Clustering above Exponential Families with Tempered Exponential Measures

Ehsan Amid, Richard Nock and Manfred K. Warmuth

Google Research

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 $\mathcal{C} \doteq \{\boldsymbol{c}_j\}_{j\in[k]}$ minimizing loss to *m*-sample, $\mathbb{E}_{i\sim[m]}[\min_{j\in[k]} D(\boldsymbol{\theta}_i\|\boldsymbol{c}_j)]$

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 $\mathcal{C} \doteq \{\boldsymbol{c}_j\}_{j\in[k]}$ minimizing loss to *m*-sample, $\mathbb{E}_{i\sim[m]}[\min_{j\in[k]} D(\boldsymbol{\theta}_i \| \boldsymbol{c}_j)]$

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 ⇒ embeds *k*-means in broad data generating processes

Google Research

Exponential families & *k*-means clustering

<u>Why this work</u>: *k*-means = popular clustering, partition of space $\mathcal{X} \subseteq \mathbb{R}^d$ by finding set of centers $\mathcal{C} \doteq \{\boldsymbol{c}_j\}_{j\in[k]}$ minimizing loss to *m*-sample, $\mathbb{E}_{i\sim[m]}[\min_{j\in[k]} D(\boldsymbol{\theta}_i \| \boldsymbol{c}_j)]$

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 ⇒ 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

Google Research

Exponential families & k-means clustering

Why this work: k-means = popul... by finding

set of ... $(\boldsymbol{\theta}_i \| \boldsymbol{c}_j)]$

If D a B... ulation

minimiz... -theoretic

loss bet... road data

generat...

But

Can eve... amilies, to deformed and

q-expon... families, keeping the Bregman divergence formulation of loss

Exponential families & *k*-means clustering

Why this work: *k*-means = popular ... by finding

set of ... $(\boldsymbol{\theta}_i \| \boldsymbol{c}_j)]$

If *D* a B ... ulation

minimiz ... -theoretic

loss bet ... road data

generat

Can eve ... to deformed and

q-expon ... 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)

Google Research

Exponential families & *k*-means clustering

Why this work: *k*-means = popul... by finding

set of ... $(\boldsymbol{\theta}_i \| \boldsymbol{c}_j)]$

If *D* a B... ulation

minimiz... -theoretic

loss bet... road data

generat...

Can eve... ...milies, to deformed and

q-expon... 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 ... by finding set of ... $(\boldsymbol{\theta}_i \| \boldsymbol{c}_j)]$

If *D* a B...

minimiz...

loss bet...

generat...

Can eve...

q-expon... families, keeping the Breg...

Our objective: get a *generalisation* of the *complete* f... work, with optional additio... clustering suc...

Generalizing

- 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

$$\tilde{\mathcal{P}}_{t|\boldsymbol{\hbar}} \doteq \left\{ \tilde{p} \,\middle|\, \begin{array}{l} \mathbb{E}_{\tilde{P}}[\boldsymbol{\phi}] \doteq \int \boldsymbol{\phi}(\boldsymbol{x})\, \tilde{p}(\boldsymbol{x})\, \mathrm{d}\xi = \boldsymbol{\hbar}, \\ \int \tilde{p}(\boldsymbol{x}) \qquad \mathrm{d}\xi = 1, \\ \tilde{p}(\boldsymbol{x}) \geq 0, \forall \boldsymbol{x} \in \mathcal{X}. \end{array} \right\}$$

