A Review of t-exponential family

The t-exponential family has been regarded as a useful generalization of the exponential family. To introduce the t-exponential family, one need to first define the t-exponential function and t-logarithm function,

$$\exp_t(x) = \begin{cases} \exp(x) & \text{if } t = 1 \\ [1 + (1-t)x]_+^{\frac{1}{1-t}} & \text{otherwise.} \end{cases} \tag{49}$$

$$\log_t(x) := \begin{cases} \log(x) & \text{if } t = 1 \\ \left(x^{1-t} - 1\right) / (1-t) & \text{otherwise.} \end{cases} \tag{50}$$

where $[x]_+$ be x if the $x > 0$ and 0 otherwise. Figure 4 depicts the $\exp_t$ function, which shows a slower decay than the $\exp$ function for $t > 1$.

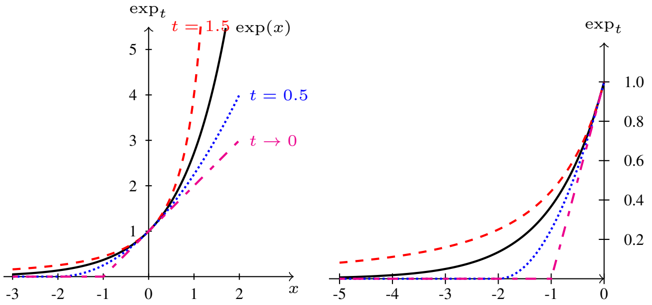

Figure 4: Left: $\exp_t$ Function. Right: Zoom of $\exp_t$ function in domain of [-5,0].

The t-exponential family is then defined as

$$p(x;\theta) := \exp_t\left(\langle \Phi(x), \theta\rangle - g_t(\theta)\right). \tag{51}$$

Although $g_t(\theta)$ cannot usually be analytically obtained, it still preserves convexity. In addition, it is very close to being a moment generating function,

$$\nabla_\theta g_t(\theta) = \mathbb{E}_q\left[\Phi(x)\right]. \tag{52}$$

where $q(x)$ is called the escort distribution of $p(x)$, which is defined as:

$$q(x;\theta) := p(x;\theta)^t / Z(\theta) \tag{53}$$

Here $Z(\theta) = \int p^t(x;\theta)dx$ is the normalizing constant which ensures that the escort distribution integrates to 1. A general version of this result appears as Lemma 3.8 in Sears [12] and a version specialized to the generalized ϕ-exponential families appears as Proposition 5.2 in [17].

A prominent member of the t-exponential family is the Student's t-distribution [13] as shown in the following example.

Example 5 (Student's t-distribution) *A k-dimensional Student's t-distribution $p(\mathbf{x}) = St(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma},v)$ with $v > 2$ degrees of freedom has the following probability density function:*

$$St(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma},v) = \frac{\Gamma\left((v+k)/2\right)}{(\pi v)^{k/2}\,\Gamma(v/2)|\,\boldsymbol{\Sigma}\,|^{1/2}} \cdot \left(1 + (\mathbf{x}-\boldsymbol{\mu})^\top (v\,\boldsymbol{\Sigma})^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)^{-(v+k)/2}. \tag{54}$$

Let $-(v+k)/2 = 1/(1-t)$ and

$$\Psi = \left(\frac{\Gamma\left((v+k)/2\right)}{(\pi v)^{k/2}\,\Gamma(v/2)|\,\boldsymbol{\Sigma}\,|^{1/2}}\right)^{-2/(v+k)}$$

then (54) *becomes*

$$St(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma},v) = \left(1+(1-t)\left\langle\Phi(\mathbf{x}),\boldsymbol{\theta}\right\rangle - g_t(\boldsymbol{\theta})\right)^{1/(1-t)} = \exp_t\left(\left\langle\Phi(\mathbf{x}),\boldsymbol{\theta}\right\rangle - g_t(\boldsymbol{\theta})\right).$$

where

$$\begin{aligned}\mathbf{K} &= (v\,\boldsymbol{\Sigma})^{-1}, \Phi(\mathbf{x}) = [\mathbf{x};\mathbf{x}\,\mathbf{x}^\top], \boldsymbol{\theta} = [\boldsymbol{\theta}_1,\boldsymbol{\theta}_2]\\ \boldsymbol{\theta}_1 &= -2\Psi\,\mathbf{K}\,\boldsymbol{\mu}/(1-t), \boldsymbol{\theta}_2 = \Psi\,\mathbf{K}/(1-t)\\ g_t(\boldsymbol{\theta}) &= -\left(\frac{\Psi}{1-t}\right)\left(\boldsymbol{\mu}^\top\,\mathbf{K}\,\boldsymbol{\mu}+1\right)+\frac{1}{1-t}\end{aligned}$$

