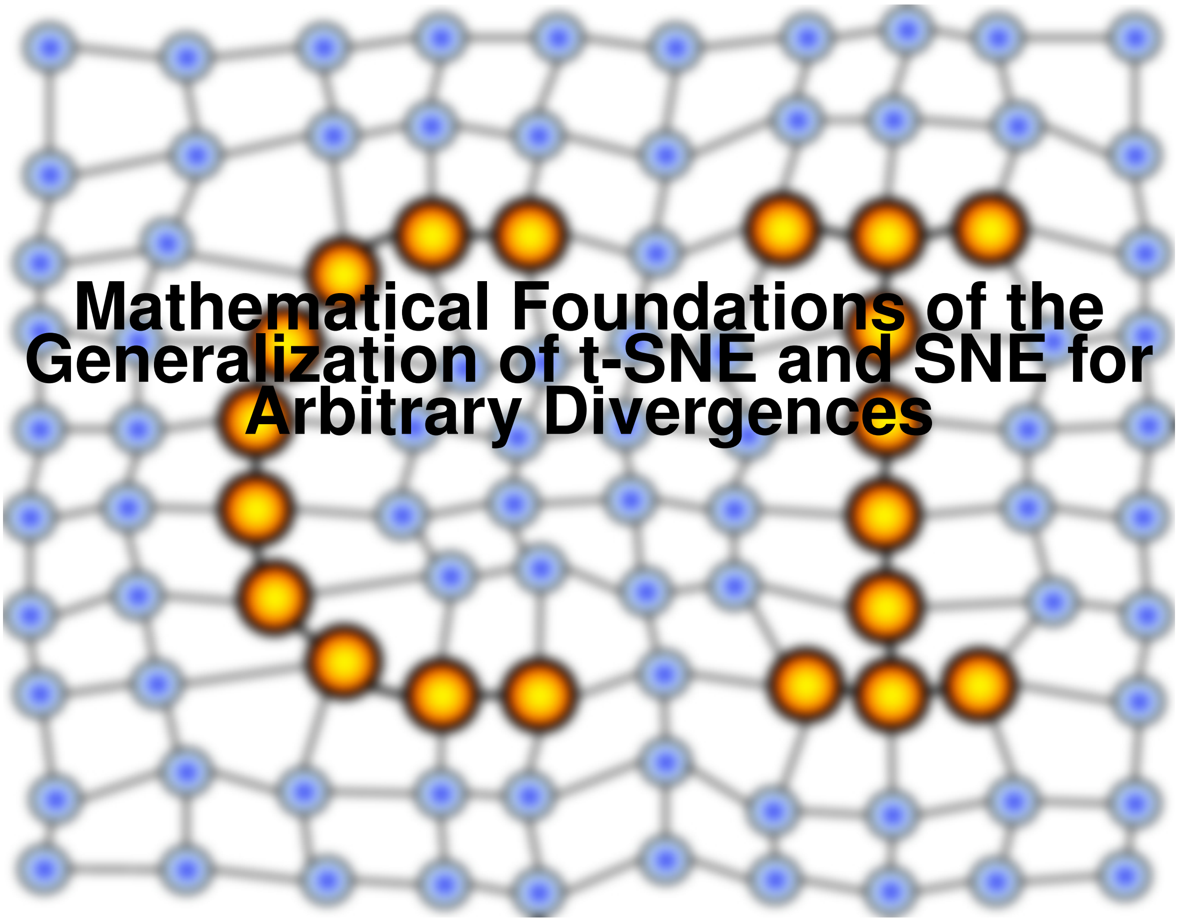
MACHINE LEARNING REPORTS



Mathematical Foundations of the Generalization of t-SNE and SNE for Arbitrary Divergences

Report 02/2010

Submitted: 01.04.2010
Published:26.04.2010

T. Villmann and S. Haase
University of Applied Sciences Mittweida, Department of MPI
Technikumplatz 17, 09648 Mittweida - Germany

Machine Learning Reports,Research group on Computational Intelligence,
http://www.mni.hs-mittweida.de/

Abstract

In this paper we offer a systematic approach of the mathematical treatment of the t-Distributed Stochastic Neighbor Embedding (t-SNE) as well as Stochastic Neighbor Embedding method. In thsi way the theory behind becomes better visible and allows an easy adaptation or exchange of the several modules contained. In particular, this concerns the underlying mathematical structure of the cost function used in the model (divergence function) which can now indepentently treated from the other components like data similarity measures or data distributions. Thereby we focus on the utilization of different divergences. This approach requires the consideration of the Fréchet-derivatives of the divergences in use. In this way, the approach can easily adapted to user specific requests.

Machine Learning Reports,Research group on Computational Intelligence,
http://www.mni.hs-mittweida.de/

1 Introduction

Methods of dimensionality reduction are challenging tools in the fields of data analysis and machine learning used in visualization as well as data compression and fusion [**?**],.

Generally, dimensionality reduction methods convert a high dimensional data set $X = \{x\}$ into low dimensional data $\Xi = \{\xi\}$. A probabilistic approach to visualize the structure of complex data sets, preserving neighbor similarities is Stochastic Neighbor Embedding (SNE), proposed by HINTON and ROWEIS [**?**]. In [9] VAN DER MAATEN and HINTON presented a technique called t-SNE, which is a variation of SNE considering another statistical model assumption for data distributions. Both methods have in common, that a probability distribution over all potential neighbors of a data point in the high-dimensional space is analyzed and described by their pairwise dissimilarities. Both, t-SNE and SNE (in a symmetric variant [9]), originally minimize a Kullback-Leibler divergence between a joint probability distribution in the high-dimensional space and a joint probability distribution in the low-dimensional space as the underlying cost function, using a gradient descent method. The pairwise similarities in the high-dimensional original data space are set to

$$p = p_{\xi\eta} = \frac{p_{\eta|\xi} + p_{\xi|\eta}}{2 \cdot \int 1 \, d\eta'} \tag{1.1}$$

with conditional probabilities

$$p_{\eta|\xi} = \frac{\exp\left(-\left\|\xi - \eta\right\|^2 \;/\; 2\sigma_\xi^2\right)}{\int \exp\left(-\left\|\xi - \eta'\right\|^2 \;/\; 2\sigma_\xi^2\right) d\eta'} \, .$$

SNE and t-SNE differ in the model assumptions according to the distribution in the low-dimensional mapping space, later defined more precisely.