+constraint to maximize entropy

$$H(P) \doteq -\int p \log p \mathrm{d}\xi$$

⇒ get an exponential family

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

$\boldsymbol{\theta} = \nabla G^{-1}(\boldsymbol{\hbar})$

natural parameter

cumulant

From Exponential families to Tempered Exponential Measures

<u>Axiomatic characterization</u>

Set of probability measures satisfying a constraint on their expectation

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

+constraint to maximize entropy

$$H(P) \doteq -\int p \log p \mathrm{d}\xi$$

$\Rightarrow$ get an exponential family

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

$\boldsymbol{\theta} = \nabla G^{-1}(\boldsymbol{\hbar})$

natural parameter

cumulant

Set of unnormalized measures satisfying a constraint on their expectation

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

$$\boxed{t \in [0,1], t^* \doteq 1/(2-t)}$$

From Exponential families to Tempered Exponential Measures

<u>Axiomatic characterization</u>

Set of probability measures satisfying a constraint on their expectation

$$\tilde{\mathcal{P}}_{t|\boldsymbol{\hbar}} \doteq \left\{ \tilde{p} \,\middle|\, \begin{array}{l} \mathbb{E}_{\tilde{P}}[\boldsymbol{\phi}] \doteq \int \boldsymbol{\phi}(\boldsymbol{x})\, \tilde{p}(\boldsymbol{x})\, \mathrm{d}\xi = \boldsymbol{\hbar}, \\ \int \tilde{p}(\boldsymbol{x}) \quad\quad \mathrm{d}\xi = 1, \\ \tilde{p}(\boldsymbol{x}) \geq 0, \forall \boldsymbol{x} \in \mathcal{X}. \end{array} \right\}$$

+constraint to maximize entropy

$$H(P) \doteq -\int p \log p \mathrm{d}\xi$$

$\Rightarrow$ get an exponential family

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

$\boldsymbol{\theta} = \nabla G^{-1}(\boldsymbol{\hbar})$

natural parameter

cumulant

Set of unnormalized measures satisfying a constraint on their expectation

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

+maximize a generalized Tsallis entropy

$$H_t(\tilde{P}) \doteq -\int (\tilde{p} \log_t \tilde{p} - \log_{t-1} \tilde{p}) \mathrm{d}\xi$$

tempered log $\longrightarrow$ $\log_t(z) \doteq \dfrac{1}{1-t}(z^{1-t} - 1)$

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

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

+constraint to maximize entropy

$$H(P) \doteq -\int p \log p \mathrm{d}\xi$$

⇒ get an exponential family

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

$\boldsymbol{\theta} = \nabla G^{-1}(\boldsymbol{\hbar})$

natural parameter

cumulant

Set of unnormalized measures satisfying a constraint on their expectation

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

+maximize a generalized Tsallis entropy

$$H_t(\tilde{P}) \doteq -\int (\tilde{p} \log_t \tilde{p} - \log_{t-1} \tilde{p}) \mathrm{d}\xi$$

$$\log_t(z) \doteq \frac{1}{1-t}\left(z^{1-t} - 1\right)$$

⇒ get a tempered exponential measure

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

$\boldsymbol{\theta} = \nabla G_t^{-1}(\boldsymbol{\hbar})$

cumulant

Google Research

Tempered Exponential Measures

Alternative expression

$$\tilde{p}_{t|\boldsymbol{\theta}}(\boldsymbol{x}) \propto \exp_t(\boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x}) \ominus_t G_t(\boldsymbol{\theta}))$$

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

Set of unnormalized measures satisfying a constraint on their expectation

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

+maximize a generalized Tsallis entropy

$$H_t(\tilde{P}) \doteq -\int (\tilde{p} \log_t \tilde{p} - \log_{t-1} \tilde{p}) \mathrm{d}\xi$$

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

⇒ get a tempered exponential measure

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

$\boldsymbol{\theta} = \nabla G_t^{-1}(\boldsymbol{\hbar})$

cumulant

Alternative expression

$$\tilde{p}_{t|\boldsymbol{\theta}}(\boldsymbol{x}) \propto \exp_t(\boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x}) \ominus_t G_t(\boldsymbol{\theta}))$$

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

Cumulant in closed form:

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

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

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

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

tempered exp

$$[.]_+ \doteq \max\{0, .\} \qquad \lim_{t \to 1} \exp_t = \exp$$

Tempered Exponential Measures

Set of unnormalized measures satisfying a constraint on their expectation

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

+maximize a generalized Tsallis entropy

$$H_t(\tilde{P}) \doteq -\int (\tilde{p}\log_t \tilde{p} - \log_{t-1} \tilde{p})\mathrm{d}\xi$$

$$\log_t(z) \doteq \frac{1}{1-t}\left(z^{1-t} - 1\right)$$

⇒ get a tempered exponential measure

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

$$\boldsymbol{\theta} = \nabla G_t^{-1}(\boldsymbol{\hbar})$$

cumulant

Google Research

Tempered Exponential Measures

Alternative expression

$$\tilde{p}_{t|\boldsymbol{\theta}}(\boldsymbol{x}) \propto \exp_t(\boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x}) \ominus_t G_t(\boldsymbol{\theta}))$$

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

Cumulant in closed form:

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

Total mass in closed form:

$$\int \tilde{p}_{t|\boldsymbol{\theta}}(\boldsymbol{x})\mathrm{d}\xi = 1 + (1-t)(G_t(\boldsymbol{\theta}) - \boldsymbol{\theta}^\top \boldsymbol{\hbar})$$

Set of unnormalized measures satisfying a constraint on their expectation

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

+maximize a generalized Tsallis entropy

$$H_t(\tilde{P}) \doteq - \int (\tilde{p} \log_t \tilde{p} - \log_{t-1} \tilde{p})\mathrm{d}\xi$$

$$\log_t(z) \doteq \frac{1}{1-t}\left(z^{1-t} - 1\right)$$

⇒ get a tempered exponential measure

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

$\boldsymbol{\theta} = \nabla G_t^{-1}(\boldsymbol{\hbar})$

cumulant

Google Research

Tempered Exponential Measures

Alternative expression

$$\tilde{p}_{t|\boldsymbol{\theta}}(\boldsymbol{x}) \propto \exp_t(\boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x}) \ominus_t G_t(\boldsymbol{\theta}))$$

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

<u>Cumulant in closed form</u>:

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

<u>Total mass in closed form</u>:

$$\int \tilde{p}_{t|\boldsymbol{\theta}}(\boldsymbol{x})\mathrm{d}\xi = 1 + (1-t)(G_t(\boldsymbol{\theta}) - \boldsymbol{\theta}^\top \boldsymbol{\hbar})$$

$\tilde{p}_{t|\boldsymbol{\theta}}^{1/t^*}$ = co-density

Set of unnormalized measures satisfying a constraint on their expectation

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

+maximize a generalized Tsallis entropy

$$H_t(\tilde{P}) \doteq -\int (\tilde{p}\log_t \tilde{p} - \log_{t-1}\tilde{p})\mathrm{d}\xi$$

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

⇒ get a tempered exponential measure

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

$\boldsymbol{\theta} = \nabla G_t^{-1}(\boldsymbol{\hbar})$

cumulant

Information Geometric Distortions (gen. Bregman divs)

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

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

$$f \doteq -\log_t$$

$$x \oslash_t y \doteq [x^{1-t} - y^{1-t} + 1]_+^{\frac{1}{1-t}}$$

Google Research

Information Geometric Distortions (gen. Bregman divs)

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

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

$$f \doteq -\log_t$$

$$x \oslash_t y \doteq [x^{1-t} - y^{1-t} + 1]_+^{\frac{1}{1-t}}$$

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

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

} **Bregman divergence**

Google Research

Clustering: population minimizers

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

$$\underbrace{\boldsymbol{\theta}_\mathrm{l}}_{\text{left population minimizer}} \doteq \arg\min_{\boldsymbol{\theta}} L_\mathrm{l}(\boldsymbol{\theta}) \quad ; \quad \underbrace{\boldsymbol{\theta}_\mathrm{r}}_{\text{right population minimizer}} \doteq \arg\min_{\boldsymbol{\theta}} L_\mathrm{r}(\boldsymbol{\theta})$$