The escort of Student's t-distribution is,

$$q(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{Z}St(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma},v)^t = St(\mathbf{x};\boldsymbol{\mu},v\,\boldsymbol{\Sigma}/(v+2),v+2)$$

Interestingly, the mean of the Student's t-pdf is $\boldsymbol{\mu}$, and its variance is $v\,\boldsymbol{\Sigma}/(v-2)$ while the mean and variance of the escort are $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ respectively.

B Proof of Theorem 2

Theorem For any μ, define $\theta(\mu)$ (if exists) to be the parameter of the t-exponential family s.t.

$$\mu = \mathbb{E}_{q(x;\theta(\mu))}\left[\Phi(x)\right] = \int \Phi(x)q(x;\theta(\mu))\,dx. \tag{55}$$

$$\text{Then}\quad g_t^*(\mu) = \begin{cases}-H_t(p(x;\theta(\mu))) \text{ if } \theta(\mu) \text{ exists}\\ +\infty \text{ otherwise}.\end{cases} \tag{56}$$

where $g_t^*(\mu)$ denotes the Fenchel dual of $g_t(\theta)$. By duality it also follows that

$$g_t(\theta) = \sup_\mu\left\{\langle\mu,\theta\rangle - g_t^*(\mu)\right\}. \tag{57}$$

Proof In view of (3) and (9),

$$\mu = \mathbb{E}_{q(x;\theta(\mu))}\left[\Phi(x)\right] = \nabla_\theta g_t(\theta).$$

We only need to consider the case when $\theta(\mu)$ exists since otherwise $g_t^*(\mu)$ is trivially defined as $+\infty$. When $\theta(\mu)$ exists, clearly $\theta(\mu) \in (\nabla g_t)^{-1}(\mu)$. Therefore, we have,

$$\begin{aligned}\sup_\theta\left\{\langle\mu,\theta\rangle - g_t(\theta)\right\} &= \sup_\theta\left\{\left\langle\mathbb{E}_{q(x;\theta(\mu))}\left[\Phi(x)\right],\theta\right\rangle - g_t(\theta)\right\}\\ &= \left\langle\mathbb{E}_{q(x;\theta(\mu))}\left[\Phi(x)\right],\theta(\mu)\right\rangle - g_t(\theta(\mu)) &&(58)\\ &= \int q(x;\theta(\mu))\left(\langle\Phi(x),\theta(\mu)\rangle - g_t(\theta(\mu))\right)dx\\ &= \int q(x;\theta(\mu))\log_t p(x;\theta(\mu))dx &&(59)\\ &= -H_t(p(x;\theta(\mu)))\end{aligned}$$

Equation (58) follows because of the duality between $\theta(\mu)$ and μ, while (59) is because $\log_t p(x;\theta(\mu)) = (\langle\Phi(x),\theta(\mu)\rangle - g_t(\theta(\mu)))$. ■

C Proof of Theorem 4

Theorem The relative t-entropy is the Bregman divergence defined on the negative t-entropy $-H_t(p)$.

Proof First, we know the concavity of H_t from Theorem 2 which leads to the convexity of $-H_t$. In addition, since $p(x)$ and $q(x)$ are one-to-one mapped, $H_t(p)$ can also work with $q(x)$ equivalently. Let us take the functional derivative of $H_t(p)$ with respect to the $q(x)$,

$$\begin{aligned}
\frac{dH_t(p(x))}{dq(x)} &= -\frac{d\left(\int q(z)\log_t p(z)dz\right)}{dq(x)} \\
&= -\log_t p(x) - \int q(z)\frac{d\log_t p(z)}{dq(x)}dz \\
&= -\log_t p(x) - \int \frac{q(z)}{p(z)^t}\frac{dp(z)}{dq(x)}dz \qquad (60) \\
&= -\log_t p(x) - \frac{1}{\int p(z)^t dz}\int \frac{dp(z)}{dq(x)}dz \qquad (61) \\
&= -\log_t p(x) \qquad (62)
\end{aligned}$$

where, (60) comes from $d\log_t(x)/dx = 1/x^t$ by definition of $\log_t$ function; (61) is because $q(z) = p(z)^t/\int p(z)^t dz$; and (62) is because $\int p(z)dz = 1$.