In this article we provide the mathematical framework for the generalization of t-SNE and SNE, in such a way, that an arbitrary divergence can be used as cost-function for the gradient descent instead of the Kullback-Leibler divergence. The methodology is based on the Fréchet-derivative of the used divergence for the cost function [16],[17].

2 Derivation of the general cost function gradient for t-SNE and SNE

2.1 The t-SNE gradient

Let $D(p||q)$ be a divergence for non-negative integrable measure functions $p = p(r)$ and $q = q(r)$ with a domain V and $\xi, \eta \in \Xi$ distributed according to Π_Ξ [2]. Further, let $r(\xi,\eta):\Xi\times\Xi\to\mathbb{R}$ with the distribution $\Pi_r=\phi(r,\Pi_\Xi)$.

Moreover the constraints $p(r)\leq 1$ and $q(r)\leq 1$ hold for all $r{\in}V$. We denote such measure functions as *positive measures*. The *weight* of the functional p is defined as

$$W(p)=\int_V p(r)\,dr. \tag{2.1}$$

Positive measures p with weight $W(p)=1$ are denoted as (probability) density functions.

Let us define

$$r=\|\xi-\eta\|^2\ . \tag{2.2}$$

For *t-SNE*, q is obtained by means of a Student t-distribution, such that

$$q'=q\left(r\left(\xi',\eta'\right)\right)=\frac{\left(1+r\left(\xi',\eta'\right)\right)^{-1}}{\int\int\left(1+r\left(\xi,\eta\right)\right)^{-1}d\xi d\eta}$$

which we will abbreviate below for reasons of clarity as

$$\begin{aligned} q\left(r'\right) &= \frac{\left(1+r'\right)^{-1}}{\int\int\left(1+r\right)^{-1}d\xi d\eta} \\ &= f\left(r'\right)\cdot I^{-1}\ . \end{aligned}$$

Now let us consider the derivative of D with respect to ξ:

$$\begin{aligned} \frac{\partial D}{\partial\xi} &= \frac{\partial D\left(p,q\left(r\left(\xi,\eta\right)\right)\right)}{\partial\xi} \\ &= \int\int\frac{\delta D}{\delta r'}\frac{\partial r'}{\partial\xi}\,d\xi'd\eta' \\ &= \int\int\frac{\delta D}{\delta r\left(\xi',\eta'\right)}\frac{\partial r\left(\xi',\eta'\right)}{\partial\xi}\,d\xi'd\eta' \\ &= \int\int\frac{\delta D}{\delta r\left(\xi',\eta'\right)}\left[2\delta_{\xi',\xi}\left(\xi'-\eta'\right)-2\delta_{\eta',\xi}\left(\xi'-\eta'\right)\right]\,d\xi'd\eta' \\ &= 2\int\left[\frac{\delta D}{\delta r\left(\xi,\eta'\right)}\left(\xi-\eta'\right)+\int\frac{\delta D}{\delta r\left(\xi',\eta'\right)}\delta_{\eta',\xi}\left(\eta'-\xi'\right)\,d\xi'\right]\,d\eta' \\ &= 2\int\frac{\delta D}{\delta r\left(\xi,\eta'\right)}\left(\xi-\eta'\right)\,d\eta'+2\int\frac{\delta D}{\delta r\left(\xi',\xi\right)}\left(\xi-\xi'\right)\,d\xi' \\ &= 4\int\frac{\delta D}{\delta r\left(\xi,\eta\right)}\left(\xi-\eta\right)\,d\eta \end{aligned} \tag{2.3}$$

We now have to consider $\frac{\delta D}{\delta r(\xi,\eta)}$. Again, using the chain rule for functional derivatives we get

$$\frac{\delta D}{\delta r\left(\xi,\eta\right)} = \int\int\frac{\delta D}{\delta q\left(r\left(\xi',\eta'\right)\right)}\frac{\delta q\left(r\left(\xi',\eta'\right)\right)}{\delta r\left(\xi,\eta\right)}d\xi'd\eta' \tag{2.4}$$

$$= \int\frac{\delta D}{\delta q\left(r\prime\right)}\frac{\delta q\left(r\prime\right)}{\delta r}\Pi_{r'}\,dr' \tag{2.5}$$

whereby

$$\frac{\delta q\left(r'\right)}{\delta r}=\frac{\delta f\left(r'\right)}{\delta r}\cdot I^{-1}-f\left(r'\right)\cdot I^{-2}\frac{\delta I}{\delta r}$$

holds, with

$$\frac{\delta f\left(r'\right)}{\delta r}=-\delta_{r,r'}\left(1+r\right)^{-2}\text{ and }\frac{\delta I}{\delta r}=-\left(1+r\right)^{-2}$$

