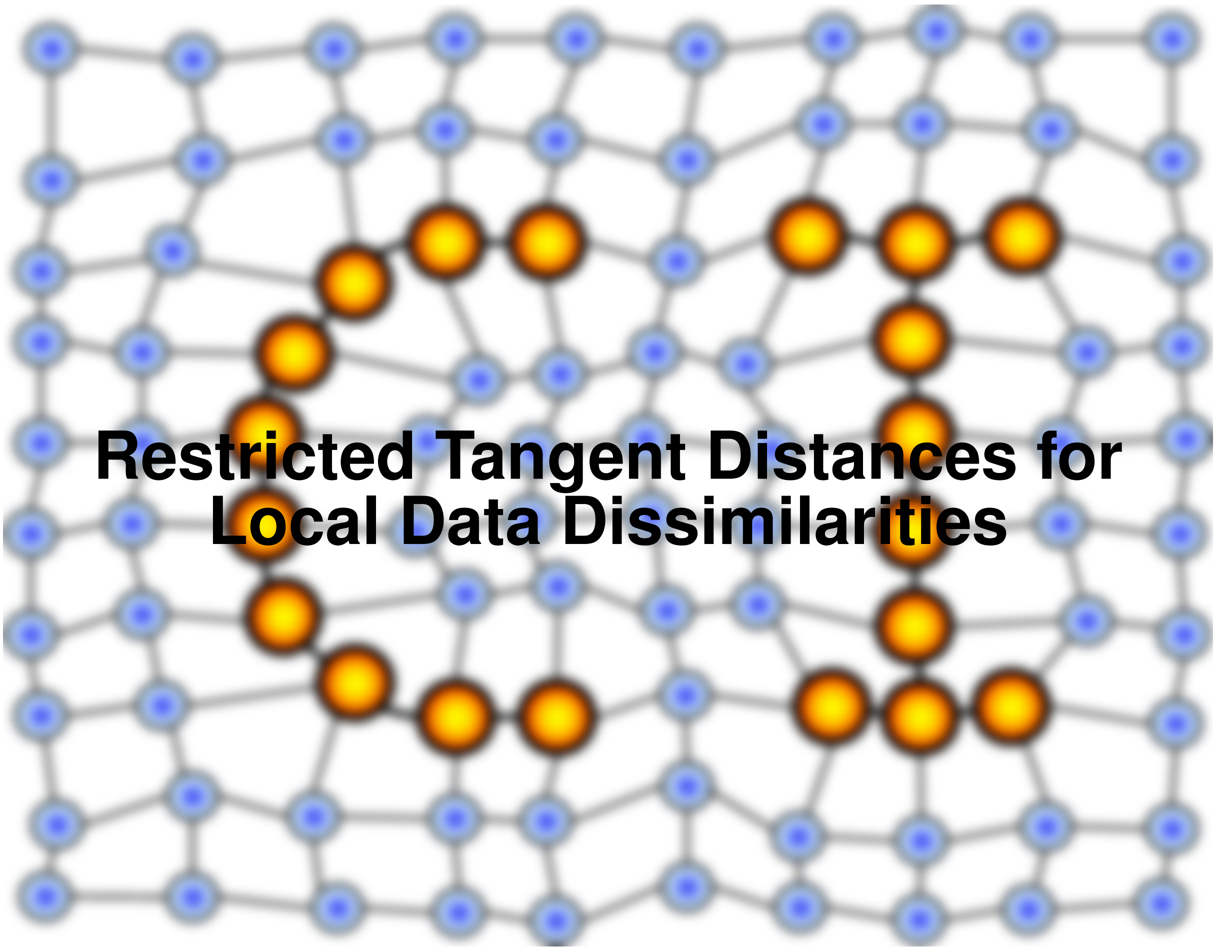
MACHINE LEARNING REPORTS



Restricted Tangent Distances for Local Data Dissimilarities

Report 01/2017
Submitted: 10.07.2016
Published: 15.12.2016

Sascha Saralajew [1] and Thomas Villmann[1]
(1) University of Applied Sciences Mittweida, Technikumplatz 17, 09648 Mittweida, Germany

Machine Learning Reports
http://www.techfak.uni-bielefeld.de/∼fschleif/mlr/mlr.html

Impressum

Publisher:	University of Applied Sciences Mittweida Technikumplatz 17, 09648 Mittweida, Germany
Editor:	Prof. Dr. Thomas Villmann Dr. Frank-Michael Schleif
Techical-Editor:	Dr. Frank-Michael Schleif
Contact:	fschleif@techfak.uni-bielefeld.de
URL:	http://techfak.uni-bielefeld.de/∼ fschleif/mlr/mlr.html
ISSN:	1865-3960

Restricted Tangent Distances for Local Data Dissimilarities

– Mathematical Treatment of the Corresponding Constrained Optimization Problem –

Sascha Saralajew[1] and Thomas Villmann[2]

[1] Dr. Ing. h.c. F. Porsche AG, Weissach (Germany),

[2] University of Applied Sciences Mittweida (Germany)

Abstract

In this technical report, we consider the tangent distance concept more deeply from a mathematical point of view. Particularly, we extend the approach to be a minimization problem over a restricted domain whereas tangent distances were treated as unrestricted problems so far. The resulting distance allows to measure the shortest distance from a given point to a well-determined *subset* of an affine subspace. In the context of tangent distances, the interpretation is that we keep care about where the defined tangent subspace approximation of the manifold structure is valid in dependence on the approximation error. Contrary, the classical tangent distance concept consists of the distance measurement regarding the whole affine subspace and, further, regarding to the whole tangent subspace approximation of the manifold structure. Therefore, the fact that the tangent subspace approximation is not globally valid is disregarded.

We introduce in this paper the restricted tangent distance over a r-orthotope domain in $\mathbb{R}^n$ with the underlying distance as the Euclidean metric. To obtain an applicable distance measure we construct the closed-form solution of the resulting minimization problem. Moreover, we show that the resulting distance measure is differentiable and, therefore, applicable to a (stochastic) gradient optimization machine learning framework. All the results are accompanied by the respective mathematical proofs.

1 Introduction

Automatic data processing in presence of distortions or significant data variations is still a challenging problem in machine learning. Simard proposed the tangent distance to handle such data transformations in distance based machine learning methods [1] or algorithms which are based on dissimilarities [2]. In general, Simard assumed that the data points are representatives of unknown manifold structures. Given two data points, the tangent distance is defined as the smallest Euclidean distance between the two tangential subspaces of the manifolds for the considered points. Frequently in machine learning environments, the tangent subspaces are estimated or determined in advance as a part of the pre-processing.

The tangent distance concept was refined over the last years and applied to several distinct machine learning methods [3, 4, 5]. A major advancement was the investigation of single-sided (one-sided) tangent distances [6, 7]. At those tangent distances the tangent subspace is assumed at only one point which leads to a closed-form solution of the optimization problem.

In previous works, the authors of the present paper used the single-sided tangent distance concept in the framework of Generalized Learning Vector Quantization (GLVQ) [8] and introduced this distance as the unrestricted optimization problem

$$\mathsf{d}_{\mathbb{R}^r}(\mathbf{v}, \mathfrak{w}(\boldsymbol{\theta})) = \min_{\boldsymbol{\theta}\in\mathbb{R}^r} d(\mathbf{v}, \mathbf{w} + \mathbf{W}\boldsymbol{\theta}) \tag{1}$$

with $d : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}_0^+$ being an arbitrary chosen underlying distance measure. The parameter $r = \dim(\mathbf{W})$ is the tangent subspace dimension. Here, the argument $\mathfrak{w}(\boldsymbol{\theta}) = \mathbf{w} + \mathbf{W}\boldsymbol{\theta}$ determines the tangent subspace at the point $\mathbf{w} \in \mathbb{R}^n$ with the tangent basis $\mathbf{W} \in \mathbb{R}^{n\times r}$ and the parameter vector $\boldsymbol{\theta} \in \mathbb{R}^r$. We denote this single-sided distance simply as tangent distance or tangent metric. It can be proved that this definition is equivalent to a Hausdorff-metric if the underlying distance is a translation invariant metric [9]. For example, the Euclidean metric

$$d_E(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^T (\mathbf{x} - \mathbf{y})} \tag{2}$$

satisfies this requirement. In this case, the closed-form solution of the optimization problem (1) becomes

$$\hat{\boldsymbol{\theta}} = \mathbf{W}^T (\mathbf{v} - \mathbf{w}) \tag{3}$$

provided $\mathbf{W}^T\mathbf{W} = \mathbf{I}_r$ is valid, i. e. an orthonormal basis is assumed for the subspace described by $\mathbf{W}$. In the following, we suppose that this property always holds.

Yet, the definition (1) of the tangent distance further presumes implicitly that the tangent subspace is a globally valid approximation of the unknown manifold structure [10]. Therefore, a natural extension of the model (1) is to consider a *restricted parameter domain* $D \subset \mathbb{R}^r$ for the tangent subspace when determining the distance, i. e. the previous unrestricted optimization problem

(1) becomes a restricted optimization problem

$$\mathsf{d}_D\left(\mathbf{v}, \mathfrak{w}(\boldsymbol{\theta})\right) = \min_{\boldsymbol{\theta}\in D} d(\mathbf{v}, \mathbf{w} + \mathbf{W}\boldsymbol{\theta}) \tag{4}$$

in order to reduce the approximation error.

The aim of this technical report is to solve the restricted optimization problem (4) in the case of an underlying Euclidean metric (2) and a r-orthotope[1] domain D in $\mathbb{R}^n$.

For this purpose, we define the used notations and give some principal definitions in the first section. After that, we solve the optimization problem by computing a Karush-Kuhn-Tucker point. It is followed by computing the gradients of the closed-form solution of the derived distance and concluding remarks.

2 Fundamentals

We denote by $\mathbf{I}_n$ the $n \times n$-dimensional identity matrix and by $\mathbf{e}_i$ the i-th n-dimensional unit vector[2]. Further, the n-dimensional one vector is obtained by $\mathbf{1}_n = \sum_{i=1}^{n} \mathbf{e}_i$. Additionally, we denote the n-dimensional zero vector by $\mathbf{0}_n$. The symbol x_i indicates the i-th element of the vector $\mathbf{x}$ and, similarly, the notation x_{ij} the element (i,j) of the matrix $\mathbf{X}$. The set of non-negative real values is defined by $\mathbb{R}_{\geq 0} = \{x \in \mathbb{R} | x \geq 0\}$ and, further, the set of positive real values by $\mathbb{R}_{>0} = \{x \in \mathbb{R} | x > 0\}$.

Definition 1. The Heaviside function of a real value $x \in \mathbb{R}$ is defined as

$$H(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{else} \end{cases}$$

and for a vector $\mathbf{x} \in \mathbb{R}^n$ as

$$H\left(\mathbf{x}\right) = \left(\begin{array}{cccc} H(x_1) & 0 & \cdots & 0 \\ 0 & H(x_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & H(x_n) \end{array}\right).$$

Definition 2. The signum function of a real value $x \in \mathbb{R}$ is defined as

$$sgn(x) = \begin{cases} -1 & \text{if } x < 0 \\ 0 & \text{if } x = 0 \\ 1 & \text{else} \end{cases}$$

[1] It is also denoted as r-dimensional hyperrectangle or r-dimensional box.