Clustering: population minimizers

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

$$\underbrace{\boldsymbol{\theta}_{\mathrm{l}}}_{\text{left population minimizer}} \doteq \arg\min_{\boldsymbol{\theta}} L_{\mathrm{l}}(\boldsymbol{\theta}) \quad ; \quad \underbrace{\boldsymbol{\theta}_{\mathrm{r}}}_{\text{right population minimizer}} \doteq \arg\min_{\boldsymbol{\theta}} L_{\mathrm{r}}(\boldsymbol{\theta})$$

Theorem: we have

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

$\alpha_* > 0$

Precise interval to search (*Cf* paper)
Closed forms available in particular cases

Clustering: population minimizers... and robustness

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

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

left population minimizer right population minimizer

Theorem: we have

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

Robustness: add outlier $\boldsymbol{\theta}_*$ with weight ϵ. The center moves as $\boldsymbol{\theta}_{\mathrm{l/r}}^{\mathrm{new}} - \boldsymbol{\theta}_{\mathrm{l/r}}^{\mathrm{old}} = \epsilon \cdot \boldsymbol{z}(\boldsymbol{\theta}_*)$

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

Clustering: population minimizers... and robustness

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

$$\underbrace{\boldsymbol{\theta}_\mathrm{l}}_{\text{left population minimizer}} \doteq \arg\min_{\boldsymbol{\theta}} L_\mathrm{l}(\boldsymbol{\theta}) \quad ; \quad \underbrace{\boldsymbol{\theta}_\mathrm{r}}_{\text{right population minimizer}} \doteq \arg\min_{\boldsymbol{\theta}} L_\mathrm{r}(\boldsymbol{\theta})$$

Theorem: we have

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

Robustness: add outlier $\boldsymbol{\theta}_*$ with weight ϵ. The center moves as $\boldsymbol{\theta}^{\mathrm{new}}_{\mathrm{l/r}} - \boldsymbol{\theta}^{\mathrm{old}}_{\mathrm{l/r}} = \epsilon \cdot \boldsymbol{z}(\boldsymbol{\theta}_*)$

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

Theorem: left robust *iff* robust for $t = 1$; right robust *iff*

$$\boxed{(G_t(\boldsymbol{\theta}) = \Omega(\|\boldsymbol{\theta}\|)) \wedge (t \neq 1)}$$

TEMs: two examples with all details

TEM	Support	$\boldsymbol{\lambda}$	$\boldsymbol{\theta}$	$\boldsymbol{\hbar}$	$G^\star_t(\boldsymbol{\hbar})$
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	$-\frac{t^*}{2\sigma^2}$	$(c_{t^*}\sqrt{2})^{1-t^*}\sigma^{3-t^*}$	$-\frac{t^*}{2} \cdot \left(\log_{t^{**}}(2c^2_{t^*}\hbar) - 1\right)$

TEMs: two examples with all details

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 ($\mu = 0$)	$(\log_t)^*\left(\frac{c_{t^*}}{\sqrt{-\theta}}\right)$	$\frac{t^*}{2} \cdot \left(\left(\sqrt{\frac{\hat{\theta}}{\theta}}\right)^{3-t^*} - (3-t^*) \cdot \log_{t^*}\sqrt{\frac{\hat{\theta}}{\theta}} - 1\right)$

Two distinct generalisations of Itakura-Saito divergence !

TEMs: two examples with all details

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 $(\mu = 0)$	$-\mathbb{E}_i\left[\frac{1}{(-\theta_i)^{\frac{1-t^*}{2}}}\right] / \mathbb{E}_i\left[\frac{1}{(-\theta_i)^{\frac{3-t^*}{2}}}\right]$	$-\frac{1}{(c_{t^*}\sqrt{t^*})^{1-t^*}} \cdot \mathbb{E}_i\left[(-\theta_i)^{\frac{3-t^*}{2}}\right]$