So we obtain

$$\begin{aligned}\frac{\delta q\left(r'\right)}{\delta r} &= \frac{-\delta_{r,r'}\left(1+r\right)^{-2}}{I}+f\left(r'\right)\cdot I^{-2}\cdot\left(1+r\right)^{-2}\\ &= \frac{-\delta_{r,r'}\left(1+r\right)^{-1}f\left(r\right)}{I}+\frac{f\left(r'\right)}{I}\frac{f\left(r\right)}{I}\left(1+r\right)^{-1}\\ &= -\delta_{r,r'}\left(1+r\right)^{-1}q\left(r\right)+q\left(r'\right)q\left(r\right)\left(1+r\right)^{-1}\\ &= -\left(1+r\right)^{-1}q\left(r\right)\left(\delta_{r,r'}-q\left(r'\right)\right)\ .\end{aligned}$$

Substituting these results in eq. (2.5), we get

$$\begin{aligned}\frac{\delta D}{\delta r} &= \int\frac{\delta D}{\delta q\left(r\prime\right)}\frac{\delta q\left(r\prime\right)}{\delta r}\Pi_{r'}\,dr'\\ &= -\left(1+r\right)^{-1}q\left(r\right)\int\frac{\delta D}{\delta q\left(r\prime\right)}\left(\delta_{r,r'}-q\left(r'\right)\right)\Pi_{r'}\,dr'\\ &= -\left(1+r\right)^{-1}q\left(r\right)\left(\frac{\delta D}{\delta q\left(r\right)}-\int\frac{\delta D}{\delta q\left(r\prime\right)}q\left(r'\right)\Pi_{r'}\,dr'\right)\end{aligned}$$

Finally, we can collect all partial results and get

$$\begin{aligned}\frac{\partial D}{\partial\xi} &= 4\int\frac{\delta D}{\delta r}\left(\xi-\eta\right)\,d\eta\\ &= -4\int\left(1+r\right)^{-1}q\left(r\right)\left(\frac{\delta D}{\delta q\left(r\right)}-\int\frac{\delta D}{\delta q\left(r\prime\right)}q\left(r'\right)\Pi_{r'}\,dr'\right)\left(\xi-\eta\right)\,d\eta \qquad (2.6)\end{aligned}$$

We now have the obvious advantage, that we can derive $\frac{\partial D}{\partial\xi}$ for several divergences $D\left(p||q\right)$ directly from (2.6), if the Fréchet derivative $\frac{\delta D}{\delta q\left(r\right)}$ of D with respect to $q\left(r\right)$ is known. Yet, for the most important classes of divergences, including Kullback-Leibler-, Rényi and Tsallis-divergences, these Fréchet derivatives can be found in [16].

2.2 The SNE gradient

In *symmetric SNE*, the pairwise similarities in the low dimensional map are given by [9]

$$q'_{SNE}=q_{SNE}\left(r\left(\xi',\eta'\right)\right)=\frac{\exp\left(-r\left(\xi',\eta'\right)\right)}{\int\int\exp\left(-r\left(\xi,\eta\right)\right)d\xi d\eta}$$

which we will abbreviate below for reasons of clarity as

$$\begin{aligned} q_{SNE}\left(r'\right) &= \frac{\exp\left(-r'\right)}{\int\int \exp\left(-r\right) d\xi d\eta} \\ &= g\left(r'\right)\cdot J^{-1}\,. \end{aligned}$$

Consequently, if we consider $\frac{\partial D}{\partial \xi}$, we can use the results from above for t-SNE. The only term that differs is the derivative of $q_{SNE}\left(r'\right)$ with respect to r. Therefore we get

$$\frac{\delta q_{SNE}\left(r'\right)}{\delta r} = \frac{\delta g\left(r'\right)}{\delta r}\cdot J^{-1} - g\left(r'\right)\cdot J^{-2}\frac{\delta J}{\delta r}$$

with

$$\frac{\delta g\left(r'\right)}{\delta r} = -\delta_{r,r'}\exp\left(-r\right) \text{ and } \frac{\delta J}{\delta r} = -\exp\left(-r\right)$$

which leads to

$$\begin{aligned} \frac{\delta q_{SNE}\left(r'\right)}{\delta r} &= \frac{-\delta_{r,r'}\exp\left(-r\right)}{J} + g\left(r'\right)\cdot J^{-2}\cdot\exp\left(-r\right) \\ &= \frac{-\delta_{r,r'}g\left(r\right)}{J} + \frac{g\left(r'\right)}{J}\frac{g\left(r\right)}{J} \\ &= -\delta_{r,r'}q_{SNE}\left(r\right) + q_{SNE}\left(r'\right)q_{SNE}\left(r\right) \\ &= -q_{SNE}\left(r\right)\left(\delta_{r,r'} - q_{SNE}\left(r'\right)\right)\,. \end{aligned}$$

Substituting these results in eq. (2.5), we get

$$\begin{aligned} \frac{\delta D}{\delta r} &= \int \frac{\delta D}{\delta q_{SNE}\left(r'\right)}\frac{\delta q_{SNE}\left(r'\right)}{\delta r}\Pi_{r'}\,dr' \\ &= -q_{SNE}\left(r\right)\int \frac{\delta D}{\delta q_{SNE}\left(r'\right)}\left(\delta_{r,r'} - q_{SNE}\left(r'\right)\right)\Pi_{r'}\,dr' \\ &= -q_{SNE}\left(r\right)\left(\frac{\delta D}{\delta q_{SNE}\left(r\right)} - \int \frac{\delta D}{\delta q_{SNE}\left(r'\right)}q_{SNE}\left(r'\right)\Pi_{r'}\,dr'\right) \end{aligned}$$

Finally, substituting this result in eq. (2.3), we obtain

$$\begin{aligned} \frac{\partial D}{\partial \xi} &= 4\int \frac{\delta D}{\delta r}\left(\xi - \eta\right)\,d\eta \\ &= -4\int q_{SNE}\left(r\right)\left(\frac{\delta D}{\delta q_{SNE}\left(r\right)} - \int \frac{\delta D}{\delta q_{SNE}\left(r'\right)}q_{SNE}\left(r'\right)\Pi_{r'}\,dr'\right)\left(\xi - \eta\right)\,d\eta \end{aligned} \tag{2.7}$$

as the general formulation of the SNE cost function gradient, which, again, utilizes the Fréchet-derivatives of the applied divergences as above for t-SNE.