[2] We skip an indexing of the dimension at the unit vector and suggest that the dimension is clearly given from the context.

and for a vector $\mathbf{x} \in \mathbb{R}^n$ as

$$sgn\left(\mathbf{x}\right) = \begin{pmatrix} sgn(x_1) & 0 & \cdots & 0 \\ 0 & sgn(x_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & sgn(x_n) \end{pmatrix}.$$

Definition 3. The absolute value function of a real value $x \in \mathbb{R}$ is defined as

$$|x| = \begin{cases} -x & \text{if } x < 0 \\ x & \text{else} \end{cases}$$

and for a vector $\mathbf{x} \in \mathbb{R}^n$ as

$$|\mathbf{x}| = \begin{pmatrix} |x_1| \\ |x_2| \\ \vdots \\ |x_n| \end{pmatrix}.$$

Definition 4. For vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ the following relations are defined:

- $\mathbf{x} \leqq \mathbf{y} \Longleftrightarrow x_i \leq y_i \ \forall i = 1, ..., n$
- $\mathbf{x} \geqq \mathbf{y} \Longleftrightarrow x_i \geq y_i \ \forall i = 1, ..., n$
- $\mathbf{x} < \mathbf{y} \Longleftrightarrow x_i < y_i \ \forall i = 1, ..., n$
- $\mathbf{x} > \mathbf{y} \Longleftrightarrow x_i > y_i \ \forall i = 1, ..., n$

Definition 5. The standard form of an optimization problem (f, D) is

$$\min_{\mathbf{x}} f(\mathbf{x})$$

subject to

$$\begin{aligned} g_i(\mathbf{x}) &\leq 0 \ , i = 1, 2, ..., m \\ h_j(\mathbf{x}) &= 0 \ , j = 1, 2, ..., p \end{aligned}$$

where $f(\mathbf{x}) : \mathbb{R}^n \longrightarrow \mathbb{R}$ is called the objective function to be minimized with respect to the variable $\mathbf{x}$. The constraints define the feasible domain (set of feasible points or set of feasible solutions)

$$D = \{\mathbf{x} \in \mathbb{R}^n | g_i(\mathbf{x}) \leq 0, \ h_j(\mathbf{x}) = 0, \ \ i \in \{1, 2, ..., m\}, \ \ j \in \{1, 2, ..., p\}\}$$

of the optimization problem.

Definition 6. Let (f_1, D_1) and (f_2, D_2) be two optimization problems. Further, let $\mathcal{X}_1^* \subseteq D_1$ be the set of optimal solutions for (f_1, D_1) and $\mathcal{X}_2^* \subseteq D_2$ be the set of optimal solutions for (f_2, D_2). The optimization problems are called equivalent if there exist a bijective function $q : \mathcal{X}_1^* \longrightarrow \mathcal{X}_2^*$.

In the following we restrict the tangent distance (4) to the underlying distance as the Euclidean metric (2) and define:

Definition 7. The rectangular restricted tangent distance $\mathsf{d}_{\tilde{D}}\left(\mathbf{v}, \tilde{\mathfrak{w}}(\tilde{\boldsymbol{\theta}})\right)$, abbreviated by rrTD, is defined as the minimal value of the optimization problem $\left(\tilde{f}, \tilde{D}\right)$ with the objective function

$$\tilde{f}\left(\tilde{\boldsymbol{\theta}}\right) = d_E(\mathbf{v}, \tilde{\mathbf{w}} + \mathbf{W}\tilde{\boldsymbol{\theta}})$$

and the feasible domain

$$\tilde{D} = [a_1, b_1] \times [a_2, b_2] \times ... \times [a_r, b_r] \subset \mathbb{R}^r$$

with $a_i \leq b_i$, $i = 1, 2, ..., r$.

We collect the lower bounds a_i into the vector $\mathbf{a} = (a_1, a_2, ..., a_r)^T$ and upper bounds b_i into the vector $\mathbf{b} = (b_1, b_2, ..., b_r)^T$. Note that $\tilde{D}$ is a r-orthotope and, further, the set $\left\{\mathbf{x} \in \mathbb{R}^n | \exists \tilde{\boldsymbol{\theta}} \in \tilde{D} : \mathbf{x} = \tilde{\mathbf{w}} + \mathbf{W}\tilde{\boldsymbol{\theta}}\right\}$ is also a r-orthotope in $\mathbb{R}^n$ or, in other words, a r-dimensional hyperrectangle in $\mathbb{R}^n$.

Lemma 1. *The optimization problem* (f, D) *with*

$$f\left(\boldsymbol{\theta}\right) = d_E(\mathbf{v}, \mathbf{w} + \mathbf{W}\boldsymbol{\theta})$$

where $\mathbf{w} = \tilde{\mathbf{w}} + \frac{1}{2}\mathbf{W}(\mathbf{b} + \mathbf{a})$ *and*

$$D = [-c_1, c_1] \times [-c_2, c_2] \times ... \times [-c_r, c_r] \subset \mathbb{R}^r$$

with $\mathbf{c} = \frac{1}{2}\left(\mathbf{b} - \mathbf{a}\right)$ *and the function*

$$q : \mathbb{R}^r \longrightarrow \mathbb{R}^r : q(\mathbf{x}) = \mathbf{x} + \frac{1}{2}\left(\mathbf{b} + \mathbf{a}\right)$$

is equivalent to $\left(\tilde{f}, \tilde{D}\right)$ *according to the previous definition.*

Proof. At first it is proven that for any $\mathbf{x} \in D$ the statement $q(\mathbf{x}) \in \tilde{D}$ holds:

$$\begin{aligned} \mathbf{x} &\geqq -\mathbf{c} &&\wedge \quad \mathbf{x} \leqq \mathbf{c} \\ \mathbf{x} &\geqq -\frac{1}{2}\left(\mathbf{b} - \mathbf{a}\right) &&\wedge \quad \mathbf{x} \leqq \frac{1}{2}\left(\mathbf{b} - \mathbf{a}\right) \\ q(\mathbf{x}) - \frac{1}{2}\left(\mathbf{b} + \mathbf{a}\right) &\geqq -\frac{1}{2}\left(\mathbf{b} - \mathbf{a}\right) &&\wedge \quad q(\mathbf{x}) - \frac{1}{2}\left(\mathbf{b} + \mathbf{a}\right) \leqq \frac{1}{2}\left(\mathbf{b} - \mathbf{a}\right) \\ q(\mathbf{x}) &\geqq \mathbf{a} &&\wedge \quad q(\mathbf{x}) \leqq \mathbf{b} \end{aligned}$$

Hence, the statement is valid. Further, the restriction of the function $q : D \longrightarrow \tilde{D}$ is injective since q is a shift function. Moreover, for an arbitrary $\tilde{\mathbf{x}} \in \tilde{D}$ we can conclude:

$$\begin{aligned} \tilde{\mathbf{x}} &\geqq \mathbf{a} &&\wedge \quad \tilde{\mathbf{x}} \leqq \mathbf{b} \\ \tilde{\mathbf{x}} &\geqq -\frac{1}{2}\left(\mathbf{b} - \mathbf{a}\right) + \frac{1}{2}\left(\mathbf{b} + \mathbf{a}\right) &&\wedge \quad \tilde{\mathbf{x}} \leqq \frac{1}{2}\left(\mathbf{b} - \mathbf{a}\right) + \frac{1}{2}\left(\mathbf{b} + \mathbf{a}\right) \\ \tilde{\mathbf{x}} - \frac{1}{2}\left(\mathbf{b} + \mathbf{a}\right) &\geqq -\frac{1}{2}\left(\mathbf{b} - \mathbf{a}\right) &&\wedge \quad \tilde{\mathbf{x}} - \frac{1}{2}\left(\mathbf{b} + \mathbf{a}\right) \leqq \frac{1}{2}\left(\mathbf{b} - \mathbf{a}\right) \\ \tilde{\mathbf{x}} - \frac{1}{2}\left(\mathbf{b} + \mathbf{a}\right) &\geqq -\mathbf{c} &&\wedge \quad \tilde{\mathbf{x}} - \frac{1}{2}\left(\mathbf{b} + \mathbf{a}\right) \leqq \mathbf{c} \end{aligned}$$

and, therefore, $\mathbf{x} = \tilde{\mathbf{x}} - \frac{1}{2}(\mathbf{b}+\mathbf{a}) \in D$ exist such that

$$\begin{aligned} q(\mathbf{x}) &= q\left(\tilde{\mathbf{x}} - \frac{1}{2}(\mathbf{b}+\mathbf{a})\right) \\ &= \tilde{\mathbf{x}} \end{aligned}$$

holds. This proves that the function $q : D \longrightarrow \tilde{D}$ is surjective. Injectivity and surjectivity yield bijectivity.

In the next step we consider some properties of these optimization problems: The feasible domains D and $\tilde{D}$ are compact and the objective functions of (f, D) and $\left(\tilde{f}, \tilde{D}\right)$ are continuous. Thus, the extreme-value-theorem holds and states that there exist sets of optimal solutions $\mathcal{X}^*_D$ and $\mathcal{X}^*_{\tilde{D}}$. Moreover, the objective functions are strictly convex and, therefore, the sets of optimal solutions are singletons [11]. Thus, it is sufficient to prove that the optimal solution $\left\{\boldsymbol{\theta}^*\right\} = \mathcal{X}^*_D$ is mapped to the optimal solution $\left\{\tilde{\boldsymbol{\theta}}^*\right\} = \mathcal{X}^*_{\tilde{D}}$ by $q(\boldsymbol{\theta}^*) = \tilde{\boldsymbol{\theta}}^*$. For $\boldsymbol{\theta}^*$ of (f, D) holds:

$$\begin{aligned} \min_{\boldsymbol{\theta}\in D} d_E(\mathbf{v}, \mathbf{w} + \mathbf{W}\boldsymbol{\theta}) &= d_E(\mathbf{v}, \mathbf{w} + \mathbf{W}\boldsymbol{\theta}^*) \\ &= d_E\left(\mathbf{v}, \tilde{\mathbf{w}} + \frac{1}{2}\mathbf{W}(\mathbf{b}+\mathbf{a}) + \mathbf{W}\left(q(\boldsymbol{\theta}^*) - \frac{1}{2}(\mathbf{b}+\mathbf{a})\right)\right) \\ &= d_E\left(\mathbf{v}, \tilde{\mathbf{w}} + \mathbf{W}q(\boldsymbol{\theta}^*)\right) \qquad (5) \end{aligned}$$

Now, we construct a proof by contradiction: Suppose that $q(\boldsymbol{\theta}^*) \neq \tilde{\boldsymbol{\theta}}^*$. By the previous results we know that $q(\boldsymbol{\theta}^*) \in \tilde{D}$ holds and $\tilde{\boldsymbol{\theta}}^*$ is unique. Therefore, it follows that

$$\begin{aligned} d_E\left(\mathbf{v}, \tilde{\mathbf{w}} + \mathbf{W}q(\boldsymbol{\theta}^*)\right) &> \min_{\tilde{\boldsymbol{\theta}}\in\tilde{D}} d_E\left(\mathbf{v}, \tilde{\mathbf{w}} + \mathbf{W}\tilde{\boldsymbol{\theta}}\right) \\ &= d_E\left(\mathbf{v}, \tilde{\mathbf{w}} + \mathbf{W}\tilde{\boldsymbol{\theta}}^*\right) \\ &= d_E\left(\mathbf{v}, \mathbf{w} - \frac{1}{2}\mathbf{W}(\mathbf{b}+\mathbf{a}) + \mathbf{W}\left(q^{-1}(\tilde{\boldsymbol{\theta}}^*) + \frac{1}{2}(\mathbf{b}+\mathbf{a})\right)\right) \\ &= d_E\left(\mathbf{v}, \mathbf{w} + \mathbf{W}q^{-1}(\tilde{\boldsymbol{\theta}}^*)\right) \end{aligned}$$

is valid. Since q is bijective, the assumption implies that $q^{-1}(\tilde{\boldsymbol{\theta}}^*) \neq \boldsymbol{\theta}^*$. Furthermore, we know that $q^{-1}(\tilde{\boldsymbol{\theta}}^*) \in D$ holds and $\boldsymbol{\theta}^*$ is unique. Finally, taking into account the solution Equation (5) of the *minimal* element we conclude that

$$d_E\left(\mathbf{v}, \tilde{\mathbf{w}} + \mathbf{W}q(\boldsymbol{\theta}^*)\right) \quad < \quad d_E\left(\mathbf{v}, \mathbf{w} + \mathbf{W}q^{-1}(\tilde{\boldsymbol{\theta}}^*)\right)$$

which contradicts the previous inequality. Therefore, the assumption $q(\boldsymbol{\theta}^*) \neq \tilde{\boldsymbol{\theta}}^*$ must be wrong and, hence, we show $q(\boldsymbol{\theta}^*) = \tilde{\boldsymbol{\theta}}^*$. □

Applying the Lemma 1, we know that it is sufficient to solve the optimization problem (f, D) for the centered r-orthotope D to get the solution of

$\left(\tilde{f}, \tilde{D}\right)$. More precisely, the information of the non-centered r-orthotope $\tilde{D}$ can be embedded into the point $\mathbf{w}$ of the restricted tangent distance. Hence, we will concentrate on the optimization problem (f, D)

$$\begin{aligned} \mathsf{d}_D\left(\mathbf{v}, \mathfrak{w}(\boldsymbol{\theta})\right) &= \min_{\boldsymbol{\theta}\in D} f\left(\boldsymbol{\theta}\right) \\ &= \min_{\boldsymbol{\theta}\in D} d_E(\mathbf{v}, \mathbf{w} + \mathbf{W}\boldsymbol{\theta}) \end{aligned}$$

over the feasible domain

$$D = [-c_1, c_1] \times [-c_2, c_2] \times ... \times [-c_r, c_r] \subset \mathbb{R}^r$$

with $\mathbf{c} \in \mathbb{R}^r_{\geq 0}$ to solve the rrTD problem.