Then the Bregman divergence between two distributions $p(x)$ and $\tilde{\mathrm{p}}(x)$ is defined based on their escort, using the fact that $-H_t(p)$ is a convex function:

$$\begin{aligned}
D_t(p\|\,\tilde{\mathrm{p}}) &= -H_t(p) + H_t(\tilde{\mathrm{p}}) - \int \frac{dH_t(\tilde{\mathrm{p}}(x))}{d\,\tilde{\mathrm{q}}(x)}(\tilde{\mathrm{q}}(x) - q(x)) \\
&= \int q(x)\log_t p(x) - \tilde{\mathrm{q}}(x)\log_t \tilde{\mathrm{p}}(x) - \log_t \tilde{\mathrm{p}}(x)(q(x) - \tilde{\mathrm{q}}(x))dx \\
&= \int q(x)\log_t p(x) - q(x)\log_t \tilde{\mathrm{p}}(x)dx
\end{aligned}$$

■

D Mean field approximation in the t-exponential family

Mean field method is another widely used deterministic approximate method. Consider the N-dimensional multivariate t-exponential family of distribution

$$p(x;\theta) = \exp_t\left(\langle\Phi(x),\theta\rangle - g_t(\theta)\right).$$

where $x = (x_1,\ldots,x_N)$. Similar to the case of the exponential family [2], the approximation error incurred as a result of replacing p by $\tilde{\mathrm{p}}$ is given by the relative t-entropy

$$g_t(\theta) - \sup_{\tilde{\mu}}\left\{\langle\tilde{\mu},\theta\rangle + H_t(\tilde{\mathrm{p}}(x;\tilde{\theta}(\tilde{\mu})))\right\} = \inf_{\tilde{\mu}} D_t(\tilde{\mathrm{p}}\,\|p). \tag{63}$$

where

$$\tilde{\mu} = \int \Phi(x)\,\tilde{\mathrm{q}}(x;\tilde{\theta}(\tilde{\mu}))dx = \mathbb{E}_{\tilde{\mathrm{q}}}[\Phi(x)]$$

Note that unlike minimizing $D_t(p\|\,\tilde{\mathrm{p}})$ in the previous method that we introduced, the mean field method (63) is attempting to minimize $D_t(\tilde{\mathrm{p}}\,\|p)$. As in the exponential family, we choose to approximate $p(x;\theta)$ by

$$\begin{aligned}
\tilde{\mathrm{p}}(x;\tilde{\theta}(\tilde{\mu})) &= \prod_{n=1}^{N}\tilde{\mathrm{p}}_n(x_n;\tilde{\theta}_n), \text{ where} \qquad (64) \\
\tilde{\mathrm{p}}_n(x_n;\tilde{\theta}_n) &= \exp_t\left(\left\langle\Phi_n(x_n),\tilde{\theta}_n\right\rangle - g_{t,n}(\tilde{\theta}_n)\right).
\end{aligned}$$

If we fix a $n \in \{1, \ldots, N\}$ and denote $\tilde{\mathrm{p}}_j = \tilde{\mathrm{p}}_j(x_j; \tilde{\theta}_j)$, and $\tilde{\mathrm{q}}_j$ the corresponding escort distribution, then one can rewrite the KL divergence as

$$D_t(\tilde{\mathrm{p}} \,\|p) = \int \tilde{\mathrm{q}}_n \left\{ \int \log_t \tilde{\mathrm{p}}(x; \tilde{\theta}) \prod_{j \neq n} \tilde{\mathrm{q}}_j \; dx_j \right\} dx_n - \int \tilde{\mathrm{q}}_n \left\{ \int \log_t p(x; \theta) \prod_{j \neq n} \tilde{\mathrm{q}}_j \; dx_j \right\} dx_n .$$

If we keep all $\tilde{\theta}_j$ for $j \neq n$ fixed, then the KL divergence is minimized by setting

$$\int \log_t \tilde{\mathrm{p}}(x; \tilde{\theta}) \prod_{j \neq n} \tilde{\mathrm{q}}_j \; dx_j = \int \log_t p(x; \theta) \prod_{j \neq n} \tilde{\mathrm{q}}_j \; dx_j + \text{const.} \tag{65}$$