3 t-SNE gradients for various divergences

In this section we explain the t-SNE gradients for various divergences. There exist a large variety of different divergences, which can be collected into several classes according to their mathematical properties and structural behavior. Here we follow the classification proposed in [2].

For this purpose, we plug the Fréchet-derivatives of these divergences or divergence families into the general gradient formula (2.6) for t-SNE. Clearly, one can convey these results easily to the general SNE gradient (2.7) in complete analogy, since its structural similarity to the t-SNE formula (2.6).

A technical remark should be given here: In the following we will abbreviate $p(r)$ by p and $p(r')$ by p'. Further, because the integration variable r is a function $r = r(\xi, \eta)$ an integration requires the weighting according to the distribution Π_r. Thus, the integration has formally to be carried out according to the differential $d\Pi_r(r)$ (Stieltjes-integral). We shorthand this simply to dr but keeping this fact in mind, i.e. by this convention, we'll drop the distribution Π_r, if it is clear from the context.

3.1 Kullback-Leibler divergence and other Bregman divergences

As a first example we show that we obtain the same result as VAN DER MAATEN and HINTON in [9] for the *Kullback-Leibler divergence*

$$D_{KL}(p||q) = \int p \log\left(\frac{p}{q}\right) dr \; .$$

The Fréchet-derivative of D_{KL} with respect to q is given by

$$\frac{\delta D_{KL}}{\delta q} = -\frac{p}{q} \; .$$

From eq. (2.6) we see that

$$\begin{aligned}\frac{\partial D_{KL}}{\partial \xi} &= 4\int (1+r)^{-1} q \left(\frac{p}{q} - \int \frac{p\prime}{q\prime} q\prime \Pi_{r'}\, dr'\right)(\xi-\eta)\; d\eta \\ &= 4\int (1+r)^{-1} q \left(\frac{p}{q} - \int p\prime \Pi_{r'}\, dr'\right)(\xi-\eta)\; d\eta \; . \qquad (3.1)\end{aligned}$$

Since the Integral $I = \int p\prime \Pi_{r'}\, dr'$ in (3.1) can be written as an double integral over all pairs of data points $I = \int\int p\prime d\xi' d\eta'$, we see from (1.1) that the integral I equals 1. So, (3.1) simplifies to

$$\begin{aligned}\frac{\partial D_{KL}}{\partial \xi} &= 4\int (1+r)^{-1} q \left(\frac{p}{q} - 1\right)(\xi-\eta)\; d\eta \\ &= 4\int (1+r)^{-1} (p-q)(\xi-\eta)\; d\eta \; . \qquad (3.2)\end{aligned}$$

This formula is exactly the differential form of the discrete version as proposed for t-SNE in [9].

The Kullback-Leibler divergence used in original SNE and t-SNE belongs to the more general class of Bregman divergences [1]. Another famous representative of this class of divergences is the *Itakura-Saito divergence* D_{IS} [6], defined as

$$D_{IS}\left(p||q\right) = \int \left[\frac{p}{q} - \log\left(\frac{p}{q}\right) - 1\right] dr$$

with the Fréchet-derivative

$$\frac{\delta D_{IS}\left(p||q\right)}{\delta q} = \frac{1}{q^2}\left(q-p\right) \ .$$

For the calculation of the gradient $\frac{\partial D_{IS}}{\partial \xi}$ we substitute the Fréchet-derivative in eq. (2.6) and obtain

$$\begin{aligned}\frac{\partial D_{IS}}{\partial \xi} &= -4\int \left(1+r\right)^{-1} q \left(\frac{1}{q^2}\left(q-p\right) - \int \frac{q\prime - p\prime}{q\prime}\Pi_{r'}\, dr'\right)\left(\xi-\eta\right)\ d\eta \\ &= -4\int \left(1+r\right)^{-1}\left(\frac{q-p}{q} - q\int \frac{q\prime - p\prime}{q\prime}\Pi_{r'}\, dr'\right)\left(\xi-\eta\right)\ d\eta \\ &= 4\int \left(1+r\right)^{-1}\left(\frac{p}{q} - 1 + q\int\left(1 - \frac{p\prime}{q\prime}\right)\Pi_{r'}\, dr'\right)\left(\xi-\eta\right)\ d\eta \ . \end{aligned} \quad (3.3)$$

One more representative of the class of Bregman-divergences is the *norm-like divergence*[1] D_θ with the parameter θ, [10]:

$$D_\theta\left(p||q\right) = \int p^\theta + \left(\theta-1\right)\ q^\theta - \theta\ p\ q^{(\theta-1)}\ dr \ .$$

The Fréchet-derivative of D_θ with respect to q is given by

$$\frac{\delta D_\theta\left(p||q\right)}{\delta q} = \theta\ \left(1-\theta\right)\left(p-q\right)\ q^{\theta-2} \ .$$

Again, we are interested in the gradient $\frac{\partial D_\theta}{\partial \xi}$, which is

$$\frac{\partial D_\theta}{\partial \xi} = 4\,\theta\ \left(\theta-1\right)\int\left(1+r\right)^{-1}\ \left(\left(p-q\right)\ q^{\theta-1} - q\int\left(p'-q'\right)\ q'^{(\theta-1)}\Pi_{r'}\ dr'\right)\left(\xi-\eta\right)\ d\eta \quad (3.4)$$

The last example of Bregman-divergences we handle in this paper is the class of $\beta-$divergences [2],[4], defined as

$$D_\beta\left(p||q\right) = \int p^\beta\left(\frac{1}{\beta-1} - \frac{1}{\beta}\right) - q^{\beta-1}\left(\frac{p}{\beta-1} + \frac{q}{\beta}\right) dr \ .$$

We use eq. (2.6) and insert the Fréchet-derivative of the $\beta-$divergences, given by

$$\frac{\delta D_\beta\left(p||q\right)}{\delta q} = q^{\beta-2}\left(q-p\right) \ .$$

[1] We note that the norm-like divergence was denoted as $\eta-$divergence in previous work [16]. We substitute the parameter η with θ in this paper to avoid confusion in denotation.

