL divergence is both an* f*-divergence and a Bregman divergence.*

V Proof of Lemma 5.2

Recall that $\mathsf{T}_i(\boldsymbol{\theta}) \doteq G_t(\boldsymbol{\theta}_i) + (\boldsymbol{\theta} - \boldsymbol{\theta}_i)^\top \nabla G_t(\boldsymbol{\theta}_i)$ the value at $\boldsymbol{\theta}$ of the tangent hyperplane to G_t at $\boldsymbol{\theta}_i$. Denote for shot

$$\begin{aligned} N(\boldsymbol{\theta}) & \doteq & 1 + (1-t)\mathbb{E}_i[\mathsf{T}_i(\boldsymbol{\theta})], \quad (53) \\ D(\boldsymbol{\theta}) & \doteq & 1 + (1-t)G_t(\boldsymbol{\theta}). \quad (54) \end{aligned}$$

We then obtain the loss for the left population minimizer:

$$L_{\mathrm{l}}(\boldsymbol{\theta}) \quad = \quad \frac{1}{1-t} \cdot \left(1 - \frac{N}{D}(\boldsymbol{\theta})\right), \quad (55)$$

which immediately yields:

Lemma A. *$\boldsymbol{\theta}$ is a critical point of $L_{\mathrm{l}}(\boldsymbol{\theta})$ iff:*

$$N(\boldsymbol{\theta}) \cdot \nabla D(\boldsymbol{\theta}) \quad = \quad D(\boldsymbol{\theta}) \cdot \nabla N(\boldsymbol{\theta}). \quad (56)$$

Lemma B. *Suppose $\exists i : \mathrm{M}_t(\boldsymbol{\theta}_i) > 0$ and G_t is strictly convex or strictly concave. Then any critical point of $L_{\mathrm{l}}(\boldsymbol{\theta})$ has $N(\boldsymbol{\theta}) \neq 0$.*

Proof. Suppose otherwise. Note that unless we are in the degenerate case where all $\boldsymbol{\theta}_i$ are equal, $D(\boldsymbol{\theta}) > N(\boldsymbol{\theta})$ from the strict convexity of G_t[4] (we recall that $N(\boldsymbol{\theta})$ is the expected value of all tangent hyperplanes at all $\boldsymbol{\theta}_i$s, at $\boldsymbol{\theta}$, which thus sits strictly below the function). So Lemma A implies $\nabla N(\boldsymbol{\theta}) = \mathbf{0}$, which, after developing, is in fact

$$\mathbb{E}_i \nabla G_t(\boldsymbol{\theta}_i) \quad = \quad \mathbf{0}, \quad (57)$$

In addition to being a critical point, the condition $N(\boldsymbol{\theta}) \neq 0$ yields $\mathbb{E}_i[\mathsf{T}_i(\boldsymbol{\theta})] = -1/(1-t)$. Using the definition of T_i and simplifying with $\mathbb{E}_i \nabla G_t(\boldsymbol{\theta}_i) = \mathbf{0}$ then reveals

$$\mathbb{E}_i[G_t(\boldsymbol{\theta}_i) + (\boldsymbol{\theta} - \boldsymbol{\theta}_i)^\top \nabla G_t(\boldsymbol{\theta}_i)] = \mathbb{E}_i[G_t(\boldsymbol{\theta}_i) - \boldsymbol{\theta}_i^\top \nabla G_t(\boldsymbol{\theta}_i)] = -\frac{1}{1-t},$$

which, using Lemma 3.3 (main file) reveals that the population for which $\boldsymbol{\theta}$ is a minimizer necessarily has

$$\begin{aligned} \mathbb{E}_i[\mathrm{M}_t(\boldsymbol{\theta}_i)] & = & 1 + (1-t)\mathbb{E}_i[G_t(\boldsymbol{\theta}_i) - \boldsymbol{\theta}_i^\top \nabla G_t(\boldsymbol{\theta}_i)] \\ & = & 1 - 1 = 0. \end{aligned}$$

Since M_t is non-negative, this leads to a contradiction with the assumption of the Lemma. □

The question is then whether such a critical point can be a population minimizer, and even more, if it is unique. Answering the first question comes from the Hessian of the loss.

Lemma C. *Removing the argument $\boldsymbol{\theta}$ for readability, we have:*

$$\mathrm{H}L_{\mathrm{l}} \quad = \quad \frac{1-t}{D^2} \cdot \left(\nabla \mathbb{E}_i[\mathsf{T}_i] \nabla G_t^\top + \nabla G_t \nabla \mathbb{E}_i[\mathsf{T}_i]^\top\right) - \frac{2(1-t)N}{D^3} \cdot \nabla G_t \nabla G_t^\top + \frac{N}{D^2} \cdot \mathrm{H}G_t. \quad (58)$$

Proof. Denote for short $\delta_{ij} \doteq [\nabla G_t(\boldsymbol{\theta}_i)]_j$ and $\nabla_j \doteq [\nabla G_t(\boldsymbol{\theta})]_j$. We check that we have $(\partial/\partial\theta'_k)\mathbb{E}_i[\mathsf{T}_i] = \delta_{ik}$, so we have