Using the fact that $\int \prod_{j \neq n} \tilde{\mathrm{q}}_j \; dx_j = 1$, we can write

$$\begin{aligned}
\int \log_t \tilde{\mathrm{p}}(x; \tilde{\theta}) \prod_{j \neq n} \tilde{\mathrm{q}}_j \; dx_j &= \frac{1}{1-t} \int \tilde{\mathrm{p}}^{1-t}(x; \tilde{\theta}) \prod_{j \neq n} \tilde{\mathrm{q}}_j \; dx_j - \frac{1}{1-t} \\
\int \log_t p(x; \theta) \prod_{j \neq n} \tilde{\mathrm{q}}_j \; dx_j &= \frac{1}{1-t} \int p^{1-t}(x; \theta) \prod_{j \neq n} \tilde{\mathrm{q}}_j \, dx_j - \frac{1}{1-t} .
\end{aligned}$$

Since $p(x; \theta)$ is t-exponential family,

$$\begin{aligned}
\int p^{1-t}(x; \theta) \prod_{j \neq n} \tilde{\mathrm{q}}_j \, dx_j &= \int (1 + (1-t) \langle \Phi(x), \theta \rangle - g_t(\theta)) \prod_{j \neq n} \tilde{\mathrm{q}}_j \, dx_j \\
&= \Big(1 + (1-t) (\Big\langle \mathbb{E}_{\tilde{\mathrm{q}}_{j \neq n}} [\Phi(x)] , \theta \Big\rangle - g_t(\theta)) \Big) ,
\end{aligned} \tag{66}$$

where we defined $\mathbb{E}_{\tilde{\mathrm{q}}_{j \neq n}} [\Phi(x)] = \int \Phi(x) \prod_{j \neq n} \tilde{\mathrm{q}}_j \, dx_j$. Similarly,

$$\begin{aligned}
&\int \tilde{\mathrm{p}}^{1-t}(x; \tilde{\theta}) \prod_{j \neq n} \tilde{\mathrm{q}}_j \; dx_j \\
&= \int \Big(1 + (1-t) \Big\langle \Phi_n(x_n), \tilde{\theta}_n \Big\rangle - g_{t,n}(\tilde{\theta}_n) \Big) \\
&\qquad \prod_{j \neq n} \Big(1 + (1-t) \Big\langle \Phi_j(x_j), \tilde{\theta}_j \Big\rangle - g_{t,j}(\tilde{\theta}_j) \Big) \tilde{\mathrm{q}}_j \; dx_j \\
&= \Big(1 + (1-t) (\Big\langle \Phi_n(x_n), \tilde{\theta}_n \Big\rangle - g_{t,n}(\tilde{\theta}_n)) \Big) \\
&\qquad \prod_{j \neq n} \Big(1 + (1-t) (\Big\langle \mathbb{E}_{\tilde{\mathrm{q}}_j} [\Phi_j(x_j)] , \tilde{\theta}_j \Big\rangle - g_{t,j}(\tilde{\theta}_j)) \Big) ,
\end{aligned} \tag{67}$$

where we defined $\mathbb{E}_{\tilde{\mathrm{q}}_j} [\Phi_j(x_j)] = \int \Phi_j(x_j) \, \tilde{\mathrm{q}}_j \, dx_j$. Putting together (66) and (67) by using (65) yields

$$\begin{aligned}
&\Big(1 + (1-t) (\Big\langle \mathbb{E}_{\tilde{\mathrm{q}}_{j \neq n}} (\Phi(x)) , \theta \Big\rangle - g_t(\theta)) \Big) \\
&= \Big(1 + (1-t) (\Big\langle \Phi_n(x_n), \tilde{\theta}_n \Big\rangle - g_{t,n}(\tilde{\theta}_n)) \Big) \\
&\qquad \prod_{j \neq n} \Big(1 + (1-t) (\Big\langle \mathbb{E}_{\tilde{\mathrm{q}}_j} [\Phi_j(x_j)], \tilde{\theta}_j \Big\rangle - g_{t,j}(\tilde{\theta}_j)) \Big) + \text{const.}
\end{aligned}$$

Absorbing all the terms which do not depend on x_n into the constant, we can rewrite the update equation for the t-exponential distributions as

$$\Big\langle \Phi_n(x_n), \tilde{\theta}_n \Big\rangle = \Big\langle \mathbb{E}_{\tilde{\mathrm{q}}_{j \neq n}} [\Phi(x)] , \theta \Big\rangle \prod_{j \neq n} \exp_t \Big(\Big\langle \mathbb{E}_{\tilde{\mathrm{q}}_j} [\Phi_j(x_j)] , \tilde{\theta}_j \Big\rangle - g_{t,j}(\tilde{\theta}_j) \Big)^{t-1} + \text{const.}$$