Thereby the gradient $\frac{\partial D_\beta}{\partial \xi}$ reads as

$$\frac{\partial D_\beta}{\partial \xi} = 4 \int (1+r)^{-1} \left(q^{\beta-1} (p-q) - q \int q'^{(\beta-1)} (p'-q') \ \Pi_{r'} \ dr' \right) (\xi - \eta) \ d\eta \ . \quad (3.5)$$

3.2 Csiszár's f-divergences

Next we will consider some divergences belonging to the class of Csiszár's f-divergences [3],[8],[14].

A famous example is the *Hellinger divergence* [8], defined as

$$D_H (p||q) = \int \left(\sqrt{p} - \sqrt{q} \right)^2 dr \ .$$

With the Fréchet-derivative

$$\frac{\delta D_H (p||q)}{\delta q} = 1 - \sqrt{\frac{p}{q}}$$

the gradient of D_H with respect to ξ is

$$\begin{aligned} \frac{\partial D_H}{\partial \xi} &= 4 \int (1+r)^{-1} \left(\sqrt{p\,q} - q - q \int \left(\sqrt{p'q'} - q' \right) \ \Pi_{r'} \ dr' \right) (\xi - \eta) \ d\eta \\ &= 4 \int (1+r)^{-1} \left(\sqrt{p\,q} - q \int \sqrt{p'q'} \, \Pi_{r'} \ dr' \right) (\xi - \eta) \ d\eta \ . \quad (3.6) \end{aligned}$$

The $\alpha-$*divergence* defines an important subclass of f-divergences

$$D_\alpha (p||q) = \frac{1}{\alpha(\alpha-1)} \int \left[p^\alpha q^{1-\alpha} - \alpha \, p + (\alpha - 1) \, q \right] dr$$

defines an important subclass of f-divergences [2], with the Fréchet-derivative

$$\frac{\delta D_\alpha (p||q)}{\delta q} = -\frac{1}{\alpha} \left(p^\alpha q^{-\alpha} - 1 \right)$$

can be handled as follows:

$$\begin{aligned} \frac{\partial D_\alpha}{\partial \xi} &= \frac{4}{\alpha} \int (1+r)^{-1} q \left(\left(p^\alpha q^{-\alpha} - 1 \right) - \int \left(p'^\alpha q'^{(-\alpha)} - 1 \right) q' \ \Pi_{r'} \ dr' \right) (\xi - \eta) \ d\eta \\ &= \frac{4}{\alpha} \int (1+r)^{-1} \left(p^\alpha q^{1-\alpha} - q - q \int \left(p'^\alpha q'^{(1-\alpha)} - q\prime \right) \Pi_{r'} \ dr' \right) (\xi - \eta) \ d\eta \\ &= \frac{4}{\alpha} \int (1+r)^{-1} \left(p^\alpha q^{1-\alpha} - q \int p'^\alpha q'^{(1-\alpha)} \Pi_{r'} \ dr' \right) (\xi - \eta) \ d\eta \ . \quad (3.7) \end{aligned}$$

A widely applied divergence, closely related to the $\alpha-$divergences, is the *Tsallis-divergence* [15], defined as

$$D_\alpha^T (p||q) = \frac{1}{1-\alpha} \left(1 - \int p^\alpha q^{1-\alpha} dr \right) \ .$$

Utilizing the Fréchet-derivative of D_α^T with respect to q, that is

$$\frac{\delta D_\alpha^T\left(p||q\right)}{\delta q} = -\left(\frac{p}{q}\right)^\alpha .$$

We now can compute the gradient of D_α^T with respect to ξ from eq. (2.6):

$$\begin{aligned}\frac{\partial D_\alpha^T}{\partial \xi} &= 4\int\left(1+r\right)^{-1} q\left(\left(\frac{p}{q}\right)^\alpha - \int\left(\frac{p'}{q'}\right)^\alpha q'\Pi_{r'}\ dr'\right)\left(\xi-\eta\right)\ d\eta \\ &= 4\int\left(1+r\right)^{-1}\left(p^\alpha q^{(1-\alpha)} - q\int p'^\alpha q'^{(1-\alpha)}\Pi_{r'}\ dr'\right)\left(\xi-\eta\right)\ d\eta\ , \qquad (3.8)\end{aligned}$$

which is also clear from eq. (3.7), since the Tsallis-divergence is a rescaled version of the $\alpha-$divergence for probability densities.