3 Solving the Optimization Problem

Theorem 1. *The set of optimal solutions $\mathcal{X}^*_D$ of (f, D) is a singleton with the element $\boldsymbol{\theta}^* \in \mathcal{X}^*_D$ defined to be*

$$\boldsymbol{\theta}^* = H\left(\mathbf{c} - \left|\hat{\boldsymbol{\theta}}\right|\right)\hat{\boldsymbol{\theta}} + sgn\left(\hat{\boldsymbol{\theta}}\right)\left(\mathbf{I}_r - H\left(\mathbf{c} - \left|\hat{\boldsymbol{\theta}}\right|\right)\right)\mathbf{c} \tag{6}$$

where $\hat{\boldsymbol{\theta}} = \mathbf{W}^T\left(\mathbf{v} - \mathbf{w}\right)$ is the solution (3) of the unrestricted tangent distance problem (1) with underlying distance as the Euclidean metric (2).

Proof. As stated in the proof of the Lemma 1, there exists a solution of (f, D) and the solution is unique. We simplify the proof by solving (f, D) for the squared distance, which is an equivalent optimization problem with q defined to be the identity function. Hence, the standard form of (f, D) is given by

$$\begin{aligned} \min_{\boldsymbol{\theta}\in\mathbb{R}^r} f(\boldsymbol{\theta}) &= \min_{\boldsymbol{\theta}\in\mathbb{R}^r} d^2_E(\mathbf{v}, \mathbf{w} + \mathbf{W}\boldsymbol{\theta}) \\ g_i(\boldsymbol{\theta}) &\leq 0 \ \ , i = 1, 2, ..., 2r \end{aligned} \tag{7}$$

with:

$$g_i(\boldsymbol{\theta}) = \begin{cases} \mathbf{e}_i^T\left(\boldsymbol{\theta} - \mathbf{c}\right) & \text{if } i \leq r \\ -\mathbf{e}_{i-r}^T\left(\boldsymbol{\theta} + \mathbf{c}\right) & \text{else} \end{cases} \tag{8}$$

Since (f, D) is a convex optimization problem and all constraints (8) are affine functions, the Slater condition is satisfied. Moreover, because both the objective function (7) and the constraints are differentiable, the vector $\boldsymbol{\theta}^*$ is optimal if and only if there exists a Karush-Kuhn-Tucker (KKT) point $(\boldsymbol{\theta}^*, \boldsymbol{\mu}^*) \in \mathbb{R}^{r+2r}$ [11]. A point $(\boldsymbol{\theta}^*, \boldsymbol{\mu}^*) \in \mathbb{R}^{r+2r}$ fits the KKT conditions and is called a KKT point of (f, D) if

$$\begin{aligned} \nabla f(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} + \sum_{i=1}^{2r}\mu_i^*\nabla g_i(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} &= \mathbf{0}_r \\ g_i(\boldsymbol{\theta}^*) &\leq 0 \ \ , i = 1, 2, ..., 2r \\ \mu_i^* &\geq 0 \ \ , i = 1, 2, ..., 2r \\ \mu_i^* g_i(\boldsymbol{\theta}^*) &= 0 \ \ , i = 1, 2, ..., 2r \end{aligned}$$

with the gradients:

$$\begin{aligned}\nabla f(\boldsymbol{\theta}) &= -2\mathbf{W}^T\left(\mathbf{v}-\mathbf{w}-\mathbf{W}\boldsymbol{\theta}\right)\\ \nabla g_i(\boldsymbol{\theta}) &= \begin{cases}\mathbf{e}_i & \text{if } i\leq r\\ -\mathbf{e}_{i-r} & \text{else}\end{cases}\end{aligned}$$

Thus, we can prove the theorem by showing the existence of a KKT point.

The stationarity condition can be simplified to

$$\begin{aligned}-2\mathbf{W}^T\left(\mathbf{v}-\left(\mathbf{w}+\mathbf{W}\boldsymbol{\theta}^*\right)\right)+\sum_{i=1}^{r}\mathbf{e}_i\left(\mu_i^*-\mu_{i+r}^*\right) &= \mathbf{0}_r\\ 2\left(\mathbf{W}^T\left(\mathbf{v}-\mathbf{w}\right)-\boldsymbol{\theta}^*\right) &= \sum_{i=1}^{r}\mathbf{e}_i\left(\mu_i^*-\mu_{i+r}^*\right)\\ 2\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right) &= \sum_{i=1}^{r}\mathbf{e}_i\left(\mu_i^*-\mu_{i+r}^*\right)\end{aligned}$$

using the property $\mathbf{W}^T\mathbf{W}=\mathbf{I}_r$.

Now, we reformulate the KKT conditions of (f, D) as

$$2\left(\hat{\theta}_i-\theta_i^*\right) = \mu_i^*-\mu_{i+r}^* \quad (9)$$

$$\begin{aligned}\theta_i^*-c_i &\leq 0\\ -\theta_i^*-c_i &\leq 0\end{aligned} \quad (10)$$

$$\begin{aligned}\mu_i^* &\geq 0\\ \mu_{i+r}^* &\geq 0\end{aligned} \quad (11)$$

$$\begin{aligned}\mu_i^*\left(\theta_i^*-c_i\right) &= 0\\ -\mu_{i+r}^*\left(\theta_i^*+c_i\right) &= 0\end{aligned} \quad (12)$$

for $i=1,2,...r$. The element θ_i^* is a function of only $\hat{\theta}_i$ and c_i and, therefore, the KKT conditions are separable regarding the index i. Hence, we simplify the proof by showing that for each $i\in\{1,2,...,r\}$ there exists a point $(\theta_i^*,\mu_i^*,\mu_{i+r}^*)\in\mathbb{R}^3$ such that the above conditions are satisfied.

In the following we assume $i\in\{1,2,...,r\}$ to be arbitrary but fixed. Since $c_i\in\mathbb{R}_{\geq 0}$, we prove that θ_i^* is primal feasible (10) by pooling of the equations to

$$\begin{aligned}\begin{matrix}\theta_i^*-c_i \leq 0\\ -\theta_i^*-c_i \leq 0\end{matrix} &\iff \begin{matrix}\theta_i^* \leq c_i\\ \theta_i^* \geq -c_i\end{matrix}\\ &\iff |\theta_i^*| \leq c_i\\ &\iff \theta_i^*\in[-c_i,c_i]\end{aligned}$$

and, further,

$$\left|H\left(c_i-\left|\hat{\theta}_i\right|\right)\hat{\theta}_i+sgn\left(\hat{\theta}_i\right)\left(1-H\left(c_i-\left|\hat{\theta}_i\right|\right)\right)c_i\right|\leq c_i\,. \quad (13)$$

The validity of the equation (13) is obtained by a case analysis:

Case 1. Suppose $\left|\hat{\theta}_i\right| > c_i$. Since $c_i \in \mathbb{R}_{\geq 0}$ it follows that

$$H\left(c_i - \left|\hat{\theta}_i\right|\right) = 0 \quad \text{and} \quad sgn(\hat{\theta}_i) = \pm 1$$

which implies

$$\theta_i^* = \begin{cases} c_i & \text{if } \hat{\theta}_i > c_i \\ -c_i & \text{if } \hat{\theta}_i < -c_i \end{cases}$$

and, hence, the equation (13) simplifies to

$$|\pm c_i| \leq c_i$$

which is true.

Case 2. Suppose $\left|\hat{\theta}_i\right| \leq c_i$. It follows that

$$H\left(c_i - \left|\hat{\theta}_i\right|\right) = 1$$

which implies

$$\theta_i^* = \hat{\theta}_i$$

and, hence, the equation (13) simplifies to

$$\left|\hat{\theta}_i\right| \quad \leq \quad c_i$$

which is true.

Combining both results for θ_i^* of the above case analysis, we obtain

$$\theta_i^* = \begin{cases} c_i & \text{if } \hat{\theta}_i > c_i \\ \hat{\theta}_i & \text{if } \left|\hat{\theta}_i\right| \leq c_i \\ -c_i & \text{else} \end{cases}$$

which is equivalent to

$$\theta_i^* = \begin{cases} c_i & \text{if } \hat{\theta}_i \geq c_i \\ \hat{\theta}_i & \text{if } \left|\hat{\theta}_i\right| < c_i \\ -c_i & \text{else} \end{cases} \tag{14}$$

Finally, we show the remaining KKT conditions again by a case analysis:

Case 1. Assume $c_i = 0$. This implies immediately that $\theta_i^* = 0$ and that both constraints (10) are active, i. e. $\theta_i^* - c_i = 0$ and $-\theta_i^* - c_i = 0$. Moreover, the complementary slackness conditions (12) are satisfied for all dual feasible points (11). The stationarity condition (9) simplifies to

$$2\hat{\theta}_i = \mu_i^* - \mu_{i+r}^* \,.$$

By defining

$$\mu_i^* = \begin{cases} 2\hat{\theta}_i & \text{if } \hat{\theta}_i \geq 0 \\ 0 & \text{else} \end{cases}$$

and

$$\mu_{i+r}^* = \begin{cases} 0 & \text{if } \hat{\theta}_i \geq 0 \\ -2\hat{\theta}_i & \text{else} \end{cases}$$

all KKT conditions are fulfilled.

Case 2. Assume $c_i > 0$ and $\hat{\theta}_i \geq c_i$. By Equation (14), we know that $\theta_i^* = c_i$. Further, it follows that the first constraint of (10) is active, i. e. $\theta_i^* - c_i = 0$ and the second constraint is inactive, i.e. $-\theta_i^* - c_i = -2c_i < 0$. Thus, the first complementary slackness condition (12) is satisfied for all dual feasible points (11) $\mu_i^* \geq 0$. By the second complementary slackness condition we can conclude that $\mu_{i+r}^* = 0$. Now, the stationarity condition (9) simplifies to

$$2\left(\hat{\theta}_i - c_i\right) = \mu_i^* .$$

Due to the assumption we can argue that $2\left(\hat{\theta}_i - c_i\right) \geq 0$, such that all KKT conditions are fulfilled.

Case 3. Assume $c_i > 0$ and $\hat{\theta}_i \leq -c_i$. By Equation (14) we know that $\theta_i^* = -c_i$. Further, it follows that the second constraint of (10) is active, i. e. $-\theta_i^* - c_i = 0$ and the first constraint is inactive, i.e. $\theta_i^* - c_i = -2c_i < 0$. Thus, the second complementary slackness condition (12) is satisfied for all dual feasible points (11) $\mu_{i+r}^* \geq 0$. By the first complementary slackness condition we can conclude that $\mu_i^* = 0$. Now, the stationarity condition (9) simplifies to

$$-2\left(\hat{\theta}_i + c_i\right) = \mu_{i+r}^* .$$

Due to the assumption we can argue that $-2\left(\hat{\theta}_i + c_i\right) \geq 0$, such that all KKT conditions are fulfilled.

Case 4. Assume $c_i > 0$ and $\left|\hat{\theta}_i\right| < c_i$. By Equation (14), we know that $\theta_i^* = \hat{\theta}_i$. Further, it follows that both constraints of (10) are inactive, i. e. $\theta_i^* - c_i < 0$ and $-\theta_i^* - c_i < 0$. Hence, the complementary slackness conditions (12) are satisfied if $\mu_i^* = 0$ and $\mu_{i+r}^* = 0$. Moreover, the stationarity condition (9) simplifies to

$$2\left(\hat{\theta}_i - \hat{\theta}_i\right) = 0$$

which is true such that all KKT conditions are fulfilled.