E Mean field approximation on the multivariate Student's t-distribution

Suppose we want to approximate a k-dimensional Student's t-distribution with degree of freedom v and parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as in (54) by k one-dimensional Student's t-distributions with degree of freedom $\tilde{\mathrm{v}}$. Recall that the t parameter of the $\exp_t$ distribution and the degree of freedom v of the Student's t-distribution are related by $\frac{1}{t-1} = \frac{v+k}{2}$. Therefore, we need to set $\frac{1}{t-1} = \frac{v+k}{2} = \frac{\tilde{\mathrm{v}}+1}{2}$, which yields $\tilde{\mathrm{v}} = v + k - 1$. We now write the approximating distribution as $\tilde{\mathrm{p}}(x;\tilde{\theta}) = \prod_n \tilde{\mathrm{p}}_n(x_n;\tilde{\theta}_n)$ where

$$\tilde{\mathrm{p}}_n(x_n;\tilde{\theta}_n) = \exp_t\left(\left\langle \tilde{\theta}_n, \Phi_n(x_n)\right\rangle - \tilde{\mathrm{g}}_{t,n}(\tilde{\theta}_n)\right).$$

If we define $\tilde{K}_n = (\tilde{\mathrm{v}}\,\tilde{\sigma}_n^2)^{-1}$ and

$$\tilde{\Psi}_n = \left(\frac{\Gamma((\tilde{\mathrm{v}}+1)/2)}{\Gamma(\tilde{\mathrm{v}}/2)(\pi\,\tilde{\mathrm{v}}\,\tilde{\sigma}_n^2)^{1/2}}\right)^{-2/(\tilde{\mathrm{v}}+1)}$$

then,

$$\tilde{\mathrm{g}}_{t,n}(\tilde{\theta}_n) = -\left(\tilde{\Psi}_n\tilde{K}_n\,\tilde{\mu}_n^2 + \tilde{\Psi}_n - 1\right)/(1-t).$$

Furthermore, $\Phi_n(x_n) = [x_n; x_n^2]$ and $\tilde{\theta}_n = [\tilde{\theta}_{n,1};\tilde{\theta}_{n,2}]$ with $\tilde{\theta}_{n,1} = -2\tilde{\Psi}_n\tilde{K}_n\,\tilde{\mu}_n/(1-t)$ and $\tilde{\theta}_{n,2} = \tilde{\Psi}_n\tilde{K}_n/(1-t)$. Now we can write

$$\begin{aligned}
\left\langle \tilde{\theta}_n, \Phi_n(x_n)\right\rangle =& \frac{1}{1-t}\tilde{\Psi}_n\cdot\left(-2\tilde{K}_n\,\tilde{\mu}_n\,x_n + \tilde{K}_n x_n^2\right)\\
\left\langle \theta, \mathbb{E}_{\tilde{\mathrm{q}}_{j\neq n}}[\Phi(\mathbf{x})]\right\rangle =& \frac{1}{1-t}\Psi\cdot\left(-2\boldsymbol{\mu}^\top\mathbf{K}\,\mathbb{E}_{\tilde{\mathrm{q}}_{j\neq n}}[\mathbf{x}] + \operatorname{tr}\left(\mathbf{K}\,\mathbb{E}_{\tilde{\mathrm{q}}_{j\neq n}}[\mathbf{x}\,\mathbf{x}^\top]\right)\right)\\
=& \frac{1}{1-t}\Psi\cdot\left(-2\boldsymbol{\mu}^\top\mathbf{k}_n\,x_n + 2\tilde{\boldsymbol{\mu}}_{j\neq n}^\top\,\mathbf{k}_{j\neq n,n}\,x_n + k_{nn}x_n^2\right) + \text{const.}
\end{aligned}$$

where $\tilde{\boldsymbol{\mu}}_{j\neq n}$ denotes the vector $\left\{\tilde{\mu}_j\right\}_{j=1\dots k,j\neq n}$, $\mathbf{k}_n$ denotes the n-th column of $\mathbf{K}$, and $\mathbf{k}_{j\neq n,n}$ denotes the n-th column of $\mathbf{K}$ after its n-th element is deleted. Recall that $\tilde{\mu}_j = \mathbb{E}_{\tilde{\mathrm{q}}_j}[x_j]$ and $\tilde{\sigma}_j^2 = \mathbb{E}_{\tilde{\mathrm{q}}_j}[x_j^2] - \mathbb{E}_{\tilde{\mathrm{q}}_j}[x_j]^2$. Therefore

