Clustering above Exponential Families with Tempered Exponential Measures

Ehsan Amid, Richard Nock and Manfred K. Warmuth
Exponential families & $k$-means clustering

Why this work: $k$-means = popular clustering, partition of space $\mathcal{X} \subseteq \mathbb{R}^d$ by finding set of centers $\mathcal{C} = \{c_j\}_{j \in [k]}$ minimizing loss to $m$-sample, $\mathbb{E}_{i \sim [m]} \left[ \min_{j \in [k]} D(\theta_i \| c_j) \right]$.
Exponential families & $k$-means clustering

Why this work: $k$-means = popular clustering, partition of space $\mathcal{X} \subset \mathbb{R}^d$ by finding set of centers $\mathcal{C} = \{c_j\}_{j \in [k]}$ minimizing loss to $m$-sample, $\mathbb{E}_{i \sim [m]} \left[ \min_{j \in [k]} D(\theta_i || c_j) \right]$.

If $D$ a Bregman divergence (Mahalanobis, Itakura-Saito, KL, etc.) then population minimizers trivial to compute and we equivalently minimise an information-theoretic loss between distributions in exponential families $\Rightarrow$ embeds $k$-means in broad data generating processes.
Exponential families & \(k\)-means clustering

**Why this work:** \(k\)-means = popular clustering, partition of space \(\mathcal{X} \subseteq \mathbb{R}^d\) by finding set of centers \(\mathcal{C} = \{c_j\}_{j \in [k]}\) minimizing loss to \(m\)-sample, \(\mathbb{E}_{i \sim [m]} \min_{j \in [k]} D(\theta_i || c_j)\)

If \(D\) a Bregman divergence (Mahalanobis, Itakura-Saito, KL, etc.) then population minimizers trivial to compute and we equivalently minimise an information-theoretic loss between distributions in exponential families \(\Rightarrow\) embeds \(k\)-means in broad data generating processes

Can even be generalized above exponential families, to deformed and \(q\)-exponential families, keeping the Bregman divergence formulation of loss
Exponential families & $k$-means clustering

Why this work: $k$-means = popular clustering, partition of space by finding set of centers $\{c_1, \ldots, c_k\}$, minimizing loss to $m$-sample,

If $D$ a Bregman divergence (Mahalanobis, Itakura-Saito, KL, etc.) then population minimizers trivial to compute and we equivalently minimise an information-theoretic loss between distributions in exponential families $\Rightarrow$ embeds $k$-means in broad data generating processes

Can even be generalized above exponential families, to deformed and $q$-exponential families, keeping the Bregman divergence formulation of loss...
Universal modeling with Bregman divergences leads to universal drawbacks, such as lack of robustness to outliers (a Bregman divergence lacks robustness).
Why this work: $k$-means = popular clustering, partition of space by finding set of centers minimizing loss to $m$-sample, if $D$ a Bregman divergence (Mahalanobis, Itakura-Saito, KL, etc.) then population minimizers trivial to compute and we equivalently minimise an information-theoretic loss between distributions in exponential families $\Rightarrow$ embeds $k$-means in broad data generating processes.

Can even be generalized above exponential families, to deformed and $q$-exponential families, keeping the Bregman divergence formulation of loss.

Our objective: get a generalisation of the complete framework, with optional additional properties for clustering such as robustness.
Exponential families & \( k \)-means clustering

Why this work: \( k \)-means = popular clustering, partition of space by finding set of \( m \)-sample, minimizing loss to \( D \) a Bregman divergence (Mahalanobis, Itakura-Saito, KL, etc.) then population minimizers trivial to compute and we equivalently minimise an information-theoretic loss between distributions in exponential families \( \Rightarrow \) embeds \( k \)-means in broad data generating processes. Can even be generalized above exponential families, to deformed and \( q \)-exponential families, keeping the Bregman divergence formulation of loss.

Our objective: get a generalisation of the complete framework, with optional additional properties for clustering such as robustness.