Now, as a last example from the class of Csiszár's f-divergences, we consider the *Rényi-divergences* [12],[13], which are also closely related to the $\alpha-$divergences and defined as

$$D_\alpha^R\left(p||q\right) = \frac{1}{\alpha-1}\log\left(\int p^\alpha q^{1-\alpha} dr\right)$$

with the corresponding Fréchet-derivative

$$\frac{\delta D_\alpha^R\left(p||q\right)}{\delta q} = -\frac{p^\alpha q^{-\alpha}}{\int p'^\alpha q'^{(1-\alpha)} dr'} .$$

Hence,

$$\begin{aligned}\frac{\partial D_\alpha^R}{\partial \xi} &= \frac{4}{\int p'^\alpha q'^{(1-\alpha)} dr'}\int\left(1+r\right)^{-1}\left(p^\alpha q^{1-\alpha} - q\int p'^\alpha q'^{(1-\alpha)}\Pi_{r'}\ dr'\right)\left(\xi-\eta\right)\ d\eta \\ &= 4\int\left(1+r\right)^{-1}\left(\frac{p^\alpha q^{1-\alpha}}{\int p'^\alpha q'^{(1-\alpha)} dr'} - q\right)\left(\xi-\eta\right)\ d\eta\ . \qquad (3.9)\end{aligned}$$

3.3 $\gamma-$*divergences*

A class of very robust divergences with respect to outliers are the $\gamma-$*divergences* [5], defined as

$$D_\gamma\left(p||q\right) = \log\left[\frac{\left(\int p^{\gamma+1} dr\right)^{\frac{1}{\gamma(\gamma+1)}}\cdot\left(\int q^{\gamma+1} dr\right)^{\frac{1}{\gamma+1}}}{\left(\int p\ q^\gamma\ dr\right)^{\frac{1}{\gamma}}}\right] .$$

The Fréchet-derivative of $D_\gamma\left(p||q\right)$ with respect to q is

$$\begin{aligned}\frac{\delta D_\gamma\left(p||q\right)}{\delta q} &= q^{\gamma-1}\left[\frac{q}{\int q^{\gamma+1} dr} - \frac{p}{\int p\ q^\gamma dr}\right] \\ &= \frac{q^\gamma}{Q_\gamma} - \frac{p\ q^{\gamma-1}}{V_\gamma} \\ &= \frac{q^\gamma V_\gamma - p\ q^{\gamma-1} Q_\gamma}{Q_\gamma V_\gamma} .\end{aligned}$$

Once again, we use eq. (2.6) to calculate the gradient of D_γ with respect to ξ :

$$\begin{aligned}\frac{\partial D_\gamma}{\partial \xi} &= -\frac{4}{Q_\gamma V_\gamma}\int (1+r)^{-1}\, q\left(q^\gamma V_\gamma - p\, q^{\gamma-1} Q_\gamma - \int \left(q'^\gamma V_\gamma - p'\, q'^{\gamma-1} Q_\gamma\right) q'\Pi_{r'}\, dr'\right)(\xi-\eta)\ d\eta \\ &= -\frac{4}{Q_\gamma V_\gamma}\int (1+r)^{-1}\, q\left(q^\gamma V_\gamma - p\, q^{\gamma-1} Q_\gamma - V_\gamma \int q'^{\gamma+1}\Pi_{r'}\, dr' + Q_\gamma \int p' q'^\gamma \Pi_{r'}\, dr'\right)(\xi-\eta)\ d\eta \\ &= -\frac{4}{Q_\gamma V_\gamma}\int (1+r)^{-1}\, q\left(q^\gamma V_\gamma - p\, q^{\gamma-1} Q_\gamma - V_\gamma Q_\gamma + Q_\gamma V_\gamma\right)(\xi-\eta)\ d\eta \\ &= 4\int (1+r)^{-1}\left(\frac{p\, q^\gamma}{\int p' q'^\gamma dr'} - \frac{q^{\gamma+1}}{\int q'^{\gamma+1} dr'}\right)(\xi-\eta)\ d\eta\ .\end{aligned} \tag{3.10}$$

For the special choice $\gamma = 1$ the $\gamma-$divergence becomes the *Cauchy-Schwarz divergence* [11],[7]:

$$D_{CS}\left(p\|q\right) = \frac{1}{2}\log\left(\int q^2 dr \cdot \int p^2 dr\right) - \log\left(\int p\cdot q\ dr\right)$$

and the gradient $\frac{\partial D_{CS}}{\partial \xi}$ for t-SNE can be directly deduced from eq. (3.10):

$$\frac{\partial D_{CS}}{\partial \xi} = 4\int (1+r)^{-1}\left(\frac{p\, q}{\int p' q'\ dr'} - \frac{q^2}{\int q'^2 dr'}\right)(\xi-\eta)\ d\eta\ . \tag{3.11}$$

Moreover, similar derivations can be made for any other divergence, since one only needs to calculate the Fréchet-derivative of the divergence and apply it to (2.6).

4 Conclusion

In this article we provide the mathematical foundation for the generalization of t-SNE and the symmetric variant of SNE. This framework enables the application of any divergence as cost-function for the gradient descent. For this purpose, we first deduced the gradient for t-SNE in a complete general case. In the result of that derivation we obtained a tool that enables us to utilize the Fréchet-derivative of any divergence. Thereafter we gave the abstract gradient also for SNE. Finally we calculated the concrete gradient for a wide range of important divergences. These results are summarized in table 1.