$$\begin{aligned}
\exp_t\left(\left\langle\tilde{\theta}_j, \mathbb{E}_{\tilde{\mathrm{q}}_j}[\Phi_j(x_j)]\right\rangle - \tilde{\mathrm{g}}_{t,j}(\tilde{\theta}_j)\right) &= \exp_t\left(\frac{\tilde{\Psi}_j\tilde{K}_j}{1-t}\cdot\left(-2\,\tilde{\mu}_j\,\mathbb{E}_{\tilde{\mathrm{q}}_j}[x_j] + \mathbb{E}_{\tilde{\mathrm{q}}_j}[x_j^2]\right) - \tilde{\mathrm{g}}_{t,j}(\tilde{\theta}_j)\right)\\
&= \exp_t\left(\frac{\tilde{\Psi}_j\tilde{K}_j}{1-t}(-2\,\tilde{\mu}_j^2 + \tilde{\sigma}_j^2) - \tilde{\mathrm{g}}_{t,j}(\tilde{\theta}_j)\right)\\
&= \exp_t\left(\frac{1}{1-t}(\frac{\tilde{\Psi}_j}{\tilde{\mathrm{v}}} + \tilde{\Psi}_j - 1)\right).
\end{aligned}$$

The last line follows because $\tilde{K}_n = (\tilde{\mathrm{v}}\,\tilde{\sigma}_n^2)^{-1}$ and by expanding $\tilde{\mathrm{g}}_{t,j}(\tilde{\theta}_j)$.

Putting everything together, the iterative updates for the Student's t-distribution are given by

$$\begin{aligned}
\tilde{\mu}_n &= \frac{1}{k_{nn}}\left(-2\boldsymbol{\mu}^\top\mathbf{k}_n + 2\tilde{\boldsymbol{\mu}}_{j\neq n}^\top\,\mathbf{k}_{j\neq n,n}\right)\\
(\tilde{\sigma}_n)^2 &= \left(\tilde{K}_n\tilde{\Psi}_n\right)^{-(\tilde{\mathrm{v}}+1)/\tilde{\mathrm{v}}}\cdot\frac{\Gamma(\tilde{\mathrm{v}}/2)^{2/\tilde{\mathrm{v}}}\pi^{1/\tilde{\mathrm{v}}}}{\Gamma((\tilde{\mathrm{v}}+1)/2)^{2/\tilde{\mathrm{v}}}\,\tilde{\mathrm{v}}}\\
\text{where, }\ \tilde{K}_n\tilde{\Psi}_n &= \Psi k_{nn}\prod_{j\neq n}\exp_t\left(\frac{1}{1-t}(\frac{\tilde{\Psi}_j}{\tilde{\mathrm{v}}} + \tilde{\Psi}_j - 1)\right)^{t-1}
\end{aligned}$$

To empirically validate the above updates, we use a 10-dimensional Student's t-distribution with degrees of freedom $v = 5$, which corresponds to setting $t = 1.13$. Overall 500 variational updates were made and the negative relative entropy $(-D_t(\tilde{\mathrm{p}}\,\|p))$ is plotted as a function of the number of iterations in Figure 5. The graph shows that the approximate distribution monotonically gets close to the real distribution until it hits a stationary point. The stationary point indicates the optimal product of one dimensional Student's t-distributions which approximate the multi-dimensional Student's t-distribution.

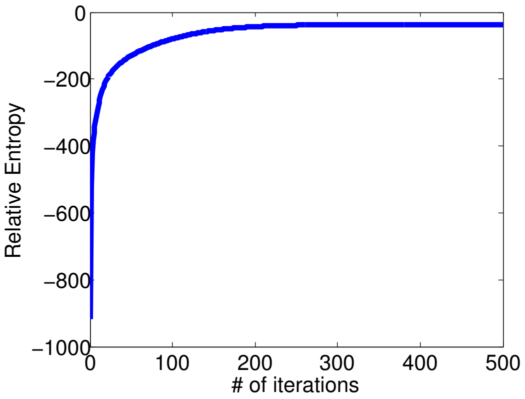

Figure 5: Negative relative entropy vs. the number of mean field updates