- Distributions (exponential families)
- The information theoretic distortion between distributions (KL divergence)
- Parameter distortions (Bregman divergences)
- The information-geometric / information theoretic link between clustering parameters and distributions
- Get additional properties (robustness)
From **Exponential families** to Tempered Exponential Measures

**Axiomatic characterization**

Set of probability measures satisfying a constraint on their expectation

\[ \tilde{P}_{t|h} = \left\{ \tilde{p} \mid \mathbb{E}_{\tilde{P}}[\phi] = \int \phi(x) \tilde{p}(x) \, d\xi = h, \int \tilde{p}(x) \, d\xi = 1, \tilde{p}(x) \geq 0, \forall x \in X. \right\} \]

+ constraint to maximize entropy

\[ H(P) = - \int p \log p \, d\xi \]

⇒ get an exponential family

\[ p_{\theta}(x) \propto \exp(\theta^\top \phi(x) - G(\theta)) \]

\[ \theta = \nabla G^{-1}(h) \]

natural parameter  \quad \text{cumulant}
From Exponential families to **Tempered Exponential Measures**

**Axiomatic characterization**

Set of probability measures satisfying a constraint on their expectation

\[
\tilde{\mathcal{P}}_{t|h} = \left\{ \tilde{\mu} \left| \begin{array}{l}
\mathbb{E}_{\tilde{\mu}}[\phi] = \int \phi(x) \tilde{\mu}(x) \, dx = h, \\
\int \tilde{\mu}(x) \, dx = 1, \\
\tilde{\mu}(x) \geq 0, \forall x \in \mathcal{X}.
\end{array} \right. \right\}
\]

+ constraint to maximize entropy

\[
H(P) = - \int p \log p \, d\xi
\]

⇒ get an exponential family

\[
p_\theta(x) \propto \exp(\theta^T \phi(x) - G(\theta))
\]

\[
\theta = \nabla G^{-1}(h)
\]

natural parameter

\[t \in [0, 1], t^* \equiv 1/(2 - t)\]
From Exponential families to Tempered Exponential Measures

**Axiomatic characterization**

Set of probability measures satisfying a constraint on their expectation

$$
\tilde{P}_{t|\mathbf{h}} = \left\{ \tilde{\rho} \mid \mathbb{E}_{\tilde{P}}[\phi] \equiv \int \phi(x) \tilde{\rho}(x) \, d\xi = \mathbf{h}, \\
\int \tilde{\rho}(x) \, d\xi = 1, \\
\tilde{\rho}(x) \geq 0, \forall x \in \mathcal{X}. \right\}
$$

+ constraint to maximize entropy

$$
H(P) \equiv - \int p \log p \, d\xi
$$

⇒ get an exponential family

$$
p_\theta(x) \propto \exp(\theta^\top \phi(x) - G(\theta))
$$

$$
\theta = \nabla G^{-1}(\mathbf{h})
$$

natural parameter

$$
\text{cumulant}
$$

**Set of unnormalized measures satisfying a constraint on their expectation**

$$
\tilde{P}_{t|\mathbf{h}} = \left\{ \tilde{\rho} \mid \mathbb{E}_{\tilde{P}}[\phi] \equiv \int \phi(x) \tilde{\rho}(x) \, d\xi = \mathbf{h}, \\
\int \tilde{\rho}(x)^{1/t} \, d\xi = 1, \\
\tilde{\rho}(x) \geq 0, \forall x \in \mathcal{X}. \right\}
$$

+ maximize a generalized Tsallis entropy

$$
H_t(\tilde{P}) \equiv - \int (\tilde{\rho} \log_t \tilde{\rho} - \log_{t-1} \tilde{\rho}) \, d\xi
$$

tempered log

$$
\log_t(z) \equiv \frac{1}{1-t} (z^{1-t} - 1)
$$

Concave, \( \lim_{t \to 1} \log_t = \log \)
From Exponential families to **Tempered Exponential Measures**