divergence family	formula	gradient for t-SNE
Kullback-Leibler divergence	$D_{KL}\left(p\Vert q\right)=\int p\cdot\log\left(\frac{p}{q}\right)dr$	$\frac{\partial D_{KL}}{\partial\xi}=4\int\left(1+r\right)^{-1}\left(p-q\right)\left(\xi-\eta\right)\ d\eta$
Itakura-Saito divergence	$D_{IS}\left(p\Vert q\right)=\int\left[\frac{p}{q}-\log\left(\frac{p}{q}\right)-1\right]dr$	$\frac{\partial D_{IS}}{\partial\xi}=4\int\left(1+r\right)^{-1}\left(\frac{p}{q}-1+q\int\left(1-\frac{p\prime}{q\prime}\right)\Pi_{r'}\ dr'\right)\left(\xi-\eta\right)\ d\eta$
norm-like divergence	$D_{\theta}\left(p\Vert q\right)=\int p^{\theta}+\left(\theta-1\right)\ q^{\theta}-\theta\ p\ q^{(\theta-1)}\ dr$	$\frac{\partial D_{\theta}}{\partial\xi}=4\ \theta\ \left(\theta-1\right)\int\left(1+r\right)^{-1}\Big(\ \left(p-q\right)\ q^{\theta-1}$ $-q\int\left(p'-q'\right)\ q'^{(\theta-1)}\Pi_{r'}\ dr'\Big)\left(\xi-\eta\right)\ d\eta$
β-divergence	$D_{\beta}\left(p\Vert q\right)=\int p\cdot\frac{p^{\beta-1}-q^{\beta-1}}{\beta-1}dr-\int\frac{p^{\beta}-q^{\beta}}{\beta}dr$	$\frac{\partial D_{\beta}}{\partial\xi}=4\int\left(1+r\right)^{-1}\Big(q^{\beta-1}\left(p-q\right)$ $-q\int q'^{(\beta-1)}\left(p'-q'\right)\ \Pi_{r'}\ dr'\Big)\left(\xi-\eta\right)\ d\eta$
Hellinger divergence	$D_{H}\left(p\Vert q\right)=\int\left(\sqrt{p}-\sqrt{q}\right)^{2}dr$	$\frac{\partial D_{H}}{\partial\xi}=4\int\left(1+r\right)^{-1}\Big(\sqrt{p\ q}$ $-q\int\sqrt{p'q'}\ \Pi_{r'}\ dr'\Big)\left(\xi-\eta\right)d\eta$
α-divergence	$D_{\alpha}\left(p\Vert q\right)=\frac{1}{\alpha(\alpha-1)}\int\left[p^{\alpha}q^{1-\alpha}-\alpha\cdot p+\left(\alpha-1\right)q\right]dr$	$\frac{\partial D_{\alpha}}{\partial\xi}=\frac{4}{\alpha}\int\left(1+r\right)^{-1}(p^{\alpha}q^{1-\alpha}$ $-q\int p'^{\alpha}q'^{(1-\alpha)}\Pi_{r'}\ dr'\Big)\left(\xi-\eta\right)\ d\eta$
Tsallis divergence	$D_{\alpha}^{T}\left(p\Vert q\right)=\frac{1}{1-\alpha}\left(1-\int p^{\alpha}q^{1-\alpha}dr\right)$	$\frac{\partial D_{\alpha}^{T}}{\partial\xi}=4\int\left(1+r\right)^{-1}\Big(p^{\alpha}q^{(1-\alpha)}$ $-q\int p'^{\alpha}q\prime^{(1-\alpha)}\Pi_{r'}\ dr'\Big)\left(\xi-\eta\right)\ d\eta$
Rényi divergence	$D_{\alpha}^{R}\left(p\Vert q\right)=\frac{1}{\alpha-1}\log\left(\int p^{\alpha}q^{1-\alpha}dr\right)$	$\frac{\partial D_{\alpha}^{R}}{\partial\xi}=4\int\left(1+r\right)^{-1}\left(\frac{p^{\alpha}q^{1-\alpha}}{\int p'^{\alpha}q'^{(1-\alpha)}dr'}-q\right)\left(\xi-\eta\right)\ d\eta$
$\gamma-$divergences	$D_{\gamma}\left(p\Vert q\right)=\log\left[\frac{\left(\int p^{\gamma+1}dr\right)^{\frac{1}{\gamma(\gamma+1)}}\cdot\left(\int q^{\gamma+1}dr\right)^{\frac{1}{\gamma+1}}}{\left(\int p\ q^{\gamma}\ dr\right)^{\frac{1}{\gamma}}}\right]$	$\frac{\partial D_{\gamma}}{\partial\xi}=4\int\left(1+r\right)^{-1}\left(\frac{p\ q^{\gamma}}{\int p'q'^{\gamma}dr'}-\frac{q^{\gamma+1}}{\int q'^{\gamma+1}dr'}\right)\left(\xi-\eta\right)\ d\eta$
Cauchy-Schwarz divergence	$D_{CS}\left(p\Vert q\right)=\frac{1}{2}\log\left(\frac{\int q^{2}dr\cdot\int p^{2}dr}{\int p\cdot qdr}\right)$	$\frac{\partial D_{CS}}{\partial\xi}=4\int\left(1+r\right)^{-1}\left(\frac{p\ q}{\int p'q'\ dr'}-\frac{q^{2}}{\int q'^{2}dr'}\right)\left(\xi-\eta\right)\ d\eta$

Table 1: Table of divergences and their t-SNE gradient

References

[1] L. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967.

[2] A. Cichocki, R. Zdunek, A. Phan, and S.-I. Amari. *Nonnegative Matrix and Tensor Factorizations*. Wiley, Chichester, 2009.

[3] I. Csiszár. Information-type measures of differences of probability distributions and indirect observations. *Studia Sci. Math. Hungaria*, 2:299–318, 1967.

[4] S. Eguchi and Y. Kano. Robustifying maximum likelihood estimation. Technical Report 802, Tokyo-Institute of Statistical Mathematics, Tokyo, 2001.

[5] H. Fujisawa and S. Eguchi. Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, 99:2053–2081, 2008.

[6] F. Itakura and S. Saito. Analysis synthesis telephony based on the maximum likelihood method. In *Proc. of the 6th. International Congress on Acoustics*, volume C, pages 17–20. Tokyo, 1968.

[7] R. Jenssen, J. Principe, D. Erdogmus, and T. Eltoft. The Cauchy-Schwarz divergence and Parzen windowing: Connections to graph theory and Mercer kernels. *Journal of the Franklin Institute*, 343(6):614–629, 2006.

[8] F. Liese and I. Vajda. On divergences and informations in statistics and information theory. *IEEE Transaction on Information Theory*, 52(10):4394–4412, 2005.