Since i was arbitrarily chosen, we have proved the theorem. □

Summarizing the above results the KKT point $(\boldsymbol{\theta}^*, \boldsymbol{\mu}^*) \in \mathbb{R}^{r+2r}$ is given by the optimal solution $\boldsymbol{\theta}^*$ and the vector $\boldsymbol{\mu}^*$ defined to be

$$\mu_i^* = \begin{cases} 2\hat{\theta}_i & \text{if } c_i = 0 \text{ and } \hat{\theta}_i \geq 0 \\ 0 & \text{if } c_i = 0 \text{ and } \hat{\theta}_i < 0 \\ 2\left(\hat{\theta}_i - c_i\right) & \text{if } c_i > 0 \text{ and } \hat{\theta}_i \geq c_i \\ 0 & \text{if } c_i > 0 \text{ and } \left|\hat{\theta}_i\right| < c_i \\ 0 & \text{if } c_i > 0 \text{ and } \hat{\theta}_i \leq -c_i \end{cases}$$

$$\mu_{i+r}^* = \begin{cases} 0 & \text{if } c_i = 0 \text{ and } \hat{\theta}_i \geq 0 \\ -2\hat{\theta}_i & \text{if } c_i = 0 \text{ and } \hat{\theta}_i < 0 \\ 0 & \text{if } c_i > 0 \text{ and } \hat{\theta}_i \geq c_i \\ 0 & \text{if } c_i > 0 \text{ and } \left|\hat{\theta}_i\right| < c_i \\ -2\left(\hat{\theta}_i + c_i\right) & \text{if } c_i > 0 \text{ and } \hat{\theta}_i \leq -c_i \end{cases}$$

for $i \in \{1, 2, ..., r\}$.

4 Computing the Derivatives

In the previous section, we defined a closed-form solution of the rrTD which can be plugged into a machine learning framework. If the training of the machine learning approach is based on the optimization of a respective energy/ cost function, like stochastic gradient learning [12], we might need the gradients of this function regarding the tunable parameters. Moreover, we assume that the tunable parameters of the rrTD are the parameters of the subset description of the affine subspace, i. e. $\mathbf{w}$, $\mathbf{W}$ and $\mathbf{c}$ in our problem. In the following we prove that the rrTD is differentiable with respect to the variables $\mathbf{w}$, $\mathbf{W}$ and $\mathbf{c}$ and derive the respective formulas for the gradients.

Theorem 2. *Let $\mathbf{v} \in \mathbb{R}^n$ arbitrary but fixed. The function $\mathsf{d}_D(\mathbf{v}, \mathfrak{w}(\boldsymbol{\theta}))$ with the variables $\mathbf{w}$, $\mathbf{W}$ and $\mathbf{c}$ is differentiable over*

$$\mathcal{D}_{\mathbf{v}} = \left\{ (\mathbf{w}, \mathbf{W}, \mathbf{c}) \in \mathbb{R}^n \times \mathbb{R}^{n \times r} \times \mathbb{R}^r_{\geq 0} \middle| \, \mathbf{v} - \mathbf{w} \neq \mathbf{W}\boldsymbol{\theta}^*, \ \mathbf{c} > \mathbf{0}_r \right\} .$$

Before we start the proof we state some useful remarks, which later needed in the proof: The set $\mathcal{D}_{\mathbf{v}}$ is constructed over the whole domain of the rrTD with two additional restrictions to ensure the differentiability.

The rrTD with the solution (6) yields

$$\begin{aligned} \mathsf{d}_D(\mathbf{v}, \mathfrak{w}(\boldsymbol{\theta})) &= \left\| \mathbf{v} - \mathbf{w} - \mathbf{W}\left(\mathbf{H}\hat{\boldsymbol{\theta}} + \mathbf{S}(\mathbf{I}_r - \mathbf{H})\mathbf{c}\right) \right\|_E \\ &= \left\| \mathbf{v} - \mathbf{w} - \mathbf{W}\mathbf{H}\mathbf{W}^T(\mathbf{v} - \mathbf{w}) - \mathbf{W}\mathbf{S}(\mathbf{I}_r - \mathbf{H})\mathbf{c} \right\|_E \\ &= \left\| \left(\mathbf{I}_n - \mathbf{W}\mathbf{H}\mathbf{W}^T\right)(\mathbf{v} - \mathbf{w}) - \mathbf{W}\mathbf{S}(\mathbf{I}_r - \mathbf{H})\mathbf{c} \right\|_E \end{aligned}$$

11

with

$$\mathbf{H} = H\left(\mathbf{c} - \left|\hat{\boldsymbol{\theta}}\right|\right) = H\left(\mathbf{c} - \left|\mathbf{W}^T(\mathbf{v}-\mathbf{w})\right|\right)$$

$$\mathbf{S} = sgn\left(\hat{\boldsymbol{\theta}}\right) = sgn\left(\mathbf{W}^T(\mathbf{v}-\mathbf{w})\right)$$

and $\|\cdot\|_E$ as the Euclidean norm. Moreover, the matrix $\mathbf{P} = \mathbf{I}_n - \mathbf{W}\mathbf{H}\mathbf{W}^T$ is idempotent:

$$\begin{aligned}\mathbf{PP} &= \left(\mathbf{I}_n - \mathbf{W}\mathbf{H}\mathbf{W}^T\right)\left(\mathbf{I}_n - \mathbf{W}\mathbf{H}\mathbf{W}^T\right)\\ &= \mathbf{I}_n - \mathbf{W}\mathbf{H}\mathbf{W}^T - \mathbf{W}\mathbf{H}\mathbf{W}^T + \mathbf{W}\mathbf{H}\mathbf{W}^T\mathbf{W}\mathbf{H}\mathbf{W}^T\\ &= \mathbf{I}_n - \mathbf{W}\mathbf{H}\mathbf{W}^T - \mathbf{W}\mathbf{H}\mathbf{W}^T + \mathbf{W}\mathbf{H}\mathbf{W}^T\\ &= \mathbf{I}_n - \mathbf{W}\mathbf{H}\mathbf{W}^T\\ &= \mathbf{P}\end{aligned}$$

We need this property to prove the following lemma.

Lemma 2. *The following statements are equivalent:*

$$\begin{aligned}&\mathsf{d}_D^2\left(\mathbf{v}, \mathfrak{w}(\boldsymbol{\theta})\right)\\ =\ & \left\|\mathbf{P}(\mathbf{v}-\mathbf{w}) - \mathbf{W}\mathbf{S}\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{c}\right\|_E^2\\ =\ & (\mathbf{v}-\mathbf{w})^T\mathbf{P}(\mathbf{v}-\mathbf{w}) + \left(\mathbf{c}^T - 2(\mathbf{v}-\mathbf{w})^T\mathbf{W}\mathbf{S}\right)\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{c} \qquad (15)\end{aligned}$$

Proof. By expansion we obtain:

$$\begin{aligned}&\mathsf{d}_D^2\left(\mathbf{v}, \mathfrak{w}(\boldsymbol{\theta})\right)\\ =\ & \left(\mathbf{P}(\mathbf{v}-\mathbf{w}) - \mathbf{W}\mathbf{S}\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{c}\right)^T\left(\mathbf{P}(\mathbf{v}-\mathbf{w}) - \mathbf{W}\mathbf{S}\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{c}\right)\\ =\ & (\mathbf{v}-\mathbf{w})^T\mathbf{PP}(\mathbf{v}-\mathbf{w}) - (\mathbf{v}-\mathbf{w})^T\mathbf{PWS}\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{c}\\ & -\mathbf{c}^T\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{S}\mathbf{W}^T\mathbf{P}(\mathbf{v}-\mathbf{w}) + \mathbf{c}^T\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{S}\mathbf{W}^T\mathbf{W}\mathbf{S}\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{c}\end{aligned}$$

The expression $\mathbf{PW}$ simplifies to

$$\begin{aligned}\mathbf{PW} &= \left(\mathbf{I}_n - \mathbf{W}\mathbf{H}\mathbf{W}^T\right)\mathbf{W}\\ &= \mathbf{W}\left(\mathbf{I}_r - \mathbf{H}\right)\end{aligned}$$

using the property $\mathbf{W}^T\mathbf{W} = \mathbf{I}_r$ and, similarly,

$$\mathbf{W}^T\mathbf{P} = \left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{W}^T .$$

Thus the above expression yields:

$$\begin{aligned}&\mathsf{d}_D^2\left(\mathbf{v}, \mathfrak{w}(\boldsymbol{\theta})\right)\\ =\ & (\mathbf{v}-\mathbf{w})^T\mathbf{P}(\mathbf{v}-\mathbf{w}) - (\mathbf{v}-\mathbf{w})^T\mathbf{W}\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{S}\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{c}\\ & -\mathbf{c}^T\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{S}\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{W}^T(\mathbf{v}-\mathbf{w}) + \mathbf{c}^T\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{SS}\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{c}\end{aligned}$$

The quantities $\mathbf{S}$ and $(\mathbf{I}_r - \mathbf{H})$ are diagonal matrices. Therefore, we can change the order of multiplication. Further, the matrix $(\mathbf{I}_r - \mathbf{H})$ is idempotent because it is additionally a 0-1-matrix. At least, the second and third part of the expression are symmetric and we obtain:

$$\begin{aligned} & \mathsf{d}_D^2\left(\mathbf{v}, \mathfrak{w}(\boldsymbol{\theta})\right) \\ = \quad & (\mathbf{v}-\mathbf{w})^T \mathbf{P}(\mathbf{v}-\mathbf{w}) - 2(\mathbf{v}-\mathbf{w})^T \mathbf{W}\mathbf{S}\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{c} + \mathbf{c}^T \mathbf{S}\mathbf{S}\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{c} \end{aligned}$$

In the last step, we simplify the term $\mathbf{c}^T \mathbf{S}\mathbf{S}\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{c}$. For this purpose, we can rewrite the elements of the diagonal matrix $\mathbf{S}\mathbf{S}\left(\mathbf{I}_r - \mathbf{H}\right)$ as

$$\left|sgn\left(\hat{\theta}_i\right)\right|\left(1 - H\left(c_i - \left|\hat{\theta}_i\right|\right)\right) = \begin{cases} 0 & \text{if } \hat{\theta}_i = 0 \text{ or } \left|\hat{\theta}_i\right| \leq c_i \\ 1 & \text{else} \end{cases}$$

Since $c_i \in \mathbb{R}_{\geq 0}$ holds, we can conclude that the above equation is equivalent to

$$1 - H\left(c_i - \left|\hat{\theta}_i\right|\right) = \begin{cases} 0 & \text{if } \hat{\theta}_i = 0 \text{ or } \left|\hat{\theta}_i\right| \leq c_i \\ 1 & \text{else} \end{cases}$$

such that

$$\mathbf{S}\mathbf{S}\left(\mathbf{I}_r - \mathbf{H}\right) = \left(\mathbf{I}_r - \mathbf{H}\right) . \tag{16}$$

Applying this result, we can complete the proof of the Lemma:

$$\begin{aligned} & \mathsf{d}_D^2\left(\mathbf{v}, \mathfrak{w}(\boldsymbol{\theta})\right) \\ = \quad & (\mathbf{v}-\mathbf{w})^T \mathbf{P}(\mathbf{v}-\mathbf{w}) - 2(\mathbf{v}-\mathbf{w})^T \mathbf{W}\mathbf{S}\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{c} + \mathbf{c}^T\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{c} \\ = \quad & (\mathbf{v}-\mathbf{w})^T \mathbf{P}(\mathbf{v}-\mathbf{w}) + \left(\mathbf{c}^T - 2(\mathbf{v}-\mathbf{w})^T \mathbf{W}\mathbf{S}\right)\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{c} \end{aligned}$$

□

Now, we are able to compute the formal partial derivatives[3] of $\mathsf{d}_D\left(\mathbf{v}, \mathfrak{w}(\boldsymbol{\theta})\right)$. We apply straightforward the rules of differentiation [13]. Doing so, we have to determine the derivative of the square root for the differentiation for each variable. For example we have to consider

$$\frac{\partial \mathsf{d}_D\left(\mathbf{v}, \mathfrak{w}(\boldsymbol{\theta})\right)}{\partial \mathbf{w}} = \frac{1}{2\mathsf{d}_D\left(\mathbf{v}, \mathfrak{w}(\boldsymbol{\theta})\right)} \frac{\partial \mathsf{d}_D^2\left(\mathbf{v}, \mathfrak{w}(\boldsymbol{\theta})\right)}{\partial \mathbf{w}} \tag{17}$$

To simplify the following calculations, we consider the squared rrTD, only, and keep the coefficient (17) in mind.