**Axiomatic characterization**

Set of probability measures satisfying a constraint on their expectation

\[ \tilde{P}_{t|\mathbf{h}} = \left\{ \tilde{p} \mid \mathbb{E}_{\tilde{p}}[\phi] = \int \phi(x) \tilde{p}(x) \, d\xi = \mathbf{h}, \right\} \]

+ constraint to maximize entropy

\[ H(P) \doteq - \int p \log p \, d\xi \]

⇒ get an exponential family

\[ p_\theta(x) \propto \exp(\theta^\top \phi(x) - G(\theta)) \]

\[ \theta = \nabla G^{-1}(\mathbf{h}) \]

**cumulant**

**Set of unnormalized measures satisfying a constraint on their expectation**

\[ \tilde{P}_{t|\mathbf{h}} = \left\{ \tilde{p} \mid \mathbb{E}_{\tilde{p}}[\phi] = \int \phi(x) \tilde{p}(x) \, d\xi = \mathbf{h}, \right\} \]

+ maximize a generalized Tsallis entropy

\[ H_t(\tilde{P}) \doteq - \int (\tilde{p} \log_t \tilde{p} - \log_{t-1} \tilde{p}) \, d\xi \]

\[ \log_t(z) \doteq \frac{1}{1-t} (z^{1-t} - 1) \]

⇒ get a **tempered exponential measure**

\[ \tilde{p}_{t|\theta}(x) \propto \frac{\exp_t(\theta^\top \phi(x))}{\exp_t(G_t(\theta))} \]

\[ \theta = \nabla G_t^{-1}(\mathbf{h}) \]

**cumulant**
From Exponential families to Tempered Exponential Measures

Axiomatic characterization

Set of probability measures satisfying a constraint on their expectation:

\[ \tilde{p}_{t|\theta}(x) \propto \exp_t(\theta^T \phi(x) \Theta_t G_t(\theta)) \]

\[ z \Theta_t x = \frac{z - x}{1 + (1 - t)x} \]

\[ \tilde{P}_{t|h} = \left\{ \tilde{p} \left| \begin{array}{l} \mathbb{E}_{\tilde{p}}[\phi] = \int \phi(x) \tilde{p}(x) \, d\xi = h, \\ \int \tilde{p}(x)^{1/t^*} \, d\xi = 1, \\ \tilde{p}(x) \geq 0, \forall x \in \mathcal{X}. \end{array} \right\} \]

+maximize a generalized Tsallis entropy:

\[ H_t(\tilde{P}) = -\int (\tilde{p} \log_t \tilde{p} - \log_{t-1} \tilde{p}) \, d\xi \]

\[ \log_t(z) = \frac{1}{1 - t} (z^{1-t} - 1) \]

\[ \Rightarrow \text{get a tempered exponential measure} \]

\[ \tilde{p}_{t|\theta}(x) \propto \frac{\exp_t(\theta^T \phi(x))}{\exp_t(G_t(\theta))} \]

\[ \theta = \nabla G_t^{-1}(h) \]

Tempered Exponential Measures

Set of unnormalized measures satisfying a constraint on their expectation:

\[ \tilde{p}_{t|h} = \left\{ \tilde{p} \left| \begin{array}{l} \mathbb{E}_{\tilde{p}}[\phi] = \int \phi(x) \tilde{p}(x) \, d\xi = h, \\ \int \tilde{p}(x)^{1/t^*} \, d\xi = 1, \\ \tilde{p}(x) \geq 0, \forall x \in \mathcal{X}. \end{array} \right\} \]

+maximize a generalized Tsallis entropy

\[ H_t(\tilde{P}) = -\int (\tilde{p} \log_t \tilde{p} - \log_{t-1} \tilde{p}) \, d\xi \]

\[ \log_t(z) = \frac{1}{1 - t} (z^{1-t} - 1) \]

\[ \Rightarrow \text{get a tempered exponential measure} \]

\[ \tilde{p}_{t|\theta}(x) \propto \frac{\exp_t(\theta^T \phi(x))}{\exp_t(G_t(\theta))} \]

\[ \theta = \nabla G_t^{-1}(h) \]
Exponential families to Tempered Exponential Measures