$$\begin{aligned} & \frac{\partial}{\partial \theta'_j}[\nabla_{\boldsymbol{\theta}} L_{\mathrm{l}}]_k \\ & = \frac{\partial}{\partial \theta'_j}\left(\frac{(1+(1-t)\mathbb{E}_i[\mathsf{T}_i])\nabla_k}{(1+(1-t)G_t(\boldsymbol{\theta}))^2} - \frac{\mathbb{E}_i \delta_{ik}}{1+(1-t)G_t(\boldsymbol{\theta})}\right) \\ & = \frac{\left\{ \begin{array}{c} ((1-t)\delta_{ij}\nabla_k + (1+(1-t)\mathbb{E}_i[\mathsf{T}_i])\nabla_{kj})\,(1+(1-t)G_t(\boldsymbol{\theta}))^2 \\ -2(1-t)\,((1+(1-t)\mathbb{E}_i[\mathsf{T}_i])\nabla_k)\,(1+(1-t)G_t(\boldsymbol{\theta}))\nabla_j \end{array} \right\}}{(1+(1-t)G_t(\boldsymbol{\theta}))^4} + \frac{(1-t)\mathbb{E}_i\delta_{ik}\nabla_j}{(1+(1-t)G_t(\boldsymbol{\theta}))^2} \\ & = \frac{(1-t)\delta_{ij}\nabla_k + N\nabla_{kj}}{D^2} - \frac{2(1-t)N\nabla_k\nabla_j}{D^3} + \frac{(1-t)\mathbb{E}_i\delta_{ik}\nabla_j}{D^2} \\ & = \frac{(1-t)}{D^2} \cdot (\delta_{ij}\nabla_k + \nabla_j\delta_{ik}) - \frac{2(1-t)N}{D^3} \cdot \nabla_k\nabla_j + \frac{N}{D^2} \cdot \nabla_{kj}; \end{aligned}$$

noting our convention yields $\nabla_{kj} = [\mathrm{H}G_t]_{jk}$, we get the statement of the Lemma. □

[4] If strictly concave, $D(\boldsymbol{\theta}) < N(\boldsymbol{\theta})$, which yields to the same result.

The next one introduces the convexity of G_t to elicit the nature of the critical points.:

Lemma D. *At any critical point of L_{l}, the convexity of L_{l} is the same as the convexity of G_t iff $N \geqslant 0$ (and it is opposed, meaning convex↔concave, otherwise).*

Proof. Using Lemma A, (58) simplifies to

$$\begin{aligned} \mathrm{H}L_{\mathrm{l}} &= \frac{(1-t)N}{D^3} \cdot \left(\nabla G_t \nabla G_t^\top + \nabla G_t \nabla G_t^\top\right) - \frac{2(1-t)N}{D^3} \cdot \nabla G_t \nabla G_t^\top + \frac{N}{D^2} \cdot \mathrm{H}G_t \\ &= \frac{N}{D^2} \cdot \mathrm{H}G_t, \end{aligned}$$

yielding the statement of the Lemma. □

Hence, if $N \geqslant 0$, all critical points are population minimizers. In the next Lemma, we show a condition for unicity.

Lemma E. *Suppose G_t is strictly convex. Any optimum of L_{l} is unique.*

Proof. Let us consider any two such minimizers $\boldsymbol{\theta}', \boldsymbol{\theta}''$. We thus have simultaneously from Lemma A:

$$N(\boldsymbol{\theta}') \cdot \nabla D(\boldsymbol{\theta}') = D(\boldsymbol{\theta}') \cdot \nabla N(\boldsymbol{\theta}') = (1-t) D(\boldsymbol{\theta}') \cdot \mathbb{E}_i \nabla G_t(\boldsymbol{\theta}_i), \quad (59)$$

$$N(\boldsymbol{\theta}'') \cdot \nabla D(\boldsymbol{\theta}'') = D(\boldsymbol{\theta}'') \cdot \nabla N(\boldsymbol{\theta}'') = (1-t) D(\boldsymbol{\theta}'') \cdot \mathbb{E}_i \nabla G_t(\boldsymbol{\theta}_i), \quad (60)$$

$$\frac{\frac{D(\boldsymbol{\theta}')-1}{1-t} - \frac{N(\boldsymbol{\theta}')-1}{1-t}}{D(\boldsymbol{\theta}')} = L_{\mathrm{l}}(\boldsymbol{\theta}') = L_{\mathrm{l}}(\boldsymbol{\theta}'') = \frac{\frac{D(\boldsymbol{\theta}'')-1}{1-t} - \frac{N(\boldsymbol{\theta}'')-1}{1-t}}{D(\boldsymbol{\theta}'')}. \quad (61)$$

We note (61) is equivalent, after simplification, to

$$\frac{N(\boldsymbol{\theta}')}{D(\boldsymbol{\theta}')} = \frac{N(\boldsymbol{\theta}'')}{D(\boldsymbol{\theta}'')}. \quad (62)$$

Also, (59) and (60) bring:

$$\frac{N(\boldsymbol{\theta}')}{D(\boldsymbol{\theta}')} \cdot \left.\frac{\partial D(\boldsymbol{\theta})}{\partial \theta_i}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}'} = \frac{N(\boldsymbol{\theta}'')}{D(\boldsymbol{\theta}'')} \cdot \left.\frac{\partial D(\boldsymbol{\theta})}{\partial \theta_i}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}''}, \forall i \in [d],$$

and thus simplifies with (62) to $\left.\frac{\partial D(\boldsymbol{\theta})}{\partial \theta_i}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}'} = \left.\frac{\partial D(\boldsymbol{\theta})}{\partial \theta_i}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}''}, \forall i \in [d]$, or in gradient form after using the definition of D and simplifying,

$$\nabla G_t(\boldsymbol{\theta}') = \nabla G_t(\boldsymbol{\theta}''),$$

but since G_t is strictly convex, ∇G_t is bijective and this implies $\boldsymbol{\theta}' = \boldsymbol{\theta}''$, and completes the proof of the Lemma. □

Folding together all Lemmata, we get that if G_t is strictly convex, $\boldsymbol{\theta}$ is the unique left population minimizer if $\nabla L_{\mathrm{l}}(\boldsymbol{\theta}) = \mathbf{0}$ and $N(\boldsymbol{\theta}) > 0$. (56) can be reformulated as:

$$\nabla G_t(\boldsymbol{\theta}) = \frac{D(\boldsymbol{\theta})}{N(\boldsymbol{\theta})} \cdot \mathbb{E}_i \nabla G_t(\boldsymbol{\theta}_i), \quad (63)$$