Using the result of Lemma 2, the partial derivative with respect to $\mathbf{w}$ is given by:

[3] Here, "formal derivative" means that we ignore singularities and domain restrictions.

$$\begin{aligned}
&\frac{\partial \mathsf{d}_D^2\left(\mathbf{v}, \mathfrak{w}(\boldsymbol{\theta})\right)}{\partial \mathbf{w}} \\
=\ & \frac{\partial\left((\mathbf{v}-\mathbf{w})^T\mathbf{P}(\mathbf{v}-\mathbf{w}) + \left(\mathbf{c}^T - 2(\mathbf{v}-\mathbf{w})^T\mathbf{W}\mathbf{S}\right)\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{c}\right)}{\partial \mathbf{w}} \\
=\ & -2\mathbf{P}(\mathbf{v}-\mathbf{w}) + \frac{\partial\left(\mathbf{c}^T - 2(\mathbf{v}-\mathbf{w})^T\mathbf{W}\mathbf{S}\right)\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{c}}{\partial \mathbf{w}} \\
=\ & -2\mathbf{P}(\mathbf{v}-\mathbf{w}) + \frac{\partial\left(\mathbf{c}^T\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{c} - 2(\mathbf{v}-\mathbf{w})^T\mathbf{W}\mathbf{S}\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{c}\right)}{\partial \mathbf{w}}
\end{aligned}$$

The derivative of the Heaviside function is almost everywhere zero, such that it remains:

$$\begin{aligned}
&\frac{\partial \mathsf{d}_D^2\left(\mathbf{v}, \mathfrak{w}(\boldsymbol{\theta})\right)}{\partial \mathbf{w}} \\
=\ & -2\mathbf{P}(\mathbf{v}-\mathbf{w}) - 2\frac{\partial\left((\mathbf{v}-\mathbf{w})^T\mathbf{W}\mathbf{S}\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{c}\right)}{\partial \mathbf{w}} \\
=\ & -2\mathbf{P}(\mathbf{v}-\mathbf{w}) - 2\frac{\partial\left((\mathbf{v}^T\mathbf{W}\mathbf{S}\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{c} - \mathbf{w}^T\mathbf{W}\mathbf{S}\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{c}\right)}{\partial \mathbf{w}}
\end{aligned}$$

Again, the first expression is zero. Further, after applying the product rule of differentiation we observe that only one remaining expression is not equal to zero:

$$\begin{aligned}
&\frac{\partial \mathsf{d}_D^2\left(\mathbf{v}, \mathfrak{w}(\boldsymbol{\theta})\right)}{\partial \mathbf{w}} \\
=\ & -2\mathbf{P}(\mathbf{v}-\mathbf{w}) + 2\frac{\partial \mathbf{w}^T\mathbf{W}\mathbf{S}\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{c}}{\partial \mathbf{w}} \\
=\ & -2\mathbf{P}(\mathbf{v}-\mathbf{w}) + 2\frac{\partial \mathbf{w}^T}{\partial \mathbf{w}}\mathbf{W}\mathbf{S}\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{c} \\
=\ & -2\mathbf{P}(\mathbf{v}-\mathbf{w}) + 2\mathbf{W}\mathbf{S}\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{c} \\
=\ & -2(\mathbf{v}-\mathbf{w}) + 2\mathbf{W}\boldsymbol{\theta}^* \qquad (18)
\end{aligned}$$

Similar, we obtain the partial derivative with respect to $\mathbf{W}$ by:

$$\begin{aligned}
&\frac{\partial \mathsf{d}_D^2\left(\mathbf{v}, \mathfrak{w}(\boldsymbol{\theta})\right)}{\partial \mathbf{W}} \\
=\ & \frac{\partial\left((\mathbf{v}-\mathbf{w})^T\mathbf{P}(\mathbf{v}-\mathbf{w}) + \left(\mathbf{c}^T - 2(\mathbf{v}-\mathbf{w})^T\mathbf{W}\mathbf{S}\right)\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{c}\right)}{\partial \mathbf{W}} \\
=\ & -2(\mathbf{v}-\mathbf{w})(\mathbf{v}-\mathbf{w})^T\mathbf{W}\mathbf{H} + \frac{\partial\left(\mathbf{c}^T\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{c} - 2(\mathbf{v}-\mathbf{w})^T\mathbf{W}\mathbf{S}\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{c}\right)}{\partial \mathbf{W}} \\
=\ & -2(\mathbf{v}-\mathbf{w})(\mathbf{v}-\mathbf{w})^T\mathbf{W}\mathbf{H} - 2(\mathbf{v}-\mathbf{w})^T\frac{\partial \mathbf{W}\mathbf{S}\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{c}}{\partial \mathbf{W}} \\
=\ & -2(\mathbf{v}-\mathbf{w})(\mathbf{v}-\mathbf{w})^T\mathbf{W}\mathbf{H} - 2(\mathbf{v}-\mathbf{w})^T\frac{\partial \mathbf{W}}{\partial \mathbf{W}}\mathbf{S}\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{c} \\
=\ & -2(\mathbf{v}-\mathbf{w})(\mathbf{v}-\mathbf{w})^T\mathbf{W}\mathbf{H} - 2(\mathbf{v}-\mathbf{w})\mathbf{c}^T\mathbf{S}\left(\mathbf{I}_r - \mathbf{H}\right) \\
=\ & -2(\mathbf{v}-\mathbf{w})\left(\boldsymbol{\theta}^*\right)^T \qquad (19)
\end{aligned}$$

If $\mathbf{H} = \mathbf{I}_r$ the expressions are equal to the unrestricted case of the tangent distance with underlying Euclidean metric [8, 9] which should be intuitively true.

At least we derive the partial derivative with respect to $\mathbf{c}$:

$$\begin{aligned}
&\frac{\partial \mathsf{d}_D^2\left(\mathbf{v}, \mathfrak{w}(\boldsymbol{\theta})\right)}{\partial \mathbf{c}} \\
=\ & \frac{\partial\left((\mathbf{v}-\mathbf{w})^T\mathbf{P}(\mathbf{v}-\mathbf{w}) + \left(\mathbf{c}^T - 2(\mathbf{v}-\mathbf{w})^T\mathbf{W}\mathbf{S}\right)\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{c}\right)}{\partial \mathbf{c}} \\
=\ & \frac{\partial\left(\mathbf{c}^T\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{c} - 2(\mathbf{v}-\mathbf{w})^T\mathbf{W}\mathbf{S}\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{c}\right)}{\partial \mathbf{c}} \\
=\ & 2\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{c} - 2(\mathbf{v}-\mathbf{w})^T\mathbf{W}\mathbf{S}\frac{\partial\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{c}}{\partial \mathbf{c}} \\
=\ & 2\left(\mathbf{I}_r - \mathbf{H}\right)\left(\mathbf{c} - \mathbf{S}\mathbf{W}^T(\mathbf{v}-\mathbf{w})\right)
\end{aligned}$$

Applying the Equation (16) we can simplify the formula to:

$$\begin{aligned}
\frac{\partial \mathsf{d}_D^2\left(\mathbf{v}, \mathfrak{w}(\boldsymbol{\theta})\right)}{\partial \mathbf{c}} &= 2\mathbf{S}\left(\mathbf{S}\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{c} + \mathbf{H}\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}\right) \\
&= 2\mathbf{S}\left(\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\right) \\
&= -2\left|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\right| \qquad (20)
\end{aligned}$$

Now we have collected all preliminaries to proof the Theorem 2.

Proof of Theorem 2. It is well-known that a function $f(\mathbf{x})$ over an *open* set $\mathbf{x} \in D(f) \subseteq \mathbb{R}^n$ with the properties:

1. $f(\mathbf{x})$ is continuous over $D(f)$;
2. the partial derivatives $\frac{\partial f(\mathbf{x})}{\partial x_i}$ exist over $D(f)$ for all i;
3. the partial derivatives $\frac{\partial f(\mathbf{x})}{\partial x_i}$ are continuous over $D(f)$ for all i;

is differentiable [14]. Note that the three conditions are sufficient for the differentiability of f. We prove our theorem by validating these three conditions.

At first, it is obvious that the set

$$\mathcal{D}_{\mathbf{v}} = \left\{ (\mathbf{w}, \mathbf{W}, \mathbf{c}) \in \mathbb{R}^n \times \mathbb{R}^{n\times r} \times \mathbb{R}^r_{\geq 0} \middle|\ \mathbf{v} - \mathbf{w} \neq \mathbf{W}\boldsymbol{\theta}^*,\ \mathbf{c} > \mathbf{0}_r \right\}$$

in the theorem is an open set since we exclude the boundary points of the domain of $\mathbf{c}$ from the general domain of d_D. Thus, the overall assumption of the statement is fulfilled. This assumption is necessary since a function is, in general, not differentiable at boundary points.

Condition 1. The Euclidean distance d_E is a continuous function. Further, the rrTD can be expressed as Euclidean distance:

$$\mathsf{d}_D\left(\mathbf{v}, \mathfrak{w}(\boldsymbol{\theta})\right) = d_E\left(\mathbf{v}, \mathbf{w} + \mathbf{W}\boldsymbol{\theta}^*\right)$$

The optimal solution $\boldsymbol{\theta}^*$ is a function of the variables $\mathbf{v}$, $\mathbf{w}$, $\mathbf{W}$ and $\mathbf{c}$, i. e. $\boldsymbol{\theta}^* = \boldsymbol{\theta}^*\left(\mathbf{v}, \mathbf{w}, \mathbf{W}, \mathbf{c}\right)$. By equation (14) it is clear, that this function is continuous. Therefore, the argument $\mathbf{w} + \mathbf{W}\boldsymbol{\theta}^*$ is also continuous and, finally, d_D is continuous since it is a composition of continuous functions.

Condition 2. Above we derived the formulas of the formal partial derivatives. Now, we show that they are valid over $\mathcal{D}_{\mathbf{v}}$ and, further, we prove their existence. Obviously, the derivative of a real square root only exists if the argument is greater zero. Due to the metric properties of d_E we know that:

$$\begin{aligned} d_E\left(\mathbf{v}, \mathbf{w} + \mathbf{W}\boldsymbol{\theta}^*\right) = 0 \quad &\Longleftrightarrow \quad \mathbf{v} = \mathbf{w} + \mathbf{W}\boldsymbol{\theta}^* \\ &\Longleftrightarrow \quad \mathbf{v} - \mathbf{w} = \mathbf{W}\boldsymbol{\theta}^* \end{aligned}$$

Since we exclude such points in $\mathcal{D}_{\mathbf{v}}$, the derivative of the (real) square root of d_D (see Equation (17)) exists over $\mathcal{D}_{\mathbf{v}}$.

We obtain the formulas of the partial derivatives by neglecting the discontinuities of the Heaviside and signum function and the break point of the absolute value function. Beside these points, the partial derivatives exist over $\mathcal{D}_{\mathbf{v}}$, because of the well-known rules of differentiation. Now, it remains to prove that the derivative also exists at the discontinuities and the break point. Interestingly, the closed-form solution of $\boldsymbol{\theta}^*$ is continuous although it is a composition of discontinuous functions. Moreover, formulating the Equation (14) of the closed-form solution θ_i^* in a piecewise manner, it becomes obvious that the function has a break point if $\left|\hat{\theta}_i\right| = c_i$. Therefore, we only have to prove that the partial derivatives exist at these exceptional points. For this purpose, by the previous discussion, it is sufficient to prove that the partial derivatives exist over the squared rrTD.

