Clustering above Exponential Families with Tempered Exponential Measures

EHSAN AMID, RICHARD NOCK, MANFRED K. WARMUTH
{eamid, richardnock, manfred}@google.com

PROBLEM STATEMENT

k-means is a popular clustering method. +Crucial link w/ exponential families:

objects clustered = expectation parameters;

distortion used = square Euclidean distance = KL divergence between Gaussians w/ identity covariance;

This can be generalized to exponential families, where the distortion becomes a general Bregman divergence. This provides a wide range of data generating distributions.
But some roadblocks remain in all cases, like the lack of robustness to outliers of population minimizers (e.g. the right population minimizer is always the arithmetic average, non robust). Also, the current generalizations of exponential families (q-exponential families or even deformed exponential families) fail at removing roadblocks.

In this work, we provide a new attempt at getting the same complete framework **and** gain additional (controllable) robustness for population minimizers, *via* a new generalization of exponential families to **non-normalized** measures.

CONTRIBUTIONS

1. Generalizing exponential family to tempered exponential family while maintaining *divisive* normalizations
2. Information-theoretic distortions between distributions (tempered KL div.)
3. Parameter-based clustering distortions (generalizing Bregman divergences)
4. The information theoretic/geometric link between the distortions
5. Robustness of the population minimizers

TEMPERED EXPONENTIAL MEASURES AND CO-DENSITIES

From left to right: tempered logarithm, tempered exponential, perspective transform of tempered exponential ($t^* \doteq 1/(2-t)$, $[.]_+ \doteq \max\{0, .\}$, same can be defined for tempered log)

$$\log_t(z) \doteq \frac{1}{1-t}\left(z^{1-t} - 1\right), \quad \exp_t(z) \doteq [1 + (1-t)z]_+^{1/(1-t)} \quad (\exp_t)^*(z) \doteq t^* \exp_{t^*}\left(\frac{z}{t^*}\right).$$

Theorem (TEMs & their co-densities) — let

$$\tilde{\mathcal{P}}_{t|\boldsymbol{\hbar}} \doteq \left\{ \tilde{p} \;\middle|\; \begin{array}{l} \mathbb{E}_{\tilde{P}}[\boldsymbol{\varphi}] \doteq \int \boldsymbol{\varphi}(\boldsymbol{x})\,\tilde{p}(\boldsymbol{x})\,\mathrm{d}\xi = \boldsymbol{\hbar}, \\ \int \tilde{p}(\boldsymbol{x})^{1/t^*}\,\mathrm{d}\xi = 1, \\ \tilde{p}(\boldsymbol{x}) \geqslant 0, \forall \boldsymbol{x} \in \mathcal{X}. \end{array} \right\}, \quad \begin{array}{l} \boldsymbol{\varphi} : \mathcal{X} \to \mathbb{R}^d \text{ sufficient statistics,} \\ \boldsymbol{\hbar} \in \mathbb{R}^d \text{ expectation parameter,} \\ \xi \text{ base measure.} \end{array}$$

and (Capital $\tilde{P}$ the measure of density $\tilde{p}$ wrt ξ)

$$H_t(\tilde{P}) \;\doteq\; -\int \psi_t(\tilde{p}(\boldsymbol{x}))\,\mathrm{d}\xi, \quad \text{where} \quad \psi_t(z) \doteq z \log_t z - \log_{t-1} z$$

a (generalized notion of) Tsallis entropy. Then 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

$$\tilde{p}_{t|\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{\exp_t(\boldsymbol{\theta}^\top \boldsymbol{\varphi}(\boldsymbol{x}))}{\exp_t(G_t(\boldsymbol{\theta}))} \quad \text{where} \quad G_t(\boldsymbol{\theta}) = (\log_t)^* \int (\exp_t)^*(\boldsymbol{\theta}^\top \boldsymbol{\varphi}(\boldsymbol{x}))\mathrm{d}\xi$$

G_t = (convex) cumulant ensuring normalization of the **dual** $\tilde{p}^{1/t^*}$ (co-density);
$\boldsymbol{\theta}$ = natural parameter (correspondence $\boldsymbol{\theta} = \nabla G_t^{-1}(\boldsymbol{\hbar})$ like exponential families).