[9] L. Maaten and G. Hinten. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[10] F. Nielsen and R. Nock. Sided and symmetrized bregman centroids. *IEEE Transaction on Information Theory*, 55(6):2882–2903, 2009.

[11] J. C. Principe, J. F. III, and D. Xu. Information theoretic learning. In S. Haykin, editor, *Unsupervised Adaptive Filtering*. Wiley, New York, NY, 2000.

[12] A. Renyi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, 1961.

[13] A. Renyi. *Probability Theory*. North-Holland Publishing Company, Amsterdam, 1970.

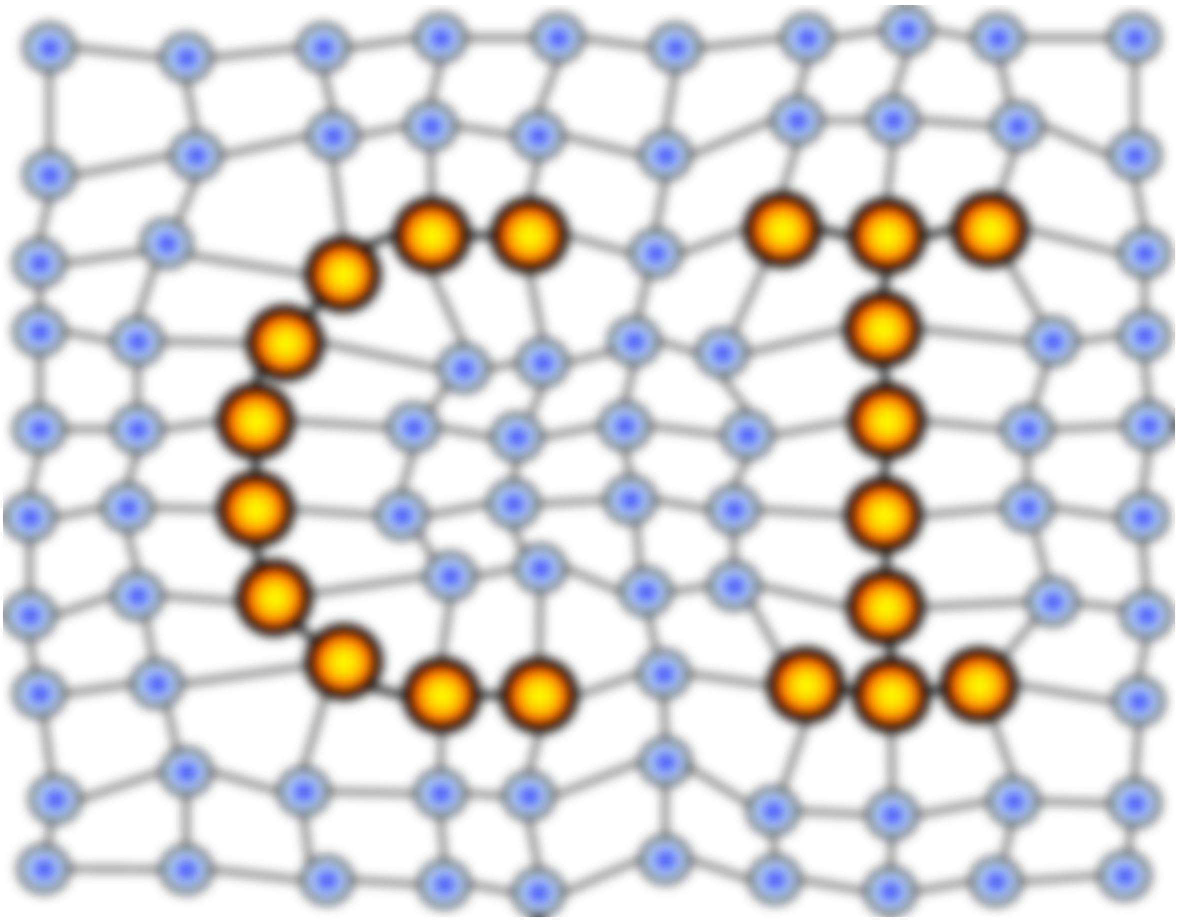
[14] I. Taneja and P. Kumar. Relative information of type s, Csiszár's f -divergence, and information inequalities. *Information Sciences*, 166:105–125, 2004.

[15] C. Tsallis. Possible generalization of Bolzmann-Gibbs statistics. *Journal oft Mathematical Physics*, 52:479–487, 1988.

[16] T. Villmann and S. Haase. Mathematical aspects of divergence based vector quantization using fréchet-derivatives - extended and revised version -. *Machine Learning Reports*, 4(MLR-01-2010):1–35, 2010. ISSN:1865-3960, http://www.uni-leipzig.de/˜compint/mlr/mlr_01_2010.pdf.

[17] T. Villmann, S. Haase, F.-M. Schleif, B. Hammer, and M. Biehl. The mathematics of divergence based online learning in vector quantization. In F. Schwenker and N. Gayar, editors, *Artificial Neural Networks in Pattern Recognition – Proc. of 4th IAPR Workshop (ANNPR'2010)*, volume 5998 of *LNAI*, pages 108–119, Berlin, 2010. Springer.

MACHINE LEARNING REPORTS



Machine Learning Reports,Research group on Computational Intelligence,

http://www.mni.hs-mittweida.de/

Impressum

Machine Learning Reports ISSN: 1865-3960

▽ Publisher/Editors

Prof. Dr. rer. nat. Thomas Villmann & Dr. rer. nat. Frank-Michael Schleif
University of Applied Sciences Mittweida
Technikumplatz 17, 09648 Mittweida - Germany•
http://www.mni.hs-mittweida.de/

▽ Acknowledgments

We would like to thank the reviewers for their time and patience.