Now, define function α:

$$\alpha(\boldsymbol{\theta}) = \frac{D(\boldsymbol{\theta})}{N(\boldsymbol{\theta})}. \quad (64)$$

We note from (63) that $\alpha_* = \alpha(\boldsymbol{\theta}_{\mathrm{l}})$ and we also note from (55) that we also have

$$L_{\mathrm{l}}(\boldsymbol{\theta}) = \frac{1}{1-t} \cdot \left(1 - \frac{1}{\alpha(\boldsymbol{\theta})}\right), \quad (65)$$

so we conclude, if $t \leqslant 1$,

$$\alpha_* \leqslant \min_i \alpha(\boldsymbol{\theta}_i), \quad (66)$$

which provides a convenient upperbound which, in addition to the fact that $\alpha_* \geqslant 1$, provides a convenient initialization interval for a line search of α_*.

VI Proof of Lemma 5.3

We recall the right population minimizer:

$$\boldsymbol{\theta}_{\mathrm{r}}^{\mathrm{old}} = \mathbb{E}_i\left[\frac{1}{1+(1-t)G_t(\boldsymbol{\theta}_i)}\cdot\boldsymbol{\theta}_i\right]. \tag{67}$$

If we add a new point $\boldsymbol{\theta}_*$ with a weight ε and downweight the other points' weights proportionally, the new right population minimizer is

$$\boldsymbol{\theta}_{\mathrm{r}}^{\mathrm{new}} = (1-\varepsilon)\cdot\boldsymbol{\theta}_{\mathrm{r}}^{\mathrm{old}}+\varepsilon\cdot\frac{1}{1+(1-t)G_t(\boldsymbol{\theta}_*)}\cdot\boldsymbol{\theta}_*, \tag{68}$$

and so

$$\boldsymbol{\theta}_{\mathrm{r}}^{\mathrm{new}}-\boldsymbol{\theta}_{\mathrm{r}}^{\mathrm{old}} = \varepsilon\cdot\underbrace{\left(\frac{1}{1+(1-t)G_t(\boldsymbol{\theta}_*)}\cdot\boldsymbol{\theta}_*-\boldsymbol{\theta}_{\mathrm{r}}^{\mathrm{old}}\right)}_{\doteq\boldsymbol{z}(\boldsymbol{\theta}_*)}, \tag{69}$$

and so if $G_t(\boldsymbol{\theta}_i) = \Omega(\|\boldsymbol{\theta}_*\|)$, $\|\boldsymbol{z}(\boldsymbol{\theta}_*)\| \leqslant (1/(1+(1-t)G_t(\boldsymbol{\theta}_*)))\cdot\|\boldsymbol{\theta}_*\|+\|\boldsymbol{\theta}_{\mathrm{r}}^{\mathrm{old}}\| = O(1)$ and the right population minimizer is robust.

VII Proof of Lemma 5.4

Denote for short

$$\boldsymbol{\theta}_{\mathrm{l}}^{\mathrm{old}} \doteq \arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) \doteq \frac{\mathbb{E}_i[D_{G_t}(\boldsymbol{\theta}\|\boldsymbol{\theta}_i)]}{1+(1-t)G_t(\boldsymbol{\theta})} \tag{70}$$

$$\boldsymbol{\theta}_{\mathrm{l}}^{\mathrm{new}} \doteq \arg\min_{\boldsymbol{\theta}} L^{\varepsilon}(\boldsymbol{\theta}) \doteq \frac{(1-\varepsilon)\mathbb{E}_i[D_{G_t}(\boldsymbol{\theta}\|\boldsymbol{\theta}_i)]+\varepsilon D_{G_t}(\boldsymbol{\theta}\|\boldsymbol{\theta}_*)}{1+(1-t)G_t(\boldsymbol{\theta})}. \tag{71}$$

Also, we let $\nabla_\varepsilon \doteq (1-\varepsilon)\cdot\nabla G_t(\boldsymbol{\theta}_{\mathrm{l}}^{\mathrm{old}})+\varepsilon\cdot\nabla G_t(\boldsymbol{\theta}_*)$ and $\mathbb{E}(\boldsymbol{\theta}) \doteq \mathbb{E}_i[\mathsf{T}_i(\boldsymbol{\theta})]$, $\mathbb{E}_\varepsilon(\boldsymbol{\theta}) \doteq (1-\varepsilon)\mathbb{E}(\boldsymbol{\theta})+\varepsilon\mathsf{T}_*(\boldsymbol{\theta})$, where by extension $\mathsf{T}_*(\boldsymbol{\theta}) \doteq G_t(\boldsymbol{\theta}_*)+(\boldsymbol{\theta}-\boldsymbol{\theta}_*)^\top\nabla G_t(\boldsymbol{\theta}_*)$. We also use the following Taylor expansion:

(A) $\nabla G_t(\boldsymbol{\theta}_{\mathrm{l}}^{\mathrm{new}})-\nabla G_t(\boldsymbol{\theta}_{\mathrm{l}}^{\mathrm{old}}) = \mathrm{H}_t(\boldsymbol{\theta}_{\mathrm{l}}^{\mathrm{new}}-\boldsymbol{\theta}_{\mathrm{l}}^{\mathrm{old}}) \doteq \varepsilon\mathrm{H}_t\boldsymbol{z}_t(\boldsymbol{\theta}_*)$, where H_t is a value of the Hessian of G_t.