Axiomatic characterization

Set of probability measures satisfying a constraint on their expectation + constraint to maximize entropy

⇒ get an exponential family

Cumulant in closed form:

\[ G_t(\theta) = (\log_t)^* \int (\exp_t)^*(\theta^T \phi(x)) d\xi \]

Alternative expression

\[ \tilde{p}_{t|\theta}(x) \propto \exp_t(\theta^T \phi(x)) \Theta_t G_t(\theta) \]

\[ z \Theta_t x \doteq \frac{z - x}{1 + (1 - t)x} \]

Set of unnormalized measures satisfying a constraint on their expectation + maximize a generalized Tsallis entropy

\[ \tilde{p}_{t|h} = \left\{ \begin{array}{l} \tilde{p} \mid \mathbb{E}_{\tilde{p}}[\phi] \doteq \int \phi(x) \tilde{p}(x) d\xi = h, \\ \int \tilde{p}(x)^{1/t^*} d\xi = 1, \\ \tilde{p}(x) \geq 0, \forall x \in \mathcal{X}. \end{array} \right\} \]

Cumulant in closed form:

\[ (\log_t)^*(z) \doteq t^* \log_{t^*} \left( \frac{z}{t^*} \right) \]

\[ (\exp_t)^*(z) \doteq t^* \exp_{t^*} \left( \frac{z}{t^*} \right) \]

\[ \exp_t(z) \doteq [1 + (1 - t)z]^{1/(1-t)} \]

\[ \log_t(z) \doteq \frac{1}{1 - t} (z^{1-t} - 1) \]

⇒ get a tempered exponential measure

\[ \tilde{p}_{t|\theta}(x) \propto \frac{\exp_t(\theta^T \phi(x))}{\exp_t(G_t(\theta))} \]

\[ \theta = \nabla G_t^{-1}(h) \]

[(Google Research) cumulant]
Alternative expression
\[ \tilde{p}_t|\theta(x) \propto \exp_t(\theta^\top \phi(x) \ominus_t G_t(\theta)) \]
\[ z \ominus_t x = \frac{z - x}{1 + (1 - t)x} \]

Cumulant in closed form:
\[ G_t(\theta) = (\log_t)^* \int (\exp_t)^* (\theta^\top \phi(x)) d\xi \]

Total mass in closed form:
\[ \int \tilde{p}_t|\theta(x) d\xi = 1 + (1 - t)(G_t(\theta) - \theta^\top \bar{h}) \]

Tempered Exponential Measures

Set of unnormalized measures satisfying a constraint on their expectation
\[ \tilde{p}_t|\bar{h} = \left\{ \tilde{p} \middle| \begin{array}{l} \mathbb{E}_\tilde{P}[\phi] = \int \phi(x) \tilde{p}(x) d\xi = \bar{h}, \\ \tilde{p}(x) \geq 0, \forall x \in \mathcal{X}. \end{array} \right\} \]

+maximize a generalized Tsallis entropy
\[ H_t(\tilde{P}) = -\int (\tilde{p} \log_t \tilde{p} - \log_{t-1} \tilde{p}) d\xi \]

\[ \log_t(z) = \frac{1}{1 - t} (z^{1-t} - 1) \]

⇒ get a tempered exponential measure
\[ \tilde{p}_t|\theta(x) \propto \frac{\exp_t(\theta^\top \phi(x))}{\exp_t(G_t(\theta))} \]
\[ \theta = \nabla G_t^{-1}(\bar{h}) \]

cumulant
Exponential families

Axiomatic characterization of a set of probability measures satisfying a constraint on their expectation

\[ \int \tilde{p}_{t|\theta}(x) \, d\xi = 1 + (1 - t) (G_t(\theta) - \theta^T \mathbf{h}) \]