The formula of the squared rrTD can be expressed by

$$\begin{aligned} \mathsf{d}_D^2\left(\mathbf{v}, \mathfrak{w}(\boldsymbol{\theta})\right) &= \left(\mathbf{v} - \mathbf{w} - \mathbf{W}\boldsymbol{\theta}^*\right)^T\left(\mathbf{v} - \mathbf{w} - \mathbf{W}\boldsymbol{\theta}^*\right) \\ &= \left(\mathbf{v} - \mathbf{w}\right)^T\left(\mathbf{v} - \mathbf{w}\right) - 2\left(\boldsymbol{\theta}^*\right)^T\mathbf{W}^T\left(\mathbf{v} - \mathbf{w}\right) + \left(\boldsymbol{\theta}^*\right)^T\boldsymbol{\theta}^* \\ &= \sum_{i=1}^{n}\left(v_i - w_i\right)^2 + \sum_{j=1}^{r}\left(\left(\theta_j^*\right)^2 - 2\theta_j^*\hat{\mathbf{w}}_j^T\left(\mathbf{v} - \mathbf{w}\right)\right) \\ &= \sum_{i=1}^{n}\left(v_i - w_i\right)^2 + \sum_{j=1}^{r}\left(\left(\theta_j^*\right)^2 - 2\theta_j^*\hat{\theta}_j\right) \end{aligned} \tag{21}$$

using the notation $\mathbf{W} = \left(\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2, \ldots, \hat{\mathbf{w}}_j, \ldots, \hat{\mathbf{w}}_r\right)$ where $\hat{\mathbf{w}}_j$ is the j-th column vector of $\mathbf{W}$.

We proof the existence of the partial derivatives at the break points by showing that the two one-sided limits from above and below of the partial differential quotient exist and are equal. Doing so we can conclude that the two-sided limit exists and it is equal to both one-sided limits. Further, we can easily show that the value of the limits is equal to the value of the general formula of the partial derivative of the respective component, which implies the correctness of the derived formulas, respectively.

In the following, we denote by arguments like $\theta_j^*(x)$, $\mathbf{c}(x)$, $\mathbf{W}(x)$, etc. the argument which is obtained by substituting the variable of interest by the new variable x. The slogan "variable of interest" means the variable which is considered during the limit computation. Moreover, we write

$$\mathsf{d}\left(\mathbf{v},\mathbf{w},\mathbf{W},\mathbf{c}\right)=\mathsf{d}_D\left(\mathbf{v},\mathfrak{w}(\boldsymbol{\theta})\right)$$

to indicate the dependencies regarding these variables. Finally, to simplify the mathematical notation of the proof we abbreviate

$$\theta_j=\theta_j^*\,.$$

Using these conventions, the Equation (21) simplifies to:

$$\begin{aligned}\mathsf{d}_D^2\left(\mathbf{v},\mathfrak{w}(\boldsymbol{\theta})\right) &= \mathsf{d}^2\left(\mathbf{v},\mathbf{w},\mathbf{W},\mathbf{c}\right)\\ &= \sum_{i=1}^{n}\left(v_i-w_i\right)^2+\sum_{j=1}^{r}\left(\theta_j^2-2\theta_j\hat{\theta}_j\right) \qquad (22)\end{aligned}$$

Without loss of generality, we assume $l\in\{1,2,...,r\}$ to be arbitrary but fixed. Further, we analyze the partial derivative

$$\frac{\partial\mathsf{d}^2\left(\mathbf{v},\mathbf{w},\mathbf{W},\mathbf{c}\right)}{\partial c_l}$$

at the break point $\hat{\theta}_l=c_l$. It follows immediately that $\theta_l=c_l$. The limit from below is given by:

$$\begin{aligned}&\lim_{x\nearrow c_j}\frac{\mathsf{d}^2\left(\mathbf{v},\mathbf{w},\mathbf{W},\mathbf{c}(x)\right)-\mathsf{d}^2\left(\mathbf{v},\mathbf{w},\mathbf{W},\mathbf{c}\right)}{x-c_l}\\ =&\lim_{x\nearrow c_l}\frac{\theta_l^2(x)-2\theta_l(x)\hat{\theta}_l-\theta_l^2+2\theta_l\hat{\theta}_l}{x-c_l}\\ =&\lim_{x\nearrow c_l}\frac{\left(\theta_l(x)-\theta_l\right)\left(\left(\theta_l(x)+\theta_l\right)-2c_l\right)}{x-c_l}\\ =&\lim_{x\nearrow c_l}\frac{\left(x-c_l\right)\left(\left(x+c_l\right)-2c_l\right)}{x-c_l}\\ =&\lim_{x\nearrow c_l}\left(\left(x+c_l\right)-2c_l\right)\\ =&\ 2c_l-2c_l\\ =&\ 0\end{aligned}$$

17

Similarly, we obtain the limit from above:

$$\begin{aligned} & \lim_{x \searrow c_l} \frac{\mathsf{d}^2\left(\mathbf{v}, \mathbf{w}, \mathbf{W}, \mathbf{c}(x)\right) - \mathsf{d}^2\left(\mathbf{v}, \mathbf{w}, \mathbf{W}, \mathbf{c}\right)}{x - c_l} \\ = & \lim_{x \searrow c_l} \frac{\left(\theta_l(x) - \theta_l\right)\left(\left(\theta_l(x) + \theta_l\right) - 2c_l\right)}{x - c_l} \\ = & \lim_{x \searrow c_l} \frac{\left(\hat{\theta}_l - c_l\right)\left(\left(\hat{\theta}_l + c_l\right) - 2c_l\right)}{x - c_l} \\ = & \lim_{x \searrow c_l} \frac{\left(\hat{\theta}_l - \hat{\theta}_l\right)\left(\left(\hat{\theta}_l + \hat{\theta}_l\right) - 2\hat{\theta}_l\right)}{x - c_l} \\ = & \ 0 \end{aligned}$$

Hence, the two one-sided limits are equal and, moreover, equivalent to the component l of the formula (20):

$$\begin{aligned} \left.\frac{\partial \mathsf{d}^2\left(\mathbf{v}, \mathbf{w}, \mathbf{W}, \mathbf{c}\right)}{\partial c_l}\right|_{\hat{\theta}_l = c_l} &= \left. 2\left(1 - H\left(c_l - \left|\hat{\theta}_l\right|\right)\right)\left(c_l - \left|\hat{\theta}_l\right|\right)\right|_{\hat{\theta}_l = c_l} \\ &= 0 \end{aligned}$$

Analogously, we derive the same result for the assumption $\hat{\theta}_l = -c_l$. Thus, the partial derivatives with respect to $\mathbf{c}$ exist over $\mathcal{D}_{\mathbf{v}}$.

We continue by proving that the partial derivatives exist for

$$\frac{\partial \mathsf{d}^2\left(\mathbf{v}, \mathbf{w}, \mathbf{W}, \mathbf{c}\right)}{\partial \mathbf{W}}.$$

Without loss of generality, we assume $l \in \{1, 2, ..., r\}$ and $k \in \{1, 2, ..., n\}$ to be arbitrary but fixed. Again, we analyze the partial derivative at the break point $\hat{\theta}_l = c_l$ such that the equality $\theta_l = \hat{\theta}_l$ follows immediately. Further, the function $\hat{\theta}_l(x)$ is linear in x and can be expressed by:

$$\begin{aligned} \hat{\theta}_l(x) &= \hat{\mathbf{w}}_l^T(x)\left(\mathbf{v} - \mathbf{w}\right) \\ &= \left(\hat{\mathbf{w}}_l + \mathbf{e}_k\left(x - W_{kl}\right)\right)^T\left(\mathbf{v} - \mathbf{w}\right) \\ &= \hat{\mathbf{w}}_l^T\left(\mathbf{v} - \mathbf{w}\right) + \left(v_k - w_k\right)\left(x - W_{kl}\right) \end{aligned}$$

Now, without loss of generality, we can assume that $x \nearrow W_{kl}$ implies the in-

18

equality $\hat{\theta}_l(x) \leq c_l$. The limit from below is given by:

$$\begin{aligned}
& \lim_{x \nearrow W_{kl}} \frac{\mathsf{d}^2(\mathbf{v}, \mathbf{w}, \mathbf{W}(x), \mathbf{c}) - \mathsf{d}^2(\mathbf{v}, \mathbf{w}, \mathbf{W}, \mathbf{c})}{x - W_{kl}} \\
= & \lim_{x \nearrow W_{kl}} \frac{\sum_{j=1}^{r} \left(\theta_j^2(x) - 2\theta_j(x)\hat{\theta}_j(x)\right) - \sum_{j=1}^{r} \left(\theta_j^2 - 2\theta_j\hat{\theta}_j\right)}{x - W_{kl}} \\
= & \lim_{x \nearrow W_{kl}} \frac{\left(\hat{\theta}_l^2(x) - 2\hat{\theta}_l^2(x)\right) - \left(\hat{\theta}_l^2 - 2\hat{\theta}_l^2\right)}{x - W_{kl}} \\
= & \lim_{x \nearrow W_{kl}} \frac{-\left(\hat{\theta}_l^2(x) - \hat{\theta}_l^2\right)}{x - W_{kl}} \\
= & \lim_{x \nearrow W_{kl}} \frac{-\left(\hat{\theta}_l(x) - \hat{\theta}_l\right)\left(\hat{\theta}_l(x) + \hat{\theta}_l\right)}{x - W_{kl}} \\
= & \lim_{x \nearrow W_{kl}} \frac{-\left(x - W_{kl}\right)\left(v_k - w_k\right)\left(\hat{\theta}_l(x) + \hat{\theta}_l\right)}{x - W_{kl}} \\
= & -2\left(v_k - w_k\right)\hat{\theta}_l
\end{aligned}$$

By the last assumption, it follows that $x \searrow W_{kl}$ implies the inequality $\hat{\theta}_l(x) \geq c_l$ and, hence, the limit from above is obtained as

$$\begin{aligned}
& \lim_{x \searrow W_{kl}} \frac{\mathsf{d}^2(\mathbf{v}, \mathbf{w}, \mathbf{W}(x), \mathbf{c}) - \mathsf{d}^2(\mathbf{v}, \mathbf{w}, \mathbf{W}, \mathbf{c})}{x - W_{kl}} \\
= & \lim_{x \searrow W_{kl}} \frac{\sum_{j=1}^{r} \left(\theta_j^2(x) - 2\theta_j(x)\hat{\theta}_j(x)\right) - \sum_{j=1}^{r} \left(\theta_j^2 - 2\theta_j\hat{\theta}_j\right)}{x - W_{kl}} \\
= & \lim_{x \searrow W_{kl}} \frac{\left(\hat{\theta}_l^2 - 2\hat{\theta}_l\hat{\theta}_l(x)\right) - \left(\hat{\theta}_l^2 - 2\hat{\theta}_l^2\right)}{x - W_{kl}} \\
= & \lim_{x \searrow W_{kl}} \frac{-2\hat{\theta}_l\left(\hat{\theta}_l(x) - \hat{\theta}_l\right)}{x - W_{kl}} \\
= & \lim_{x \searrow W_{kl}} \frac{-2\left(x - W_{kl}\right)\left(v_k - w_k\right)\hat{\theta}_l}{x - W_{kl}} \\
= & -2\left(v_k - w_k\right)\hat{\theta}_l
\end{aligned}$$

Again, the two one-sided limits are equal and, moreover, equivalent to the component (k, l) of the formula (19):

$$\begin{aligned}
& \left.\frac{\partial \mathsf{d}^2\left(\mathbf{v}, \mathbf{w}, \mathbf{W}, \mathbf{c}\right)}{\partial W_{kl}}\right|_{\hat{\theta}_l = c_l} \\
= \quad & -2(v_k - w_k)\left.\left(\hat{\theta}_l H\left(c_l - \left|\hat{\theta}_l\right|\right) + c_l sgn\left(\hat{\theta}_l\right)\left(1 - H\left(c_l - \left|\hat{\theta}_l\right|\right)\right)\right)\right|_{\hat{\theta}_l = c_l} \\
= \quad & -2\left(v_k - w_k\right)\hat{\theta}_l
\end{aligned}$$

Analogously, we derive the same result for the assumption $\hat{\theta}_l = -c_l$. In conclusion, the partial derivatives with respect to $\mathbf{W}$ exist over $\mathcal{D}_{\mathbf{v}}$.