Finally, we note

$$\nabla L^{\varepsilon}(\boldsymbol{\theta}) = \frac{1}{(1+(1-t)G_t(\boldsymbol{\theta}))^2}\cdot\left((1+(1-t)\mathbb{E}_\varepsilon(\boldsymbol{\theta}))\cdot\nabla G_t(\boldsymbol{\theta})-(1+(1-t)G_t(\boldsymbol{\theta}))\cdot\nabla_\varepsilon\right). \tag{72}$$

Using the definition of $\boldsymbol{\theta}_{\mathrm{l}}^{\mathrm{new}}$ and (A), we get to:

$$\begin{aligned}\mathbf{0} &= (1+(1-t)G_t(\boldsymbol{\theta}_{\mathrm{l}}^{\mathrm{new}}))^2\cdot\nabla L^{\varepsilon}(\boldsymbol{\theta}_{\mathrm{l}}^{\mathrm{new}})\\ &= ((1+(1-t)\mathbb{E}_\varepsilon(\boldsymbol{\theta}_{\mathrm{l}}^{\mathrm{new}}))\cdot\nabla G_t(\boldsymbol{\theta}_{\mathrm{l}}^{\mathrm{new}})-(1+(1-t)G_t(\boldsymbol{\theta}_{\mathrm{l}}^{\mathrm{new}}))\cdot\nabla_\varepsilon)\\ &= \Big((1+(1-t)\mathbb{E}_\varepsilon(\boldsymbol{\theta}_{\mathrm{l}}^{\mathrm{new}}))\cdot(\nabla G_t(\boldsymbol{\theta}_{\mathrm{l}}^{\mathrm{old}})+\varepsilon\mathrm{H}_t\boldsymbol{z}_t(\boldsymbol{\theta}_*))-(1+(1-t)G_t(\boldsymbol{\theta}_{\mathrm{l}}^{\mathrm{new}}))\cdot\nabla_\varepsilon\Big).\end{aligned} \tag{73}$$

We get the relationship satisfied by the influence function:

$$\begin{aligned}&\boldsymbol{z}_t(\boldsymbol{\theta}_*)\\ &= \frac{1}{\varepsilon}\cdot\mathrm{H}_t^{-1}\left(\frac{1+(1-t)G_t(\boldsymbol{\theta}_{\mathrm{l}}^{\mathrm{new}})}{1+(1-t)\mathbb{E}_\varepsilon(\boldsymbol{\theta}_{\mathrm{l}}^{\mathrm{new}})}\cdot\nabla_\varepsilon-\nabla G_t(\boldsymbol{\theta}_{\mathrm{l}}^{\mathrm{old}})\right)\\ &= \mathrm{H}_t^{-1}\left(\frac{1+(1-t)G_t(\boldsymbol{\theta}_{\mathrm{l}}^{\mathrm{new}})}{1+(1-t)\mathbb{E}_\varepsilon(\boldsymbol{\theta}_{\mathrm{l}}^{\mathrm{new}})}\cdot\nabla G_t(\boldsymbol{\theta}_*)+\left(\frac{1-\varepsilon}{\varepsilon}\cdot\frac{1+(1-t)G_t(\boldsymbol{\theta}_{\mathrm{l}}^{\mathrm{new}})}{1+(1-t)\mathbb{E}_\varepsilon(\boldsymbol{\theta}_{\mathrm{l}}^{\mathrm{new}})}-1\right)\cdot\nabla G_t(\boldsymbol{\theta}_{\mathrm{l}}^{\mathrm{old}})\right)\\ &= \mathrm{H}_t^{-1}\left(\alpha(\boldsymbol{\theta}_{\mathrm{l}}^{\mathrm{new}})\cdot\nabla G_t(\boldsymbol{\theta}_*)+\left(\frac{1-\varepsilon}{\varepsilon}\cdot\alpha(\boldsymbol{\theta}_{\mathrm{l}}^{\mathrm{new}})-1\right)\cdot\nabla G_t(\boldsymbol{\theta}_{\mathrm{l}}^{\mathrm{old}})\right).\end{aligned} \tag{74}$$

We know from (66) that $\alpha(\boldsymbol{\theta}_1^{\mathrm{new}})$ cannot diverge as a function of the outlier $\boldsymbol{\theta}_*$, so we end up with

$$\boldsymbol{z}_t(\boldsymbol{\theta}_*) \quad = \quad Q \cdot \mathrm{H}_t^{-1}\nabla G_t(\boldsymbol{\theta}_*) + \mathrm{H}_t^{-1}\boldsymbol{v}, \tag{75}$$

where $Q \ll \infty, \|\boldsymbol{v}\| \ll \infty$. If we compute the f-mean for G_t and its influence function, then we get this time:

$$\begin{aligned}\boldsymbol{z}_1(\boldsymbol{\theta}_*) \quad &= \quad \frac{1}{\varepsilon} \cdot \mathrm{H}_1^{-1}\nabla_\varepsilon & (76)\\ &= \quad \mathrm{H}_1^{-1}\left(\nabla G_t(\boldsymbol{\theta}_*) + \frac{1-\varepsilon}{\varepsilon} \cdot \mathbb{E}_i[\nabla G_t(\boldsymbol{\theta}_i)]\right) & (77)\\ &= \quad \mathrm{H}_1^{-1}\nabla G_t(\boldsymbol{\theta}_*) + \mathrm{H}_1^{-1}\boldsymbol{v}_1, & (78)\end{aligned}$$

where $\|\boldsymbol{v}_1\| \ll \infty$. We see that $\boldsymbol{z}_t$ has bounded norm iff $\boldsymbol{z}_1$ does so, which proves the statement of the Lemma.

Remark A. *The strong convexity argument is here just to handle the influence of the Hessian via its minimal eigenvalue in* (75) *and* (78)*. We could add (realistic) assumptions on the training sample's domain to replace the strong convexity argument by strict convexity.*

VIII Proof of Theorem 7.1

We proceed in three steps.