Cumulant in closed form:

\[ G_t(\theta) = (\log_t)^* \int (\exp_t)^*(\theta^T \phi(x)) \, d\xi \]

Total mass in closed form:

\[ \tilde{p}_{t|\theta}(x) \propto \exp_t(\theta^T \phi(x)) \Theta_{t} G_t(\theta) \]

Tempered Exponential Measures

Set of unnormalized measures satisfying a constraint on their expectation

\[ \tilde{p}_{t|\mathbf{h}} = \left\{ \tilde{p} \mid \mathbb{E}_{\tilde{p}}[\phi] = \int \phi(x) \tilde{p}(x) \, d\xi = \mathbf{h}, \quad \int \tilde{p}(x)^{1/t^*} \, d\xi = 1, \quad \tilde{p}(x) \geq 0, \forall x \in \mathcal{X}. \right\} \]

Maximize a generalized Tsallis entropy

\[ H_t(\tilde{P}) = -\int (\tilde{p} \log_t \tilde{p} - \log_{t-1} \tilde{p}) \, d\xi \]

\[ \log_t(z) \equiv \frac{1}{1-t} (z^{1-t} - 1) \]

Get a tempered exponential measure

\[ \tilde{p}_{t|\theta}(x) \propto \frac{\exp_t(\theta^T \phi(x))}{\exp_t(G_t(\theta))} \]

Google Research
Information Geometric Distortions (gen. Bregman divs)

Information-theoretic distortion between two TEMs, generalizing (reverse) KL divergence:

$$F_t(\tilde{P}_{t|\theta} \parallel \tilde{P}_{t|\theta}) \doteq \int f \left( \frac{d\tilde{p}_{t|\theta}}{d\xi} \otimes_t \frac{d\tilde{p}_{t|\theta}}{d\xi} \right) \cdot d\tilde{p}_{t|\theta}$$

$$f \doteq -\log_t$$

$$x \otimes_t y \doteq [x^{1-t} - y^{1-t} + 1]_{+}^{\frac{1}{1-t}}$$
Information-theoretic distortion between two TEMs, generalizing (reverse) KL divergence:

\[
F_t(\tilde{P}_{t|\hat{\theta}} \Vert \tilde{P}_{t|\theta}) = \int f \left( \frac{d\tilde{p}_{t|\hat{\theta}}}{d\xi} \otimes_t \frac{d\tilde{p}_{t|\theta}}{d\xi} \right) \cdot d\tilde{p}_{t|\theta}
\]

Theorem: for any 2 members of the same TEM family, \( F_t(\tilde{P}_{t|\hat{\theta}} \Vert \tilde{P}_{t|\theta}) = B_{G_t}(\hat{\theta} \Vert \theta) \), with

\[
B_{G_t}(\hat{\theta} \Vert \theta) \equiv \frac{G_t(\hat{\theta}) - G_t(\theta) - (\hat{\theta} - \theta)^\top \nabla G_t(\theta)}{1 + (1 - t)G_t(\hat{\theta})}
\]

\( f \equiv -\log_t \)

\( x \otimes_t y \equiv [x^{1-t} - y^{1-t} + 1]_+^{\frac{1}{1-t}} \)
Given training sample $\{\theta_i\}_{i=1}^m$, we seek its left and right population minimizers, i.e. having $L_1(\theta) \doteq \mathbb{E}_i[B_{G_i}(\theta||\theta_i)]$; $L_r(\theta) \doteq \mathbb{E}_i[B_{G_i}(\theta_i||\theta)]$, we want to compute

$$\theta_1 \doteq \arg\min_{\theta} L_1(\theta) \quad ; \quad \theta_r \doteq \arg\min_{\theta} L_r(\theta)$$

left population minimizer
right population minimizer
Clustering: population minimizers