Compared to other generalizations of exponential families (deformed & q-exp. families), key parameters in closed form, *e.g.* cumulant (G_t), **total mass** of TEM ($\int \tilde{p}\mathrm{d}\xi$), *Cf* paper.

Theorem (information theoretic / geometric link) — let

$$F_t(\tilde{P}_{t|\hat{\boldsymbol{\theta}}} \| \tilde{P}_{t|\boldsymbol{\theta}}) \doteq \int f\left(\frac{\mathrm{d}\tilde{p}_{t|\hat{\boldsymbol{\theta}}}}{\mathrm{d}\xi} \oslash_t \frac{\mathrm{d}\tilde{p}_{t|\boldsymbol{\theta}}}{\mathrm{d}\xi}\right) \cdot \mathrm{d}\tilde{p}_{t|\boldsymbol{\theta}}, \quad \begin{array}{l} f \doteq -\log_t, \\ x \oslash_t y \doteq (x^{1-t} - y^{1-t} + 1)_+^{\frac{1}{1-t}} \text{ if } x, y \geqslant 0 \\ \text{(else undefined)} \end{array}$$

Then $\forall \tilde{P}_{t|\hat{\boldsymbol{\theta}}}, \tilde{P}_{t|\boldsymbol{\theta}}$ of the same TEM family,

$$F_t(\tilde{P}_{t|\hat{\boldsymbol{\theta}}} \| \tilde{P}_{t|\boldsymbol{\theta}}) = B_{G_t}(\hat{\boldsymbol{\theta}} \| \boldsymbol{\theta}), \quad \text{where} \quad B_{G_t}(\hat{\boldsymbol{\theta}} \| \boldsymbol{\theta}) \doteq \frac{\overbrace{G_t(\hat{\boldsymbol{\theta}}) - G_t(\boldsymbol{\theta}) - (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top \nabla G_t(\boldsymbol{\theta})}^{\text{Bregman divergence !}}}{1 + (1-t)G_t(\hat{\boldsymbol{\theta}})}.$$

TEM (example)	Support	$\boldsymbol{\lambda}$	$\boldsymbol{\theta}$	$G_t(\boldsymbol{\theta})$	$B_{G_t}(\hat{\boldsymbol{\theta}} \| \boldsymbol{\theta})$
t-exponential	$\left[0, \frac{3-2t}{(1-t)\lambda}\right]$	λ	$\frac{-\lambda}{3-2t}$	$-\log_{2-t}\left((-\theta)^{\frac{1}{2-t}}\right)$	$t^* \cdot \left(\left(\frac{\hat{\theta}}{\theta}\right)^{2-t^*} - (2-t^*) \cdot \log_{t^*}\left(\frac{\hat{\theta}}{\theta}\right) - 1\right)$

which Bregman divergence for $t \to 1$? ☺

POPULATION MINIMIZERS

$\{\boldsymbol{\theta}_i\}_{i=1}^m$ = training set of parameters endowed with an implicit distribution. Two losses for the so-called left and right population minimizers:

$$L_{\mathrm{l}}(\boldsymbol{\theta}) \doteq \mathbb{E}_i[B_{G_t}(\boldsymbol{\theta} \| \boldsymbol{\theta}_i)] \;;\; L_{\mathrm{r}}(\boldsymbol{\theta}) \doteq \mathbb{E}_i[B_{G_t}(\boldsymbol{\theta}_i \| \boldsymbol{\theta})], \; \mathbb{E}_i[.] = \text{average over the training sample.}$$