Step 1: $\forall \boldsymbol{\theta}, \forall \varepsilon > 0, \exists t < 1 : G_t(\boldsymbol{\theta}) \geqslant G_1(\boldsymbol{\theta}) - \varepsilon$. We rely on the inequalities:[5]

$$\begin{aligned}\forall t \in [0,1], \forall z \geqslant 0, \log(z) \quad &\leqslant \quad (\log_t)^*(z), & (79)\\ \exists a, b \in \mathbb{R} \text{ s.t. } \forall t \in [0,1], \forall z \in \mathbb{R}, \underbrace{(1-(1-t)g_{a,b}(z))_+}_{=\exp_t^{1-t}(-g(z))} \exp(z) \quad &\leqslant \quad (\exp_t)^*(z), & (80)\end{aligned}$$

where $g_{a,b}(z) \doteq az^2 - bz + 1$. From (79) and (80), we get the inequalities in:

$$\begin{aligned}G_t(\boldsymbol{\theta}) \quad &= \quad (\log_t)^* \int (\exp_t)^* \left(\boldsymbol{\theta}^\top \boldsymbol{\varphi}\right) \mathrm{d}\xi\\ &\geqslant \quad \log \int (\exp_t)^* \left(\boldsymbol{\theta}^\top \boldsymbol{\varphi}\right) \mathrm{d}\xi\\ &\geqslant \quad \log \int \exp_t^{1-t}\left(-a(\boldsymbol{\theta}^\top \boldsymbol{\varphi})^2 + b(\boldsymbol{\theta}^\top \boldsymbol{\varphi}) - 1\right) \exp\left(\boldsymbol{\theta}^\top \boldsymbol{\varphi}\right) \mathrm{d}\xi\\ &\quad = \log \int \left[t + (1-t)\left(-a(\boldsymbol{\theta}^\top \boldsymbol{\varphi})^2 + b(\boldsymbol{\theta}^\top \boldsymbol{\varphi})\right)\right]_+ \exp\left(\boldsymbol{\theta}^\top \boldsymbol{\varphi}\right) \mathrm{d}\xi\\ &= \quad G_1(\boldsymbol{\theta}) + \log \int \left[t + (1-t)\left(-a(\boldsymbol{\theta}^\top \boldsymbol{\varphi})^2 + b(\boldsymbol{\theta}^\top \boldsymbol{\varphi})\right)\right]_+ \exp\left(\boldsymbol{\theta}^\top \boldsymbol{\varphi} - G_1(\boldsymbol{\theta})\right) \mathrm{d}\xi\\ &\doteq \quad G_1(\boldsymbol{\theta}) + \log Q_t(\boldsymbol{\theta}), & (81)\end{aligned}$$

with $Q_t(\boldsymbol{\theta}) \doteq \mathbb{E}_1\left[f_t(\boldsymbol{\theta}^\top \boldsymbol{\varphi})\right]$ ($\mathbb{E}_1$ indicating the expectation for $t = 1$, *i.e.*, the exponential family) and

$$f_t(z) \quad \doteq \quad [t + (1-t) \cdot z\,(b - az)]_+ \,. \tag{82}$$

Since f_t does not take negative values, for any $0 \leqslant \delta < 1$, if we let $\mathcal{X}_\delta \doteq \{\boldsymbol{x} \in \mathcal{X} : f_t(\boldsymbol{\theta}^\top \boldsymbol{\varphi}(\boldsymbol{x})) \geqslant \delta\}$, then we have, for μ_1 the probability measure associated to $t = 1$,

$$Q_t(\boldsymbol{\theta}) \quad \geqslant \quad \delta \cdot \mu_1(\mathcal{X}_\delta). \tag{83}$$

Also, for any $a, b, z \in \mathbb{R}$, $f_1(z) = 1$ and $f_t(z)$ is continuous in t and z so

$$\forall z_* > 0, \forall 0 \leqslant \delta < 1, \exists t < 1 \text{ s.t. } f_t([-z_*, z_*]) \subseteq [\delta, +\infty),$$

[5] The proofs, at the end of the proof of Theorem 7.1, elicit a, b.

and for any applicable $t_\delta < 1$, any $t \in [t_\delta, 1)$ is also valid. This implies that $\forall 0 \leqslant \delta < 1$, we can always find $t_\delta < 1$ close enough to 1 such that $\mu_1(\mathcal{X}_\delta) \geqslant \delta$ by picking z_* large enough. So, for any $0 \leqslant \delta < 1$, we can find t_δ such that $Q_t(\boldsymbol{\theta}) \geqslant \delta \cdot \delta = \delta^2$, and if we choose $\delta \doteq \exp(-\varepsilon/2)$, then $\log Q_t(\boldsymbol{\theta}) \geqslant -\varepsilon$ and considering (81), we obtain:

$$\forall \boldsymbol{\theta}, \forall \varepsilon > 0, \exists t < 1 : G_t(\boldsymbol{\theta}) \geqslant G_1(\boldsymbol{\theta}) - \varepsilon, \tag{84}$$

i.e., we have completed the proof of **Step 1**.

Step 2: $\forall \boldsymbol{\theta}, \forall \varepsilon > 0, \exists t < 1 : G_t(\boldsymbol{\theta}) \leqslant G_1(\boldsymbol{\theta}) + \varepsilon$. We rely on the inequalities:[6]

$$\begin{aligned} \forall t \in [0,1], \forall z \geqslant 0, (\log_t)^*(z) &\leqslant \log(z) - u(t)z + v(t)(1 + \log^2 z + (z-1)^2), & (85)\\ (\exp_t)^*(z) &\leqslant \exp(z), \forall t \leqslant 1, \forall z \in \mathbb{R}, & (86)\end{aligned}$$

where $u(t), v(t)$ are two continuous functions of t satisfying $u(1) = v(1) = 0$. $\log_t$ being non-decreasing, we get with (86) the first inequality of