Given training sample \( \{ \theta_i \}_{i=1}^{m} \) we seek its left and right population minimizers, i.e. having \( L_1(\theta) = \mathbb{E}_i [B_{G_t}(\theta \parallel \theta_i)] \) ; \( L_r(\theta) = \mathbb{E}_i [B_{G_t}(\theta_i \parallel \theta)] \), we want to compute

\[
\theta_l = \arg \min_{\theta} L_1(\theta) \quad ; \quad \theta_r = \arg \min_{\theta} L_r(\theta)
\]

left population minimizer

right population minimizer

Theorem: we have

\[
\theta_l = \nabla G_t^{-1} (\alpha_\star \cdot \mathbb{E}_i \nabla G_t (\theta_i))
\]

\[
\theta_r = \mathbb{E}_i \left[ \frac{1}{\exp_{t}^{1-t} (G_t(\theta_i))} \cdot \theta_i \right]
\]

\( \alpha_\star > 0 \)

Precise interval to search (Cf paper)

Closed forms available in particular cases
Clustering: population minimizers... and robustness

Given training sample $\{\theta_i\}_{i=1}^m$ we seek its left and right population minimizers, i.e. having $L_1(\theta) = \mathbb{E}_i[B_{G_t}(\theta\|\theta_i)]$; $L_r(\theta) = \mathbb{E}_i[B_{G_t}(\theta_i\|\theta)]$, we want to compute

$$\theta_l = \arg\min_{\theta} L_1(\theta) \quad ; \quad \theta_r = \arg\min_{\theta} L_r(\theta)$$

left population minimizer
right population minimizer

Theorem: we have

$$\theta_l = \nabla G_t^{-1}(\alpha_* \cdot \mathbb{E}_i \nabla G_t(\theta_i))$$

$$\theta_r = \mathbb{E}_i \left[ \frac{1}{\exp_t^{1-t}(G_t(\theta_i))} \cdot \theta_i \right]$$

Robustness: add outlier $\theta_*$ with weight $\epsilon$. The center moves as $\theta_{1/r}^{\text{new}} - \theta_{1/r}^{\text{old}} = \epsilon \cdot z(\theta_*)$

If the influence function, $z(.)$, has bounded norm, then the center is robust

Google Research
Clustering: population minimizers... and robustness

Given training sample \( \{ \theta_i \}_{i=1}^m \) we seek its left and right population minimizers, i.e. having 
\[ L_1(\theta) \doteq \mathbb{E}_i [B_{G_t}(\theta \| \theta_i)] ; \quad L_r(\theta) \doteq \mathbb{E}_i [B_{G_t}(\theta_i \| \theta)] , \]
we want to compute
\[ \theta_1 \doteq \arg \min_{\theta} L_1(\theta) ; \quad \theta_r \doteq \arg \min_{\theta} L_r(\theta) \]

Theorem: we have
\[ \theta_1 = \nabla G_t^{-1}(\alpha_* \cdot \mathbb{E}_i \nabla G_t(\theta_i)) \quad \theta_r = \mathbb{E}_i \left[ \frac{1}{\exp_t^{1-t}(G_t(\theta_i))} \cdot \theta_i \right] \]

Robustness: add outlier \( \theta_* \) with weight \( \epsilon \). The center moves as 
\[ \theta_1^{\text{new}} - \theta_1^{\text{old}} = \epsilon \cdot z(\theta_*) \]
If the influence function, \( z(\cdot) \), has bounded norm, then the center is robust