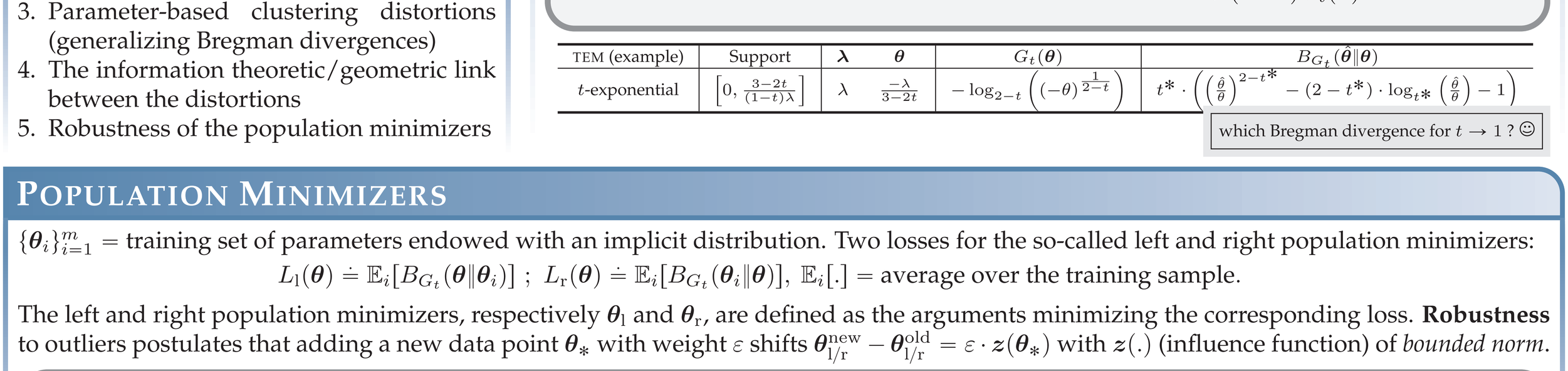
The left and right population minimizers, respectively $\boldsymbol{\theta}_{\mathrm{l}}$ and $\boldsymbol{\theta}_{\mathrm{r}}$, are defined as the arguments minimizing the corresponding loss. **Robustness** to outliers postulates that adding a new data point $\boldsymbol{\theta}_*$ with weight ε shifts $\boldsymbol{\theta}_{\mathrm{l/r}}^{\mathrm{new}} - \boldsymbol{\theta}_{\mathrm{l/r}}^{\mathrm{old}} = \varepsilon \cdot \boldsymbol{z}(\boldsymbol{\theta}_*)$ with $\boldsymbol{z}(.)$ (influence function) of *bounded norm*.

Theorem (population minimizers) — The left and right population minimizers satisfy respectively:

$$\boldsymbol{\theta}_{\mathrm{l}} = \nabla G_t^{-1}\left(\alpha_* \cdot \mathbb{E}_i \nabla G_t(\boldsymbol{\theta}_i)\right), \text{for some } \alpha_* > 0 \text{ (closed form available in some cases, } \textit{Cf} \text{ paper)} \quad ; \quad \boldsymbol{\theta}_{\mathrm{r}} = \mathbb{E}_i\left[\frac{1}{\exp_t^{1-t}(G_t(\boldsymbol{\theta}_i))} \cdot \boldsymbol{\theta}_i\right]$$

Furthermore, for $t \in [0,1]$, $\boldsymbol{\theta}_{\mathrm{l}}$ is robust iff robust for $t = 1$ and $\boldsymbol{\theta}_{\mathrm{r}}$ is robust if $G_t(\boldsymbol{\theta}) = \Omega(\|\boldsymbol{\theta}\|)$ and $t \neq 1$.

EXPERIMENTS (SEE PAPER FOR NUMERICAL RESULTS)



Robustness: cluster (red), 1 heavy outlier (green) moves away ⇒ resulting cluster center (blue)

Voronoi diagram (right center): rotating regular pentagon (diagram always affine for $t = 1$)

B_F balls resp. left, right center