$$\begin{aligned} G_t(\boldsymbol{\theta}) &\doteq (\log_t)^* \int (\exp_t)^* \left(\boldsymbol{\theta}^\top \boldsymbol{\varphi}\right) \mathrm{d}\xi \\ &\leqslant (\log_t)^* \int \exp \left(\boldsymbol{\theta}^\top \boldsymbol{\varphi}\right) \mathrm{d}\xi \\ &\leqslant \log \int \exp \left(\boldsymbol{\theta}^\top \boldsymbol{\varphi}\right) \mathrm{d}\xi - u(t) \cdot \int \exp \left(\boldsymbol{\theta}^\top \boldsymbol{\varphi}\right) \mathrm{d}\xi \\ &\quad + v(t) \cdot \left(1 + \left(\log \int \exp \left(\boldsymbol{\theta}^\top \boldsymbol{\varphi}\right) \mathrm{d}\xi\right)^2 + \left(\int \exp \left(\boldsymbol{\theta}^\top \boldsymbol{\varphi}\right) \mathrm{d}\xi - 1\right)^2\right) \\ &= G_1(\boldsymbol{\theta}) \underbrace{- u(t) \cdot \exp(G_1(\boldsymbol{\theta})) + v(t) \cdot \left(1 + G_1^2(\boldsymbol{\theta}) + (\exp(G_1(\boldsymbol{\theta})) - 1)^2\right)}_{\doteq h_t(G_1(\boldsymbol{\theta}))}. \end{aligned} \tag{87}$$

We have $h_1 = 0$ and h_t is continuous in t, so $\forall z \in \mathbb{R}, \forall \varepsilon > 0, \exists t < 1$ close enough to 1 such that $h_t(z) \leqslant \varepsilon$, implying, for $z \doteq G_1(\boldsymbol{\theta})$, $G_t(\boldsymbol{\theta}) \leqslant G_1(\boldsymbol{\theta}) + \varepsilon$. This completes the proof of **Step 2**.

We then check that we can simultaneously get **Step 1** and **Step 2** as both depend on choosing $t < 1$ close enough to 1. What remains is then:

Step 3: we show (79), (80), (85), (86).

↪ We first prove (79) and define

$$\Delta(z) \doteq (\log_t)^*(z) - \log(z).$$

We have

$$\Delta'(z) = ((2-t)z)^{-\frac{1}{2-t}} - \frac{1}{z}. \tag{88}$$

Δ' zeroes for $z_* = (2-t)^{1/(1-t)}$ (which is the global minimum for Δ), for which $(\log_t)^*(z_*) = 1$. Since $\log(1+z) \leqslant z, \forall z > 0$, we get by picking $z = 1 - t$ and reorganizing $\log z_* \leqslant 1$, and thus $\Delta(z_*) \geqslant 0$, and since z_* is the global minimum of Δ, yields (79).

↪ Since $(\log_t)^*$ and $(\exp_t)^*$ are inverses of each other, (86) follows from (79).

↪ We now prove (80). Equivalently, we show that for some $a, b \in \mathbb{R}$ we have

$$P_{a,b}(z) \doteq t + (1-t)z(b - az) \leqslant \exp(-z)\exp_t(z), \forall t \in [0,1], \forall z \in \mathbb{R} \tag{89}$$

(we note the result trivially holds for $t = 1$ so we focus on $t \in [0,1)$). If

$$a, b > 0 \tag{90}$$

[6] The proofs, at the end of the proof of Theorem 7.1, elicit functions $u(t), v(t)$.

then $P_{a,b}(z)$ is a downwards facing parabola with its maximum in $z_* \doteq b/(2a) > 0$ and it always has two roots:

$$z_\pm \doteq z_* \cdot \left(1 \pm \sqrt{1 + \frac{4at}{b^2(1-t)}}\right). \tag{91}$$

We now want to choose a, b so as to constrain

$$z_-, z_+ \in \mathbb{I}, \forall t \in [0,1), \mathbb{I} \doteq \sqrt{\frac{1}{1-t}} \cdot [-1,1]. \tag{92}$$

1. Case of z_-. Since $z_- < 0$, we just need $z_- \geqslant -\sqrt{1/(1-t)}$, which after reorganising becomes:

$$\sqrt{1 + \frac{4at}{b^2(1-t)}} \leqslant 1 + \sqrt{a} \cdot \sqrt{\frac{4a}{b^2(1-t)}}. \tag{93}$$

   To get this, it is sufficient we want the same inequality with a t factor in the rightmost square root of the RHS (since $t \leqslant 1$). Making the change of variable $Z \doteq 4at/(b^2(1-t))$, which ranges through $\mathbb{R}_+$, we thus want $\sqrt{1+Z} \leqslant 1 + \sqrt{aZ}$, which indeed holds over $\mathbb{R}_+$ if

$$a \geqslant 1. \tag{94}$$

2. Case of z_+. Since $z_+ > 0$, we just need $z_+ \leqslant \sqrt{1/(1-t)}$, which after reorganising becomes:

$$a \geqslant t + b\sqrt{1-t}. \tag{95}$$

   The RHS takes its max for $b = 2\sqrt{1-t}$, for which it equals $2-t$. Hence, to get $z_+ \leqslant \sqrt{1/(1-t)}$, we just need

$$a \geqslant 2. \tag{96}$$

Hence, if $a \geqslant 2$, then (92) holds. Given that is holds and since $\exp_t$ is an increasing function of t, to get (89), it is enough that we prove that for some $a > 2, b > 0$,

$$P_{a,b}(z) \leqslant \exp_t(-z)\exp_t(z) = \underbrace{\left(1-(1-t)^2z^2\right)^{\frac{1}{1-t}}}_{\doteq Q(z)}, \forall z \in \mathbb{I}, \forall t \in [0,1). \tag{97}$$

We then note

$$Q'(z) = -2(1-t)\cdot zQ^t(z) \quad ; \quad Q''(z) = -2(1-t)\cdot(1-(1-t^2)z^2)Q^{2t-1}(z). \tag{98}$$