Theorem: left robust iff robust for \( t = 1 \) ; right robust iff
\[ (G_t(\theta) = \Omega(\| \theta \|)) \land (t \neq 1) \]
TEMs: two examples with all details

| TEM                      | Support                              | $\lambda$ | $\theta$ | $\mathcal{h}$ | $G_t^*(\mathcal{h})$                                      |
|--------------------------|--------------------------------------|-----------|-----------|---------------|------------------------------------------------------------|
| 1D $t$-exponential       | $\left[0, \frac{3-2t}{(1-t)\lambda}\right]$ | $\lambda$ | $-\frac{\lambda}{3-2t}$ | $t^* \left(\frac{3-2t}{\lambda}\right)^{2-t^*}$ | $-t^* \cdot \left(\log \frac{1}{2-t^*} \left(\frac{\mathcal{h}}{t^*}\right) - 1\right)$ |
| 1D $t$-Gaussian ($\mu = 0$) | $\left[-\frac{1}{\sqrt{1-t}}, \frac{1}{\sqrt{1-t}}\right]$ | $\sigma^2$ | $-\frac{t^*}{2\sigma^2}$ | $(c_{t^*} \sqrt{2})^{1-t^*} \sigma^{3-t^*}$ | $-\frac{t^*}{2} \cdot \left(\log_{t^*} \left(2c_{t^*}^2 \mathcal{h}\right) - 1\right)$ |
TEMs: two examples with all details

| TEM                      | Support                       | $\lambda$  | $\theta$   | $\mathbf{h}$ | $G^*_t(\mathbf{h})$                       |
|--------------------------|-------------------------------|------------|------------|--------------|--------------------------------------------|
| 1D $t$-exponential       | $[0, \frac{3-2t}{(1-t)\lambda}]$ | $\lambda$ | $\frac{-\lambda}{3-2t}$ | $t^* \left(\frac{3-2t}{\lambda}\right)^{2-t^*}$ | $-t^* \cdot \left(\log \frac{1}{2-t^*} \left(\frac{h}{t^*}\right) - 1\right)$ |
| 1D $t$-Gaussian ($\mu = 0$) | $[-\frac{1}{\sqrt{1-t}}, \frac{1}{\sqrt{1-t}}]$ | $\sigma^2$ | $-\frac{t^*}{2\sigma^2}$ | $(c_t^* \sqrt{2})^{1-t^*} \sigma^{3-t^*}$ | $-\frac{t^*}{2} \cdot \left(\log_{t^*} (2c_t^* h) - 1\right)$ |

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

Two distinct generalisations of Itakura-Saito divergence!
### TEMs: two examples with all details

| TEM                      | Support                                      | $\lambda$   | $\theta$              | $\mathbf{h}$           | $G_t^*(\mathbf{h})$                              |
|--------------------------|----------------------------------------------|-------------|-----------------------|------------------------|-------------------------------------------------|
| $1D$ $t$-exponential     | $[0, \frac{3-2t}{(1-t)\lambda}]$            | $\lambda$  | $\frac{-\lambda}{3-2t}$ | $t^* (\frac{3-2t}{\lambda})^{2-t^*}$            | $-t^* \cdot \left( \log \frac{1}{2-t^*} \left( \frac{\mathbf{h}}{t^*} \right) - 1 \right)$ |
| $1D$ $t$-Gaussian ($\mu = 0$) | $\left[-\frac{1}{\sqrt{1-t}}, \frac{1}{\sqrt{1-t}}\right]$ | $\sigma^2$ | $\frac{-t^*}{2\sigma^2}$ | $(c_t^* \sqrt{2})^{1-t^*} \sigma^{3-t^*}$ | $-t^* \cdot \left( \log_t^* (2c_t^* \mathbf{h}) - 1 \right)$ |

| TEM                      | $G_t(\theta)$                                               | $B_{G_t}(\hat{\theta}||\theta)$ |
|--------------------------|-------------------------------------------------------------|-----------------------------------|
| $1D$ $t$-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                      | $\theta_l$ | $\theta_r$ |
|--------------------------|------------|------------|
| $1D$ $t$-exponential     | $-\mathbb{E}_i \left[ \frac{1}{(-\theta_i)^{1-t^*}} \right]/\mathbb{E}_i \left[ \frac{1}{(-\theta_i)^{2-t^*}} \right]$ | $-\mathbb{E}_i \left[ \frac{1}{(-\theta_i)^{2-t^*}} \right]$ |
| $1D$ $t$-Gaussian ($\mu = 0$) | $-\mathbb{E}_i \left[ \frac{1}{(-\theta_i)^{1-t^*}} \right]/\mathbb{E}_i \left[ \frac{1}{(-\theta_i)^{2-t^*}} \right]$ | $-\mathbb{E}_i \left[ \frac{1}{(-\theta_i)^{2-t^*}} \right]$ | $-\mathbb{E}_i \left[ \frac{1}{(c_t^* \sqrt{t})^{1-t^*}} \cdot \mathbb{E}_i \left[ (-\theta_i)^{\frac{3-t^*}{2}} \right] \right]$ |
### TEMs: two examples with all details