TEMs: two examples with all details

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	$G_t(\boldsymbol{\theta})$	$B_{G_t}(\hat{\boldsymbol{\theta}}\|\boldsymbol{\theta})$
1D t-exponential	$-\log_{2-t}\left((-\theta)^{\frac{1}{2-t}}\right)$	[illegible]
1D t-Gaussian ($\mu = 0$)	$(\log_t)^*\left(\frac{c_{t^*}}{\sqrt{-\theta}}\right)$	[illegible]

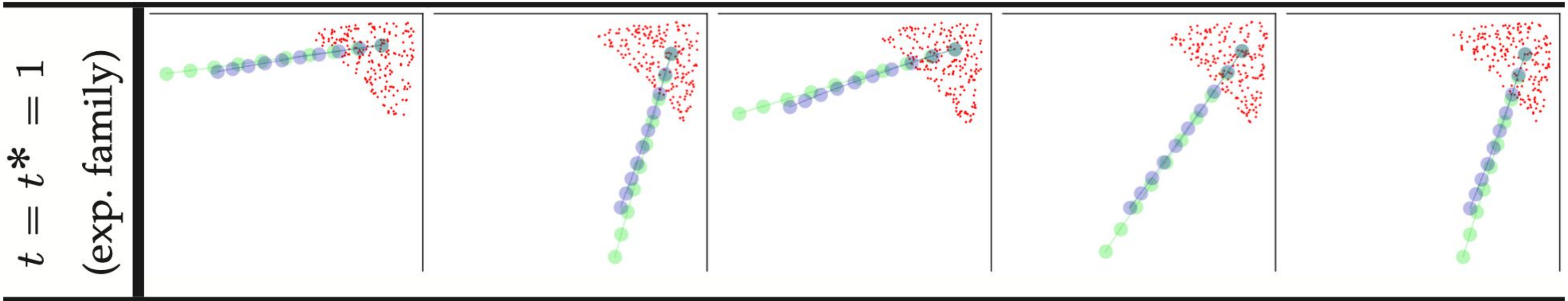
Right population minimizer: not the arithmetic average, unless t=1 (Bregman divergences)

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 ($\mu = 0$)	$-\mathbb{E}_i\left[\frac{1}{(-\theta_i)^{\frac{1-t^*}{2}}}\right] / \mathbb{E}_i\left[\frac{1}{(-\theta_i)^{\frac{3-t^*}{2}}}\right]$	$-\frac{1}{(c_{t^*}\sqrt{t^*})^{1-t^*}} \cdot \mathbb{E}_i\left[(-\theta_i)^{\frac{3-t^*}{2}}\right]$

Experiments

(more in paper)

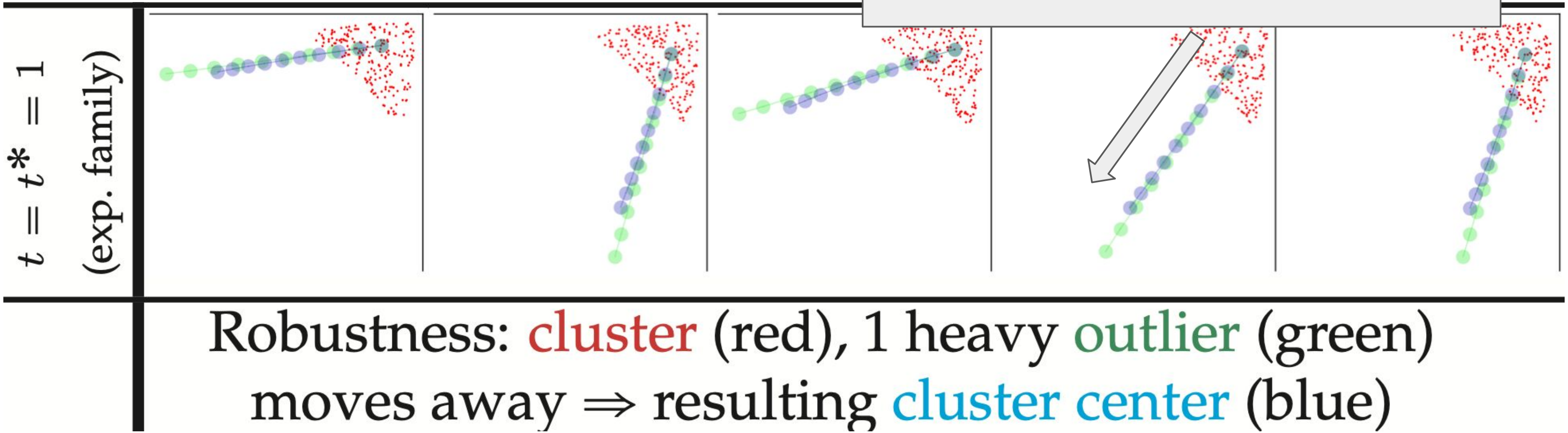
Robustness (t-exponential)