A Taylor expansion in $z=0$ then gives $Q(z) \sim_0 1-(1-t)z^2 \doteq R(z)$. Noting $R'(z) = -2(1-t)\cdot z$ and since $Q(z) \leqslant 1$, we obtain $0 \geqslant Q'(z) \geqslant R'(z)$ for $z \in \mathbb{I}_+$ and so $R(z) \leqslant Q(z), \forall z \in \mathbb{I}_+$. Since both functions are even, we thus get

$$R(z) \leqslant Q(z), \forall z \in \mathbb{I}. \tag{99}$$

To get (97), we thus just need $P_{a,b}(z) \leqslant R(z), \forall z \in \mathbb{I}$. Since $t \leqslant 1$, this inequality has the convenient t-free simplification

$$(a-1)z^2 - bz + 1 \geqslant 0, \forall z \in \mathbb{I}. \tag{100}$$

This parabola facing upwards has no root (and is thus non negative) if

$$a \geqslant 1 + \frac{b^2}{4}. \tag{101}$$

To summarize, we get (89) (and so (80)) for any choice a, b satisfying:

$$b > 0 \quad ; \quad a \geqslant \max\left\{2, 1 + \frac{b^2}{4}\right\}. \tag{102}$$

↪ We finish by showing (85). We want to show

$$\forall t \in [0,1], \forall z \geqslant 0, (\log_t)^*(z) \quad \leqslant \quad \underbrace{\log(z) - u(t)z + v(t)(1 + \log^2 z + (z-1)^2)}_{\doteq h_t(z)}. \tag{103}$$

For some $u(t) \geqslant 0, v(t) \geqslant 0$ both defined on $[0,1]$, continuous and with limit 0 in $t = 1^-$. We fix

$$u(t) \doteq 1 - {t^*}^{t^*}, \quad ; \quad v(t) \doteq (\log_t)^*(1) + u(t), \tag{104}$$

and we check they trivially satisfy those properties in addition to being strictly decreasing over $[0,1]$ and satisfying $u([0,1]) = [0, 1-1/\sqrt{2}], v([0,1]) = [0, 1/\sqrt{2}]$. We also check that (103) trivially holds for $t = 1$ so we prove the result for $t \in [0,1)$. We have

$$(\log_t)^*(1) = h_t(1); \quad (\log_t)^{*\prime}(1) = h'_t(1), \tag{105}$$

so functions $(\log_t)^*, h_t$ are tangent at $z = 1$, $\forall t \in [0,1]$. We note

$$h'_t(z) \quad = \quad \frac{1 - (u(t) + 2v(t))z + 2v(t)(z^2 + \log z)}{z}.$$

Given (105), if we can show

$$h'_t(z) \quad \leqslant \quad (\log_t)^{*\prime}(z), \forall z \leqslant 1, \forall t \in [0,1) \tag{106}$$

then, since all related functions are continuous, we obtain

$$(\log_t)^*(z) \quad \leqslant \quad h_t(z), \forall z \leqslant 1, \forall t \in [0,1). \tag{107}$$

(106) is the same as

$$\underbrace{1 - (u(t) + 2v(t))z + 2v(t)(z^2 + \log z)}_{\doteq i_t(z)} \quad \leqslant \quad \underbrace{(1 - u(t))z^{1-t^*}}_{j_t(z)}, \forall z \leqslant 1, \forall t \in [0,1). \tag{108}$$

Since $i_t(1) = j_t(1)$, (108) is guaranteed if $i'_t(z) \geqslant j'_t(z), \forall z \leqslant 1, \forall t \in [0,1)$. This condition can be formulated as (for any c):

$$4v(t)z + \frac{c}{z} + \frac{2v(t) - c}{z} \quad \geqslant \quad u(t) + 2v(t) + \frac{(1-t^*)(1-u(t))}{z^{t^*}}, \forall z \leqslant 1. \tag{109}$$

Now pick $c \doteq (u(t) + 2v(t))^2/(16v(t))$. We can check that

$$c \leqslant v(t), \forall t \in [0,1] \quad ; \quad 4v(t)z + \frac{c}{z} \geqslant u(t) + 2v(t), \forall z \geqslant 0, \forall t \in [0,1), \tag{110}$$

so we get

$$4v(t)z + \frac{c}{z} + \frac{2v(t) - c}{z} \quad \geqslant \quad u(t) + 2v(t) + \frac{v(t)}{z}, \forall z \leqslant 1, \forall t \in [0,1). \tag{111}$$

To get (109), it is thus enough, since $t^* \in [1/2, 1], u(t) \in [0,1], z \leqslant 1$, that we show $v(t) \geqslant 1 - t^*$, which after reordering, yields equivalently $(\log_t)^*(1) \geqslant {t^*}^{t^*} - t^*$. Using $t^* \doteq 1/(2-t)$ and multiplying both sides by $2 - t$ yields in compact form the requirement

$$(2-t)(\log_t)^*(1) \quad \geqslant \quad (1-t)(\log_t)^*(1), \tag{112}$$

which, since $(\log_t)^*(1) > 0$ for $t \in [0,1)$, indeed holds. In summary, we have shown:

$$(\log_t)^*(z) \quad \leqslant \quad h_t(z), \forall z \leqslant 1, \forall t \in [0,1). \tag{113}$$

There remains to cover the cases $z > 1$ and it is sufficient to change the polarity of (106), (108) and thus show

$$i_t(z) \quad \geqslant \quad j_t(z), \forall z \geqslant 1, \forall t \in [0,1). \tag{114}$$