Finally, we have to prove the existence of

$$\frac{\partial \mathrm{d}^2\left(\mathbf{v}, \mathbf{w}, \mathbf{W}, \mathbf{c}\right)}{\partial w_k}$$

where $k \in \{1, 2, ..., r\}$ is, without loss of generality, arbitrary but fixed. For this purpose, we define the following subsets:

$$\begin{aligned}
I_{=}^{-} &:= \left\{ j \in \{1, 2, \ldots, r\} \,|\, \hat{\theta}_j = -c_j \right\} \\
I_{=}^{+} &:= \left\{ j \in \{1, 2, \ldots, r\} \,|\, \hat{\theta}_j = c_j \right\} \\
I^{-} &:= \left\{ j \in \{1, 2, \ldots, r\} \,|\, \hat{\theta}_j < -c_j \right\} \\
I^{+} &:= \left\{ j \in \{1, 2, \ldots, r\} \,|\, \hat{\theta}_j > c_j \right\} \\
I &:= \left\{ j \in \{1, 2, \ldots, r\} \,|\, \left|\hat{\theta}_j\right| < c_j \right\}
\end{aligned}$$

Since the function $\hat{\theta}_j(x)$ is linear in the new variable x according to

$$\begin{aligned}
\hat{\theta}_j(x) &= \hat{\mathbf{w}}_j^T \left(\mathbf{v} - \mathbf{w} - \mathbf{e}_k(x - w_k)\right) \\
&= \hat{\mathbf{w}}_j^T \left(\mathbf{v} - \mathbf{w}\right) - W_{kj}\left(x - w_k\right)
\end{aligned} \tag{23}$$

we can conclude the implication: if the limit $x \nearrow w_k$ is valid, the value $\hat{\theta}_j(x)$ converges uniquely from above or below to $\hat{\theta}_j$. Therefore, it is intuitive to define the refined subsets:

$$\begin{aligned}
J_{<}^{-} &:= \left\{ j \in I_{=}^{-} | \hat{\theta}_j(x) \leq -c_j \text{ if } x \nearrow w_k \wedge W_{kj} \neq 0 \right\} \\
J_{>}^{-} &:= \left\{ j \in I_{=}^{-} | \hat{\theta}_j(x) \geq -c_j \text{ if } x \nearrow w_k \right\} \\
J_{<}^{+} &:= \left\{ j \in I_{=}^{+} | \hat{\theta}_j(x) \leq c_j \text{ if } x \nearrow w_k \right\} \\
J_{>}^{+} &:= \left\{ j \in I_{=}^{+} | \hat{\theta}_j(x) \geq c_j \text{ if } x \nearrow w_k \wedge W_{kj} \neq 0 \right\}
\end{aligned}$$

These reformulations allow the following implications:

$$\begin{array}{lcccccl}
j \in J_{<}^{-} & \Longrightarrow & \theta_j(x) = -c_j & \wedge & \theta_j = -c_j & \text{if } x \nearrow w_k \\
j \in J_{>}^{-} & \Longrightarrow & \theta_j(x) = \hat{\theta}_j(x) & \wedge & \theta_j = -c_j & \text{if } x \nearrow w_k \\
j \in J_{<}^{+} & \Longrightarrow & \theta_j(x) = \hat{\theta}_j(x) & \wedge & \theta_j = c_j & \text{if } x \nearrow w_k \\
j \in J_{>}^{+} & \Longrightarrow & \theta_j(x) = c_j & \wedge & \theta_j = c_j & \text{if } x \nearrow w_k
\end{array} \tag{24}$$

Since $c_j > 0$, we can write the set $\{1, 2, ..., r\}$ as the union

$$\{1, 2, \ldots, r\} = J_{<}^{-} \cup J_{>}^{-} \cup J_{<}^{+} \cup J_{>}^{+} \cup I^{-} \cup I^{+} \cup I$$

over *pairwise disjoint* sets. We take

$$J = \left\{J_<^-, J_>^-, J_<^+, J_>^+, I^-, I^+, I\right\}$$

as the family of sets.

Using Equation (22), the distance $\mathsf{d}^2\left(\mathbf{v}, \mathbf{w}, \mathbf{W}, \mathbf{c}\right)$ can be rewritten as[4]

$$\begin{aligned} & \mathsf{d}^2\left(\mathbf{v}, \mathbf{w}, \mathbf{W}, \mathbf{c}\right) \\ = & \sum_{j=1}^{n} \left(v_j - w_j\right)^2 + \sum_{M \in J} \sum_{j \in M} \left(\theta_j^2 - 2\theta_j \hat{\theta}_j\right) \\ = & \sum_{j=1}^{n} \left(v_j - w_j\right)^2 - \sum_{j \in J_<^-} c_j^2 - \sum_{j \in J_>^-} c_j^2 \\ & - \sum_{j \in J_<^+} c_j^2 - \sum_{j \in J_>^+} c_j^2 + \sum_{j \in I^-} \left(c_j^2 + 2c_j \hat{\theta}_j\right) \\ & + \sum_{j \in I^+} \left(c_j^2 - 2c_j \hat{\theta}_j\right) - \sum_{j \in I} \hat{\theta}_j^2 \end{aligned}$$

and, analogously, $\mathsf{d}^2\left(\mathbf{v}, \mathbf{w}(x), \mathbf{W}, \mathbf{c}\right)$:

$$\begin{aligned} & \mathsf{d}^2\left(\mathbf{v}, \mathbf{w}(x), \mathbf{W}, \mathbf{c}\right) \\ = & \sum_{j=1; j \neq k}^{n} \left(v_j - w_j\right)^2 + \left(v_k - x\right)^2 + \sum_{M \in J} \sum_{j \in M} \left(\theta_j^2(x) - 2\theta_j(x) \hat{\theta}_j(x)\right) \\ = & \sum_{j=1; j \neq k}^{n} \left(v_j - w_j\right)^2 + \left(v_k - x\right)^2 + \sum_{j \in J_<^-} \left(c_j^2 + 2c_j \hat{\theta}_j(x)\right) - \sum_{j \in J_>^-} \hat{\theta}_j^2(x) \\ & - \sum_{j \in J_<^+} \hat{\theta}_j^2(x) + \sum_{j \in J_>^+} \left(c_j^2 - 2c_j \hat{\theta}_j(x)\right) + \sum_{j \in I^-} \left(c_j^2 + 2c_j \hat{\theta}_j(x)\right) \\ & + \sum_{j \in I^+} \left(c_j^2 - 2c_j \hat{\theta}_j(x)\right) - \sum_{j \in I} \hat{\theta}_j^2(x) \end{aligned}$$

[4] To avoid a too heavy mathematical notation, we use the same summation index for all sums and define that the scope of the sum is only the immediately following expression.

Using all these results, we obtain for the limit from below:

$$\begin{aligned}
& \lim_{x\nearrow w_k} \frac{\mathsf{d}^2\left(\mathbf{v},\mathbf{w}(x),\mathbf{W},\mathbf{c}\right)-\mathsf{d}^2\left(\mathbf{v},\mathbf{w},\mathbf{W},\mathbf{c}\right)}{x-w_k} \\
= & \lim_{x\nearrow w_k} \left(\frac{\left(v_k-x\right)^2-\left(v_k-w_k\right)^2+2\sum_{j\in J^-_<} c_j\left(c_j+\hat{\theta}_j(x)\right)}{x-w_k} \right. \\
& +\frac{-\sum_{j\in J^-_>}\left(\hat{\theta}^2_j(x)-c^2_j\right)-\sum_{j\in J^+_<}\left(\hat{\theta}^2_j(x)-c^2_j\right)}{x-w_k} \\
& +\frac{2\sum_{j\in J^+_>} c_j\left(c_j-\hat{\theta}_j(x)\right)+2\sum_{j\in I^-} c_j\left(\hat{\theta}_j(x)-\hat{\theta}_j\right)}{x-w_k} \\
& \left. +\frac{-2\sum_{j\in I^+} c_j\left(\hat{\theta}_j(x)-\hat{\theta}_j\right)-\sum_{j\in I}\left(\hat{\theta}^2_j(x)-\hat{\theta}^2_j\right)}{x-w_k} \right)
\end{aligned}$$

We use the following simplification principles for the respective expressions to simplify the limit: **a)** Sums of the form $\sum_{j\in J^-_<} c_j\left(c_j+\hat{\theta}_j(x)\right)$ are simplified by the corresponding property of Equation (24) and the formula (23)

$$\begin{aligned}
c_j\left(c_j+\hat{\theta}_j(x)\right) &= -c_j\left(-c_j-\hat{\theta}_j(x)\right) \\
&= -c_j\left(\hat{\mathbf{w}}^T_j\left(\mathbf{v}-\mathbf{w}\right)-\hat{\mathbf{w}}^T_j\left(\mathbf{v}-\mathbf{w}(x)\right)\right) \\
&= -c_j\hat{\mathbf{w}}^T_j\mathbf{e}_k\left(x-w_k\right) \\
&= -c_jW_{kj}\left(x-w_k\right)
\end{aligned}$$

and **b)** sums of the type $\sum_{j\in J^+_<}\left(\hat{\theta}^2_j(x)-c^2_j\right)$ to:

$$\begin{aligned}
\left(\hat{\theta}^2_j(x)-c^2_j\right) &= \left(\hat{\theta}_j(x)-\hat{\theta}_j\right)\left(\hat{\theta}_j(x)+c_j\right) \\
&= \left(\hat{\mathbf{w}}^T_j\left(\mathbf{v}-\mathbf{w}(x)\right)-\hat{\mathbf{w}}^T_j\left(\mathbf{v}-\mathbf{w}\right)\right)\left(\hat{\theta}_j(x)+c_j\right) \\
&= -\hat{\mathbf{w}}^T_j\mathbf{e}_k\left(x-w_k\right)\left(\hat{\theta}_j(x)+c_j\right) \\
&= -W_{kj}\left(x-w_k\right)\left(\hat{\theta}_j(x)+c_j\right)
\end{aligned}$$

Applying both concepts a) and b) for simplification to all sums we obtain for the limit from below:

$$
\begin{aligned}
& \lim_{x \nearrow w_k} \frac{\mathsf{d}^2\left(\mathbf{v}, \mathbf{w}(x), \mathbf{W}, \mathbf{c}\right) - \mathsf{d}^2\left(\mathbf{v}, \mathbf{w}, \mathbf{W}, \mathbf{c}\right)}{x - w_k} \\
= \quad & \lim_{x \nearrow w_k} \left(\frac{-\left(2v_k - x - w_k\right)\left(x - w_k\right) - 2\sum_{j \in J^-_<} c_j W_{kj}\left(x - w_k\right)}{x - w_k} \right. \\
& + \frac{\sum_{j \in J^-_>} W_{kj}\left(x - w_k\right)\left(\hat{\theta}_j(x) - c_j\right) + \sum_{j \in J^+_<} W_{kj}\left(x - w_k\right)\left(\hat{\theta}_j(x) + c_j\right)}{x - w_k} \\
& + \frac{2\sum_{j \in J^+_>} c_j W_{kj}\left(x - w_k\right) - 2\sum_{j \in I^-} c_j W_{kj}\left(x - w_k\right)}{x - w_k} \\
& \left. + \frac{2\sum_{j \in I^+} c_j W_{kj}\left(x - w_k\right) + \sum_{j \in I} W_{kj}\left(x - w_k\right)\left(\hat{\theta}_j(x) + \hat{\theta}_j\right)}{x - w_k} \right) \\
= \quad & \lim_{x \nearrow w_k} \left(-\left(2v_k - x - w_k\right) - 2\sum_{j \in J^-_<} c_j W_{kj} + \sum_{j \in J^-_>} W_{kj}\left(\hat{\theta}_j(x) - c_j\right) \right. \\
& + \sum_{j \in J^+_<} W_{kj}\left(\hat{\theta}_j(x) + c_j\right) + 2\sum_{j \in J^+_>} c_j W_{kj} - 2\sum_{j \in I^-} c_j W_{kj} \\
& \left. +2\sum_{j \in I^+} c_j W_{kj} + \sum_{j \in I} W_{kj}\left(\hat{\theta}_j(x) + \hat{\theta}_j\right) \right) \\
= \quad & -2\left(v_k - w_k\right) - 2\sum_{j \in J^-_<} c_j W_{kj} - 2\sum_{j \in J^-_>} c_j W_{kj} \\
& +2\sum_{j \in J^+_<} c_j W_{kj} + 2\sum_{j \in J^+_>} c_j W_{kj} - 2\sum_{j \in I^-} c_j W_{kj} \\
& +2\sum_{j \in I^+} c_j W_{kj} + 2\sum_{j \in I} W_{kj}\hat{\theta}_j \\
= \quad & -2\left(v_k - w_k\right) - 2\sum_{j \in I^-_= \cup I^-} c_j W_{kj} + 2\sum_{j \in I^+_= \cup I^+} c_j W_{kj} + 2\sum_{j \in I} W_{kj}\hat{\theta}_j
\end{aligned}
$$