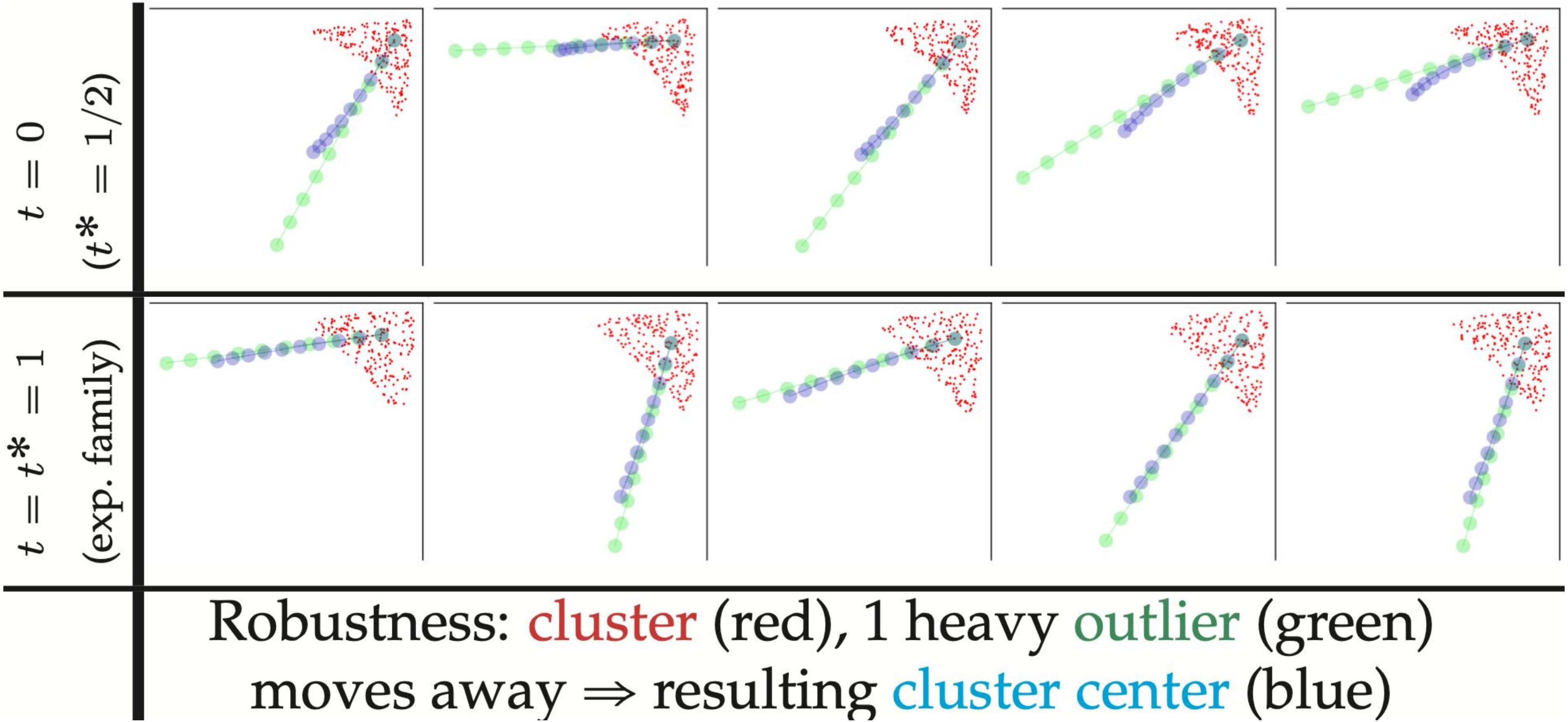
Robustness: cluster (red), 1 heavy outlier (green) moves away ⇒ resulting cluster center (blue)