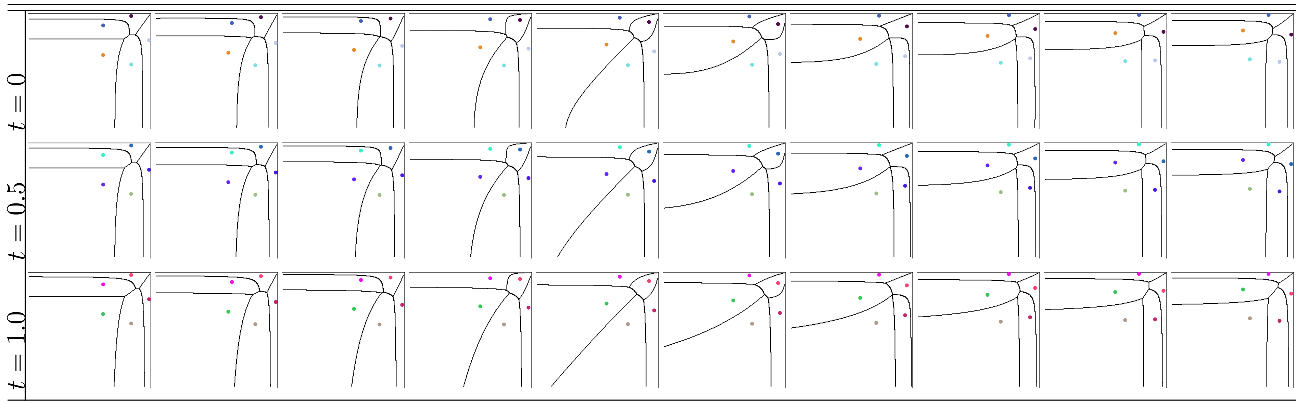


Figure 4: Voronoi diagrams associated to the left population minimizer of the 1D t-exponential (domain = $\mathbb{R}^2_{-*}$) of the vertices of a rotating regular pentagon, for $t \in \{0, 0.5, 1\}$.

We now restrict the interval to check for t. We remark that

$$h''_t(z) \quad = \quad \frac{v(t)}{z^2} \cdot \left(2z^2 - \log z - w(t)\right), \quad w(t) \doteq \frac{1 - v(t)}{v(t)};$$

We note $w(t) \in [\sqrt{2} - 1, +\infty)$. Since $v(t) \geqslant 0$, for all ts such that $(2z^2 - \log z - w(t))([1, +\infty))$ does not contain 0, h_t is convex for $z \geqslant 1$. Since $(\log_t)^*$ is concave, we shall get our result. What is the set of such ts ? The function $z \mapsto 2z^2 - \log z - w(t)$ is strictly increasing for $z \geqslant 1$. Thus, we seek t such that $w(t) \leqslant 2$, or equivalently $v(t) \geqslant 1/3$: for any t such that $v(t) \geqslant 1/3$, h_t is convex over $[1, +\infty)$ and our result (103) holds. We thus refine (115) by checking

$$i_t(z) \quad \geqslant \quad j_t(z), \forall z \geqslant 1, \forall t \in [0, 1) : v(t) \leqslant 1/3. \tag{115}$$

Since $z^{1-t^*} \leqslant z$ and $\log z \geqslant 0$ for $z \geqslant 1$, (115) is implied by showing $1 - (u(t) + 2v(t))z + 2v(t)z^2 \geqslant (1 - u(t))z$ (we recall $v(t) \geqslant 0$), which provides us with the degree-2 polynomial condition

$$1 - (1 + 2v(t))z + 2v(t)z^2 \quad \geqslant \quad 0, \tag{116}$$

and this needs to be checked for $z \geqslant 1, t \in [0, 1) : v(t) \leqslant 1/3$. We compute the roots

$$z_{\pm} \quad \doteq \quad \frac{1 + 2v(t) \pm |1 - 2v(t)|}{2}, \tag{117}$$

and check that the largest root, under the condition $v(t) \leqslant 1/3$, is $z_+ = (1/2)(1 + 2v(t) + 1 - 2v(t)) = 1$. In other words, (115) holds and we have completed the proof of (85), and thus the proof of Theorem 7.1.

IX Voronoi diagrams

Figures 4 and 5 present more detailed Voronoi diagrams for the same setting as described in the main file.

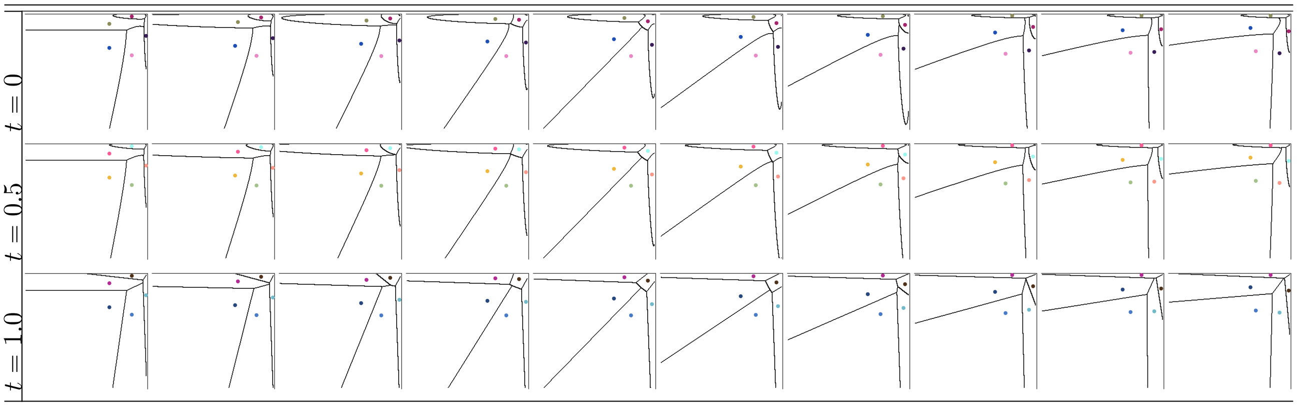


Figure 5: Voronoi diagrams associated to the right population minimizer of the 1D t-exponential (domain = $\mathbb{R}^2_{-*}$) of the vertices of a rotating regular pentagon, for $t \in \{0, 0.5, 1\}$.