In the next step, we calculate the limit from above $x \searrow w_k$ and, again, start with analogous definitions of index sets:

$$
\begin{aligned}
K^-_< \quad &:= \quad \left\{ j \in I^-_= | \hat{\theta}_j(x) \leq -c_j \text{ if } x \searrow w_k \wedge W_{kj} \neq 0 \right\} \\
K^-_> \quad &:= \quad \left\{ j \in I^-_= | \hat{\theta}_j(x) \geq -c_j \text{ if } x \searrow w_k \right\} \\
K^+_< \quad &:= \quad \left\{ j \in I^+_= | \hat{\theta}_j(x) \leq c_j \text{ if } x \searrow w_k \right\} \\
K^+_> \quad &:= \quad \left\{ j \in I^+_= | \hat{\theta}_j(x) \geq c_j \text{ if } x \searrow w_k \wedge W_{kj} \neq 0 \right\}
\end{aligned}
$$

As before, we are able to formulate the following implications regarding these sets

$$\begin{array}{rclcll}
j \in K^-_< & \Longrightarrow & \theta_j(x) = -c_j & \wedge & \theta_j = -c_j & \text{if } x \searrow w_k \\
j \in K^-_> & \Longrightarrow & \theta_j(x) = \hat{\theta}_j(x) & \wedge & \theta_j = -c_j & \text{if } x \searrow w_k \\
j \in K^+_< & \Longrightarrow & \theta_j(x) = \hat{\theta}_j(x) & \wedge & \theta_j = c_j & \text{if } x \searrow w_k \\
j \in K^+_> & \Longrightarrow & \theta_j(x) = c_j & \wedge & \theta_j = c_j & \text{if } x \searrow w_k
\end{array}$$

which allow the explicit calculation of the limits according to

$$\begin{aligned}
& \lim_{x \searrow w_k} \frac{\mathsf{d}^2\left(\mathbf{v}, \mathbf{w}(x), \mathbf{W}, \mathbf{c}\right) - \mathsf{d}^2\left(\mathbf{v}, \mathbf{w}, \mathbf{W}, \mathbf{c}\right)}{x - w_k} \\
= \quad & -2\left(v_k - w_k\right) - 2 \sum_{j \in K^-_<} c_j W_{kj} - 2 \sum_{j \in K^-_>} c_j W_{kj} \\
& +2 \sum_{j \in K^+_<} c_j W_{kj} + 2 \sum_{j \in K^+_>} c_j W_{kj} - 2 \sum_{j \in I^-} c_j W_{kj} \\
& +2 \sum_{j \in I^+} c_j W_{kj} + 2 \sum_{j \in I} W_{kj} \hat{\theta}_j \\
= \quad & -2\left(v_k - w_k\right) - 2 \sum_{j \in I^-_= \cup I^-} c_j W_{kj} + 2 \sum_{j \in I^+_= \cup I^+} c_j W_{kj} + 2 \sum_{j \in I} W_{kj} \hat{\theta}_j
\end{aligned}$$

Again, both one-sided limits are equal and, moreover, equivalent to the component k of the formula (18)

$$\begin{aligned}
& \frac{\partial \mathsf{d}^2\left(\mathbf{v}, \mathbf{w}, \mathbf{W}, \mathbf{c}\right)}{\partial w_k} \\
= \quad & -2\mathbf{e}_k^T \left(\mathbf{I}_n - \mathbf{W}\mathbf{H}\mathbf{W}^T\right)\left(\mathbf{v} - \mathbf{w}\right) + 2\mathbf{e}_k^T \mathbf{W}\mathbf{S}\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{c} \\
= \quad & -2(v_k - w_k) + 2\check{\mathbf{w}}_k \left(\mathbf{H}\hat{\boldsymbol{\theta}} + \mathbf{S}\left(\mathbf{I}_r - \mathbf{H}\right)\mathbf{c}\right) \\
= \quad & -2(v_k - w_k) + 2\check{\mathbf{w}}_k \boldsymbol{\theta}^* \qquad (25)
\end{aligned}$$

where $\check{\mathbf{w}}_k$ is the k-the row vector of $\mathbf{W}$. Thus, the partial derivatives with respect to $\mathbf{w}$ exist over $\mathcal{D}_\mathbf{v}$, which proves the overall partial differentiability of $\mathsf{d}^2_D\left(\mathbf{v}, \mathfrak{w}(\boldsymbol{\theta})\right)$.

Condition 3. Considering the Equations (17) – (20), it is straightforward to see that the partial derivatives are continuous over $\mathcal{D}_\mathbf{v}$.

□

Corollary 1. *Since the function is differentiable over $\mathcal{D}_\mathbf{v}$ it follows that the relation between the gradient $\nabla f(\mathbf{x})$ and the partial derivatives $\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}$ holds as*

$$\nabla f(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}$$

such that we can compute the gradients of $\mathrm{d}^2_D\left(\mathbf{v}, \mathfrak{w}(\boldsymbol{\theta})\right)$ *over* $\mathcal{D}_{\mathbf{v}}$ *by simply applying the formulas (17) – (20).*

5 Conclusion

We have shown in this paper that the restricted minimization problem (4) can be solved analytically and, therefore, the shortest distance problem has a close solution formula (15). Further, we have proven that the closed-form solution is differentiable almost everywhere and derive the gradients with respect to the variables of interest, see Equation (17) – (20). Using these formulas, it is possible to plug the distance measure in existing distance based machine learning frameworks, where the training procedure optimizes a given energy function by a gradient learning scheme. A respective framework is the GLVQ method. The application of the proposed concept for the GLVQ approach will be the topic of ongoing research.

Acknowledgment

The authors would like to thank Mehrdad Mohannazadeh Bakhtiari from the Computational Intelligence Group of the University of Applied Sciences Mittweida for proofreading.

References

[1] P. Simard, Y. LeCun, and J.S. Denker. Efficient pattern recognition using a new transformation distance. In S.J. Hanson, J.D. Cowan, and C.L. Giles, editors, *Advances in Neural Information Processing Systems 5*, pages 50–58. Morgan-Kaufmann, 1993.

[2] D. Nebel, M. Kaden, A. Bohnsack, and T. Villmann. Types of (dis−)similarities and adaptive mixtures thereof for improved classification learning. *Neurocomputing*, 268:42–54, 2017.

[3] T. Hastie, P. Simard, and E. Säckinger. Learning prototype models for tangent distance. In G. Tesauro, D.S. Touretzky, and T.K. Leen, editors, *Advances in Neural Information Processing Systems 7*, pages 999–1006. MIT Press, 1995.

[4] B. Haasdonk and D. Keysers. Tangent distance kernels for support vector machines. In R. Kasturi, D. Laurendeau, and C. Suen, editors, *Proceedings of the 16th International Conference on Pattern Recognition (ICPR), Québec City*, volume 2, pages 864–868. IEEE Press, Los Alamitos, California, 2002.

[5] D. Keysers, W. Macherey, H. Ney, and J. Dahmen. Adaptation in statistical pattern recognition using tangent vectors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):269–274, 2004.

[6] D. Sona, A. Sperduti, and A. Starita. Discriminant pattern recognition using transformation-invariant neurons. *Neural Computation*, 12:1355–1370, 2000.

[7] D. Keysers, J. Dahmen, T. Theiner, and H. Ney. Experiments with an extended tangent distance. In A. Sanfeliu, J. J. Villanueva, M. Vanrell, R. Alquékzar, J. Crowley, and Y. Shirai, editors, *Proceedings of the 15th International Conference on Pattern Recognition, Barcelona*, volume 2, pages 38–42. IEEE Press, Los Alamitos, California, 2000.

[8] S. Saralajew and T. Villmann. Adaptive tangent metrics in generalized learning vector quantization for transformation and distortion invariant classification learning. In *Proceedings of the International Joint Conference on Neural networks (IJCNN) , Vancouver*, pages 2672–2679. IEEE Computer Society Press, 2016.

[9] S. Saralajew, D. Nebel, and T. Villmann. Adaptive Hausdorff distances and tangent distance adaptation for transformation invariant classification learning. In A. Hirose, editor, *Proceedings of the International Conference on Neural Information Processing (ICONIP) , Kyoto*, volume 9949 of *LNCS*, pages 362–371. Springer, 2016.

[10] S. Saralajew and T. Villmann. Transfer learning in classification based on manifold models and its relation to tangent metric learning. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN), Anchorage*, pages 1756–1765. IEEE Computer Society Press, 2017.

[11] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press„ 2004.

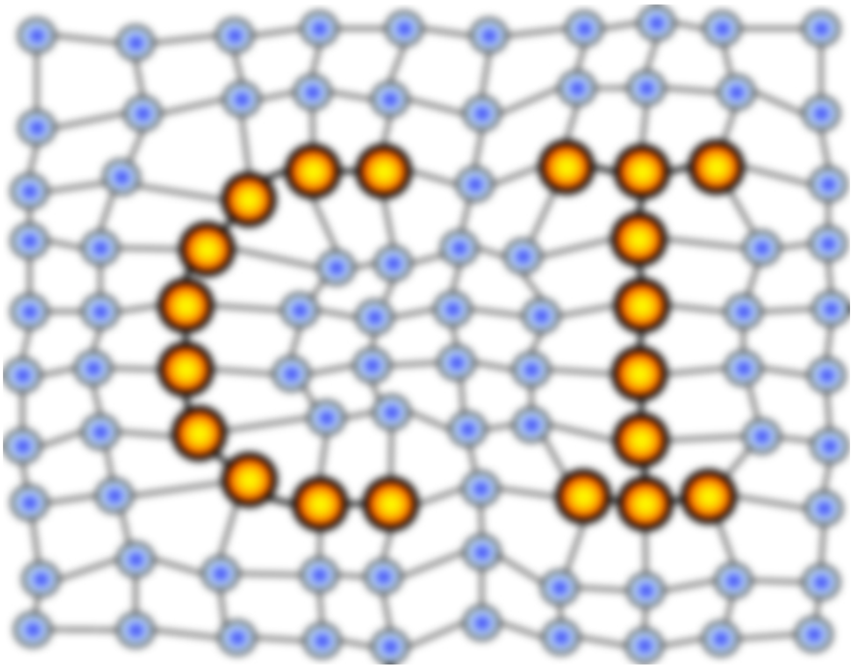
[12] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning.* MIT Press, 2016.

[13] K. Brand Petersen and M. Syskind Pedersen. The Matrix Cookbook. `http://matrixcookbook.com`, 2012.

[14] A.N. Kolmogorov and S.V. Fomin. *Reelle Funktionen und Funktionalanalysis.* VEB Deutscher Verlag der Wissenschaften, Berlin, 1975.

Machine Learning Reports

Report 01/2017



Impressum

Machine Learning Reports ISSN: 1865-3960

▽ Publisher/Editors

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

Dr. rer. nat. Frank-Michael Schleif
University of Bielefeld
Universitätsstrasse 21-23, 33615 Bielefeld, Germany
• http://www.cit-ec.de/tcs/about

▽ Acknowledgments

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