| TEM                        | Support                                      | $\lambda$ | $\theta$ | $h$                      | $G_t^*(h)$                                    |
|----------------------------|----------------------------------------------|-----------|----------|--------------------------|-----------------------------------------------|
| 1D $t$-exponential         | $[0, \frac{3-2t}{(1-t)\lambda}]$           | $\lambda$ | $\frac{-\lambda}{3-2t}$ | $t^* \left(\frac{3-2t}{\lambda}\right)^{2-t^*}$ | $-t^* \cdot \left(\log \frac{1}{2-t^*} \left(\frac{h}{t^*}\right) - 1\right)$ |
| 1D $t$-Gaussian ($\mu = 0$)| $[-\frac{1}{\sqrt{1-t}}, \frac{1}{\sqrt{1-t}}]$ | $\sigma^2$ | $\frac{-t^*}{2\sigma^2}$ | $(c_t^* \sqrt{2})^{1-t^*} \sigma^{3-t^*}$     | $-\frac{t^*}{2} \cdot \left(\log_{t^*} \left(2c_t^2 \frac{h}{t^*}\right) - 1\right)$ |

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

Right population minimizer: **not the arithmetic average**, unless $t=1$ (Bregman divergences)

| TEM                        | $\theta_l$                         | $\theta_r$                                    |
|----------------------------|------------------------------------|-----------------------------------------------|
| 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 ($\mu = 0$)| $-\mathbb{E}_i \left[\frac{1}{(-\theta_i)^{1-t^*}}\right] / \mathbb{E}_i \left[\frac{1}{(-\theta_i)^{3-t^*}}\right]$ | $-\frac{1}{(c_t^* \sqrt{t^*})^{1-t^*}} \cdot \mathbb{E}_i \left[(-\theta_i)^{3-t^*}\right]$ |
Experiments

(more in paper)
Robustness ($t$-exponential)

| $t = t^* = 1$ (exp. family) |
|-----------------------------|
|                             |

Robustness: **cluster** (red), 1 heavy **outlier** (green) moves away ⇒ resulting **cluster center** (blue)
Robustness ($t$-exponential)

Exponential family case robust, but center still moves “far”

$t = t^* = 1$

(Exp. family)

Robustness: cluster (red), 1 heavy outlier (green) moves away $\Rightarrow$ resulting cluster center (blue)
Robustness ($t$-exponential)

| $t = 0$ | $t = \frac{1}{2}$ | $t = 1$ | $t = 1$ |
|--------|-----------------|--------|--------|
| $t^* = 0$ | $t^* = \frac{1}{2}$ | $t^* = 1$ | $t^* = 1$ |

Robustness: **cluster** (red), 1 heavy **outlier** (green) moves away $\Rightarrow$ resulting **cluster center** (blue).
Robustness ($t$-exponential)

$\begin{array}{c|c}
\hline
 t = 0 & \text{TEM center remains closer to cluster, “more robust”} \\
 (t^* = 1/2) & \\
 \hline
 t = t^* = 1 & \text{exp. family} \\
 \hline
\end{array}$

Robustness: \textbf{cluster} (red), 1 heavy \textbf{outlier} (green) moves away $\Rightarrow$ resulting \textbf{cluster center} (blue)
Thank You

Ehsan Amid

Richard Nock

Manfred K. Warmuth