Robustness (t-exponential)

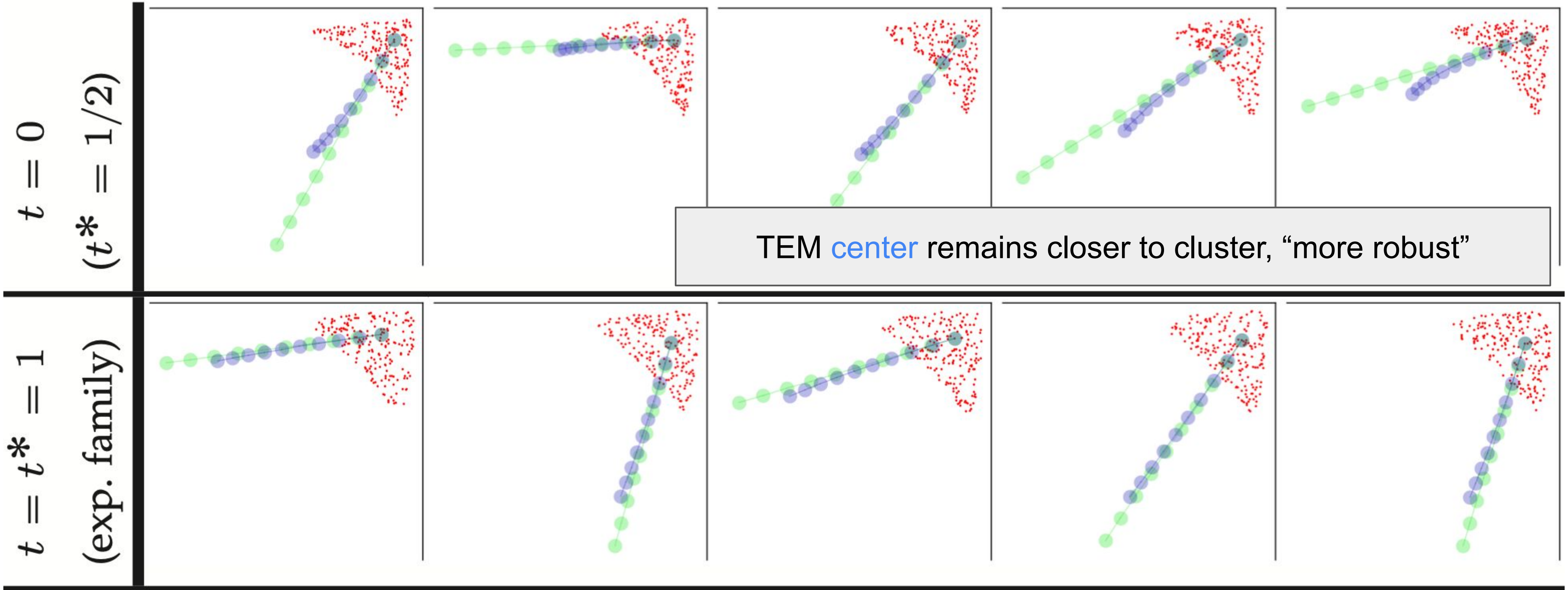
Exponential family case robust, but center still moves "far"



Robustness (t-exponential)



Robustness (*t*-exponential)



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

Thank You

Ehsan
Amid

Richard
Nock

Manfred K.
Warmuth

Google Research